Please cite this paper as:
J. Lasek, M. Gagolewski, Interpretable sport team rating models based on the gradient descent algorithm, International Journal of Forecasting 37(3), 1061–1071, 2021, doi:10.1016/j.ijforecast.2020.11.008

Interpretable sport team rating models based on the gradient descent algorithm

Jan Lasek[a,*], Marek Gagolewski[b,a]

[a] *Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland, janek.lasek@gmail.com*
[b] *School of Information Technology, Deakin University, Geelong, VIC 3220, Australia, m.gagolewski@deakin.edu.au*

Abstract

We introduce several new sport team rating models based upon the gradient descent algorithm. More precisely, the models can be formulated by maximising the likelihood of match results observed using a single step of this optimisation heuristic. The framework proposed, inspired by the prominent Elo rating system, yields an iterative version of the ordinal logistic regression as well as different variants of the Poisson regression-based models. This construction makes the update equations easy to interpret as well as adjusts ratings once new match results are observed. Thus, it naturally handles temporal changes in team strength. Moreover, a study of association football data indicates that the new models yield more accurate forecasts and are less computationally demanding than corresponding methods that jointly optimise likelihood for the whole set of matches.

Keywords: rating systems, association football, match outcome forecasting, gradient descent, Poisson regression, ordinal logistic regression, Elo rating system

1. Introduction

Sport team rating models have numerous applications, including forecasting match outcomes, providing team seedings for tournaments and qualifying rounds, scheduling tournaments, creating interesting match-ups, or even granting players work permits at the international level. They also provide contestants with a measure of their progress and overall strength. The recent surge of e-sports has made rating systems even more important – these settings are much more computationally demanding as they typically involve a large number of contestants. This further justifies efforts to design accurate and scalable rating models. Moreover, the widening critique of black-box modelling has made interpretability and transparency all the more important.

In this paper we look at the design and evaluation of intuitive and interpretable rating systems illustrated by the example of the most popular sport in the world – association football (football for short, sometimes referred to

*Corresponding author

as soccer). As a workhorse for designing such systems, we shall use a popular optimisation heuristic – the gradient (or steepest) descent algorithm (Nocedal & Wright, 2006). This simple approach lays the groundwork for building interpretable rating systems that can be easily adjusted once new match results are observed. Different models can be formulated by maximising the likelihood of match results observed using a single step of gradient descent algorithm.

As for evaluations of different team rating models, it is widely accepted that they should be compared on the basis of their accuracy in forecasting match outcomes (Barrow et al., 2013; Boshnakov et al., 2017; Lasek et al., 2013; Ley et al., 2019). The models are also often compared to the predictions derived from bookmaker odds. We follow these practices in evaluating the models proposed. We shall demonstrate that the new ratings we derive using the gradient descent approach are easily updated and are more accurate in predicting future match outcomes than their counterparts estimated jointly on the whole sample of results.

The Elo rating system is the most prominent example of an iterative rating scheme (Elo, 1961, 1978). It has a transparent interpretation and a simple rating update rule. Its elegant formulation, accuracy and interpretability contributes to its popularity and accounts for its frequent deployment for various sports (Stefani, 2011). Many extensions of this classic model have been proposed (Glickman, 1999; Herbrich et al., 2006). It has also been employed to assess areas as diverse as educational systems (Pelánek, 2016), vulnerabilities in information security (Pieters et al., 2012) and dominance hierarchies within animal colonies (Pörschmann et al., 2010). In this paper, we re-express the Elo model as a special case of the general approach to deriving rating models based on the gradient descent algorithm.

More recently, iterative update schemes for rating models were proposed in a paper by Koopman & Lit (2019), where, in the domain of econometric time series modelling, they are referred to as *score-driven models* (Creal et al., 2013). These models define a rating update equation as an autoregressive process with an extension to account for the derivative of the predictive likelihood with respect to the ratings modelled. While these ideas are similar, we propose to directly use an existing and well-founded optimisation heuristic to minimise some predefined loss functions.

Another example of an iterative rating scheme is the *pi-rating* model proposed in (Constantinou & Fenton, 2013; Constantinou et al., 2012) for football. This model decomposes a team's strength into home and away ratings and provides the rating update equations after each match as well. Its formulation is slightly more involved than that of the Elo rating model. The prediction function used in this model is based on a non-parametric estimate of the observed results' frequency based on the discrepancy of the ratings between the competing teams. On the other hand, the Elo model and other models studied in this paper offer a simple way to generate predictions, which provides a strong advantage especially for applying the model in practical settings.

In a different context, Moulton (2014) used a version of the gradient descent algorithm (Adagrad – an extension of the basic gradient descent by Duchi et al., 2011) to build a rating system for an online multiplayer video game. This approach can also be viewed as fitting the general framework of using the gradient descent algorithm for optimising the log-likelihood function. However, the exact problem setting is slightly different as the goal is to rate individual players

based on their team's score, just as the TrueSkill rating system does (Herbrich et al., 2006).

This paper is structured as follows. In Section 2 we introduce the iterative rating models (including the theoretical underpinnings of the Elo model). In Section 3 we describe a computational experiment for evaluating their quality. The final section summarises the results and concludes the paper. The implementation of all the proposed rating systems and all the steps required to reproduce the results are available online at `https://github.com/janekl/iterative-rating-systems`.

2. Iterative rating systems

2.1. The Elo rating system

The Elo rating system (Elo, 1961, 1978) is amongst the most popular methods for rating teams as well as players. There exist a few implementations of the Elo system for different sports that include discipline-specific tweaks, which we will discuss in greater detail below. We begin, however, by describing the Elo model in the most basic setting, along with the notation used throughout the paper.

Let $r_i^{(k-1)}$, $i = 1, 2, \ldots, n$, denote the ratings for a set of n teams after a total of $k-1$ matches have been played. The ratings are initialised with some default value.[1] If a team does not play in match k, its rating does not change, $r_i^{(k)} = r_i^{(k-1)}$. Otherwise, the ratings of the teams competing against each other are updated in an iterative manner after the match as follows.

Assume that we wish to generate the ratings after match k that happens between a home team i and an away team j.[2] The Elo rating system provides a simple update rule for the teams' ratings. First, the model forecasts the possible match outcome based on the current team ratings $r_i^{(k-1)}$ and $r_j^{(k-1)}$ that are up to date for match $k-1$. This forecast is formulated in terms of the probability of the home team winning the match. It is computed using the logistic function applied on the difference between the current team rating estimates

$$\mathbb{P}(O_{ij}^{(k)} = 1) = \frac{1}{1 + \exp\left(-r_i^{(k-1)} + r_j^{(k-1)}\right)}, \tag{1}$$

where $O_{ij}^{(k)}$ is a random variable standing for match outcome. In the following discussion of the Elo model we denote this probability as $p_{ij}^{(k)}$. Once the actual match result $o_{ij}^{(k)} \in \{0, 0.5, 1\}$ is available, where 1 denotes a win for team i, 0.5 – a draw and 0 – a victory for team j, the ratings are revised according to

[1] Typically $r_i^{(0)} = 1500$ for each i in the versions of the Elo model. However, the default value can be chosen arbitrarily as only the relative differences between the ratings actually matter.

[2] The distinction between the home and away teams is important in sport as a host of a match typically has an edge over the visitors (see, e.g., Neave & Wolfson (2003) or Swartz & Arce (2014) for an analysis of this effect). In an iterative model, this can be accounted for, e.g., by adding a constant value $h > 0$ to the home team rating (Sismanis, 2010).

the following rules

$$\begin{aligned} r_i^{(k)} &= r_i^{(k-1)} + K \cdot \left(o_{ij}^{(k)} - p_{ij}^{(k)} \right), \\ r_j^{(k)} &= r_j^{(k-1)} - K \cdot \left(o_{ij}^{(k)} - p_{ij}^{(k)} \right), \end{aligned} \tag{2}$$

where K is a scaling constant (and referred to as the *K-factor*). In particular, the magnitude of the updates is the same for both teams in absolute terms.

Interpretation. One of the most appealing features of the Elo rating system is that it offers an intuitive and plausible interpretation of the update rules. Namely, if the model's probability estimate $p_{ij}^{(k)}$ is lower than the actual match result $o_{ij}^{(k)}$ for team i, then the ratings are adjusted upward. The team performed better than expected, so its rating should be increased. Moreover, the higher the discrepancy between the actual result of the game and the predicted result, the greater the rating increase will be. An analogous effect is observed when team i falls short of expectations. If the actual result observed is lower than the one predicted, then its rating is decreased accordingly. This works analogously for team j.

Elo model as the gradient descent method for optimising a log-likelihood. It turns out the Elo rating system can be formulated as a special case of the stochastic gradient descent algorithm. Although this has been observed elsewhere (e.g., Pelánek 2016), in our view, it has not been recognised as widely as it should be, given the ubiquity of applications and deployments of the Elo model. We shall therefore introduce the derivations here in some detail. They will also serve as a basis for the introduction of the new models further on.

Generally, for a given set of matches $\mathcal{M}$, the goal is to estimate the individual team ratings $\mathbf{r} = (r_1, r_2, \ldots, r_n)$. For now, the superscript denoting the match index is not used. We assume that the ratings are static and are to be estimated jointly for the entire sample of matches $\mathcal{M}$. Yet, we shall see that by applying the gradient descent iterations according to the order in which the matches take place, the update equations (2) will emerge naturally and the ratings will become dynamically revised after each consecutive match.

The probability of the outcome of a match k played between teams i and j is modelled using the logistic function, just like in the original formulation:

$$p_{ij}^{(k)} = \frac{1}{1 + \exp(-r_i + r_j)}. \tag{3}$$

The negative log-likelihood of the observed results is defined as:

$$L(\mathbf{r}|\mathcal{M}) = -\sum_{(i,j,k)\in\mathcal{M}} \left(o_{ij}^{(k)} \log p_{ij}^{(k)} + \left(1 - o_{ij}^{(k)}\right) \log(1 - p_{ij}^{(k)}) \right), \tag{4}$$

Note that the above is nothing more than the case where match results are expressed by means of the binary – win or loss – logistic regression outcome model where the team ratings $\mathbf{r} = (r_1, r_2, \ldots, r_n)$ are the parameters to be estimated. Possible ties are not explicitly handled (this will be discussed in greater detail below). As such, we are describing a convex optimisation problem (Boyd &

Vandenberghe, 2004) and we may use the gradient descent (amongst others) to effectively minimise such an objective function to find the desired ratings.

In order to apply the gradient descent scheme so as to optimise the loss function specified, we need to compute the derivatives with respect to the rating parameters. Note that the objective function can be decomposed as the sum of losses for individual matches,

$$L(\mathbf{r}|\mathcal{M}) = \sum_{(i,j,k)\in\mathcal{M}} l_{ij}^{(k)}(\mathbf{r}).$$

Taking the partial derivative with respect to r_i of a single term yields

$$\frac{\partial l_{ij}^{(k)}}{\partial r_i} = p_{ij}^{(k)} - o_{ij}^{(k)} \tag{5}$$

and analogously for r_j. The stochastic gradient descent algorithm operates iteratively to find a local minimum of a function. In each step, a coordinate-wise move in the counter-gradient direction (steepest descent) is performed. Let us assume that the current estimate of team i's strength from the previous iteration of the algorithm is $r_i^{(k-1)}$. Then the update rule for the rating parameters is

$$r_i^{(k)} = r_i^{(k-1)} - \gamma \cdot \frac{\partial l_{ij}^{(k)}}{\partial r_i} = r_i^{(k-1)} - \gamma \cdot \left(p_{ij}^{(k)} - o_{ij}^{(k)}\right), \tag{6}$$

where $\gamma > 0$ is the *learning rate* in the gradient descent algorithm and the predicted match result $p_{ij}^{(k)}$ in Eq. (3) is evaluated using the previous rating values $r_i^{(k-1)}$ and $r_j^{(k-1)}$, just as Eq. (1) does. If we choose the iterations of the algorithm to correspond with the consecutive matches as they appear over time, we will get the rating updates in the Elo rating system given by Eq. (2) with $K = \gamma$. Thus, the Elo model has a theoretical background as a single scan over the dataset of matches in the order given by the dates they appear.

Presenting the Elo rating rating system in this setting also helps one to recognise more clearly how the possible ties are handled. Namely, from the log-likelihood function optimised by the Elo model given by Eq. (4), we conclude that a draw is considered a "half-win, half-loss". More precisely, a draw enters the likelihood function for the results observed as the geometric mean of team i's win and loss probabilities. This can be seen by plugging in $o_{ij}^{(k)} = \frac{1}{2}$ and exponentiating a single component in Eq. (4):

$$\exp\left(\frac{1}{2}\log p_{ij}^{(k)} + \frac{1}{2}\log(1 - p_{ij}^{(k)})\right) = \sqrt{p_{ij}^{(k)} \cdot \left(1 - p_{ij}^{(k)}\right)}. \tag{7}$$

This convention was later proposed by Glickman (1999) in the context of sport rating models. As demonstrated, it is an implicit assumption behind the Elo rating system itself.

Implementing the Elo model for football. An example adaptation of the Elo model for football was proposed by Hvattum & Arntzen (2010). The authors propose the following tweaks. First, to account for the fact that the Elo rating system does not provide the probability of a draw, we can rely on the differences

in ratings as covariates to an ordinal logistic regression model, which is used as a second-level model once the rating differences are computed.

Second, there are some differences in the way the K-factor (learning rate) is formulated. The authors suggest that the K-factor could be amplified by the difference in the goals scored by the teams

$$K = K(i, j, k) = K_0 \cdot (1 + |g_i^{(k)} - g_j^{(k)}|)^{\lambda_g}, \tag{8}$$

where $|g_i^{(k)} - g_j^{(k)}|$ denotes the absolute goal difference in a match k between teams i and j, where $K_0 > 0$ and $\lambda_g > 0$ are additional parameters. Strictly speaking, this modification is a heuristic approach for adapting the learning rate discussed above. The larger the goal difference, the greater the update to the current team ratings.

Finally, setting the initial values for the ratings $r_i^{(0)}$ corresponds to choosing prior ratings in the Elo model. The prediction and update equations remain the same as in the original formulation – in fact, the authors divided the difference in ratings by 400 and used exponentiation with base 10 in Eq. (1) but it is just a question of scaling the ratings.

We shall later compare this model to other approaches in the computational experiments section. Other implementations of the Elo model for football include, e.g., (EloRatings.net, 2020) or both official FIFA men and women world ranking implementations (FIFA.com, 2020).

2.2. Ordinal logistic regression model

We shall now turn to the second team rating system, focusing on the ordinal logistic regression (Aitchison & Silvey, 1957; Koning, 2000). We first formulate the base version of the model and next discuss how the model's parameters can be obtained by an application of the gradient descent method.

As before, let $\mathbf{r} = (r_1, r_2, \ldots, r_n)$ denote the ratings for a set of n teams. Again, the superscript to denote the current rating estimates for a given match will appear as a by-product of the gradient descent optimisation steps. Denote with

$$\Delta_{ij} = r_i - r_j + h$$

the difference in the ratings of a home team i and a team j, corrected for the *home team advantage* parameter $h > 0$. In the general setting, we assume that there is a latent variable ε_{ij} functioning as a random variable that follows the logistic distribution with mean zero and scale one. It yields a noisy version of the true rating difference $\Delta_{ij}^* = \Delta_{ij} + \varepsilon_{ij}$ which we observe in reality up to a certain interval.

The idea behind the model is that we test whether Δ_{ij}^* falls into a specified interval, corresponding to the match outcome in the following way

$$O_{ij}^{(k)} = \begin{cases} 1 & \text{for } \Delta_{ij}^* \in (c, \infty), \\ 0.5 & \text{for } \Delta_{ij}^* \in [-c, c], \\ 0 & \text{for } \Delta_{ij}^* \in (-\infty, -c), \end{cases} \tag{9}$$

where $c > 0$ is an intercept governing the draw margin. The probabilities of the

match outcomes are given by

$$\begin{aligned}
\mathbb{P}(O_{ij}^{(k)} = 1) &= 1 - \frac{1}{1 + \exp(-c + \Delta_{ij})}, \\
\mathbb{P}(O_{ij}^{(k)} = 0.5) &= \frac{1}{1 + \exp(-c + \Delta_{ij})} - \frac{1}{1 + \exp(c + \Delta_{ij})}, \\
\mathbb{P}(O_{ij}^{(k)} = 0) &= \frac{1}{1 + \exp(c + \Delta_{ij})},
\end{aligned} \tag{10}$$

The parameter estimation procedure is again based on the maximum likelihood principle. Given the outcome model, we can construct a loss function $L(\mathbf{r}, h, c|\mathcal{M}, \lambda)$, defined as the negative penalised log-likelihood for the match results observed (Tutz & Gertheiss, 2016). The penalty introduced is L_2 regularisation on the team rating parameters:

$$L(\mathbf{r}, h, c|\mathcal{M}, \lambda) = -\sum_{(i,j,k)\in\mathcal{M}} \log \mathbb{P}\left(O_{ij}^{(k)} = o_{ij}^{(k)}|\mathbf{r}, h, c\right) + \frac{\lambda}{2} \cdot \|\mathbf{r}\|_2^2, \tag{11}$$

where $\mathbb{P}\left(O_{ij}^{(k)} = o_{ij}^{(k)}|\mathbf{r}, h, c\right)$ denotes the probability attributed by the model to the actual result of a match. The use of regularisation for match outcome prediction models is generally advised because the uncertainty factor is relatively large in this domain. It usually helps to provide more accurate forecasts (Groll et al., 2015; Lasek & Gagolewski, 2018).

The parameters of the model are found by minimising the above function or, equivalently, maximising the penalised log-likelihood of the results with respect to $\mathbf{r}$, h, and c. The choice of regularisation parameter λ is discussed later. In addition to preventing overfitting, regularisation ensures that the parameters are identifiable. According to the model formulation given by Eq. (10), any shift in the rating parameters by a constant yields the same probabilities (as their differences remain equal).

Finally, it is worth noting how this model is connected to the Elo rating system. If $c = 0$ in Eq. (10), the probability of a draw becomes zero. This brings us back to the binary logistic regression considered in the previous section.

Rating updates. Let us now derive a method to iteratively estimate the model parameters using the gradient descent approach. To obtain the update equations, the partial derivatives of L with respect to the model parameters are needed. We shall focus only on optimising the ratings, considering parameters c and h as fixed. These parameters can be estimated by grid or random search as discussed in Section 3. To provide the iterative updates for the likelihood function given in Eq. (11), it is useful to first obtain the derivatives of the logarithm of probability functions (Eq. 10):

$$\begin{aligned}
\frac{\partial \log \mathbb{P}(O_{ij}^{(k)} = 1)}{\partial r_i} &= 1 - \mathbb{P}(O_{ij}^{(k)} = 1), \\
\frac{\partial \log \mathbb{P}(O_{ij}^{(k)} = 0.5)}{\partial r_i} &= \mathbb{P}(O_{ij}^{(k)} = 0) - \mathbb{P}(O_{ij}^{(k)} = 1), \\
\frac{\partial \log \mathbb{P}(O_{ij}^{(k)} = 0)}{\partial r_i} &= \mathbb{P}(O_{ij}^{(k)} = 0) - 1,
\end{aligned} \tag{12}$$

and symmetrically for the derivatives for the other team j. These derivatives have a simple closed form, unlike for the related model in the case of the probit link function in (Koopman & Lit, 2019).

Next, we introduce the update equations for the model. As in the presentation of the Elo model above, a single step of the gradient descent algorithm is performed for an individual match and all the steps are in line with the order in which matches are played over time. Additionally, we add the L_2 regularisation component to each update. Hence, the update equations take the form

$$r_i^{(k)} = \begin{cases} r_i^{(k-1)} - \gamma \cdot \left(\mathbb{P}\left(O_{ij}^{(k)} = 1\right) - 1 + \lambda r_i^{(k-1)} \right) & \text{for } o_{ij}^{(k)} = 1, \\ r_i^{(k-1)} - \gamma \cdot \left(\mathbb{P}\left(O_{ij}^{(k)} = 1\right) - \mathbb{P}\left(O_{ij}^{(k)} = 0\right) + \lambda r_i^{(k-1)} \right) & \text{for } o_{ij}^{(k)} = 0.5, \\ r_i^{(k-1)} - \gamma \cdot \left(1 - \mathbb{P}\left(O_{ij}^{(k)} = 0\right) + \lambda r_i^{(k-1)} \right) & \text{for } o_{ij}^{(k)} = 0. \end{cases} \tag{13}$$

and symmetrically for team j. We note that the match outcome probabilities, denoted with $\mathbb{P}\left(O_{ij}^{(k)} = 1\right)$ and $\mathbb{P}\left(O_{ij}^{(k)} = 0\right)$ above, are evaluated using the previous rating estimates $r_i^{(k-1)}$.

Interpretation. The update equations can be interpreted intuitively as follows. For simplicity, we assume $\lambda = 0$. First, if a home team win is observed, the home team rating update is proportional to the confidence in this event measured by the difference $1 - \mathbb{P}\left(O_{ij}^{(k)} = 1\right)$. The more unexpected this result is, the greater the decrease in team i's rating. The opposite holds if the away team wins. On the other hand, in the case of a draw, both the home and away team ratings are changed so that the rating of the team expected to win based on the expression $\mathbb{P}\left(O_{ij}^{(k)} = 1\right) - \mathbb{P}\left(O_{ij}^{(k)} = 0\right)$ is decreased and the rating of the other team is increased. Finally, the regularisation component equal to $\lambda r_i^{(k-1)}$ acts as the rating decay by shrinking it, regardless of the actual match result.

2.3. Poisson regression model

Another popular model is based on the assumption that the number of goals scored by each team in a match is a Poisson-distributed random variable. In the basic setup, Maher (1982) suggests modelling of the goals scored under the independence assumption. This was one of the first approaches specifically crafted for football and it serves as a basis for more complex models including (Crowder et al. 2002; Dixon & Coles 1997; Groll et al. 2015; Karlis & Ntzoufras 2003; Kharrat 2016; Koopman & Lit 2015; Rue & Salvesen 2000). The Maher model in introduced in greater detail below.

Assume that in match k of a competition team i (home) plays against team j (away). Let $G_i^{(k)}$ and $G_j^{(k)}$ be random variables for the goals scored by team i and j, respectively. Assuming that these random variables are independent and that they follow the Poisson distributions with expected values of μ_i and μ_j, respectively, the probability of the match outcome being x-y is:

$$\mathbb{P}\left(G_i^{(k)} = x, G_j^{(k)} = y | \mu_i, \mu_j\right) = \frac{\mu_i^x}{x!} \exp(-\mu_i) \cdot \frac{\mu_j^y}{y!} \exp(-\mu_j).$$

If we assume a log-linear model for the underlying parameters (i.e., goal scoring rates), then we can write

$$\begin{aligned}\log(\mu_i) &= c + h + a_i - d_j,\\ \log(\mu_j) &= c + a_j - d_i,\end{aligned}$$

where c is an intercept and a_i, a_j and d_i, d_j stand for attack and defence ratings of teams i and j, respectively. Hence, as opposed to the previously discussed settings, this model describes the strength of each team using two parameters. Parameter h is introduced to capture the home field advantage.

Again, the model parameters can be estimated by maximising the likelihood. Let $\mathbf{r} = (\mathbf{a}, \mathbf{d}) = (a_1, a_2, \ldots, a_n, d_1, d_2, \ldots, d_n)$ be the team strength parameters and denote with $L(\mathbf{r}, h, c|\mathcal{M}, \lambda)$ the loss function, which is the negative penalised log-likelihood of the results observed in dataset $\mathcal{M}$:

$$\begin{aligned}L(\mathbf{r}, h, c|\mathcal{M}, \lambda) = &- \sum_{(i,j,k)\in\mathcal{M}} \Bigg(\log \mathbb{P}\left(G_i^{(k)} = g_i^{(k)} | \mathbf{r}, h, c\right) \\ &+ \log \mathbb{P}\left(G_j^{(k)} = g_j^{(k)} | \mathbf{r}, h, c\right) \Bigg) \\ &+ \lambda \cdot \left(\frac{\|\mathbf{r}\|_2^2}{2} - \rho \cdot \mathbf{a}^\top \mathbf{d} \right),\end{aligned} \tag{14}$$

where the observed match outcome between teams i and j in the k-th round is $g_i^{(k)}$ to $g_j^{(k)}$ (in terms of goals scored). The third term is the extension proposed in (Lasek & Gagolewski, 2018) with $\rho \in [-1, 1]$ being a correlation between the attack and defence ratings. We note that the regularisation term also enables the identification of the model parameters here.

Rating updates. For the Poisson model we can adopt analogous ideas so as to arrive at the iterative version of the model that constitutes a standalone rating system. To obtain the update equations, we consider the parameters c, h and λ as fixed, focusing on the attack and defence ratings as the parameters over which we optimise the likelihood.

In a single step of the gradient descent algorithm for optimising the loss function given by Eq. (14), for single match at round k between the teams i and j, ending in score $g_i^{(k)}$ to $g_j^{(k)}$, assuming that the current team strength estimates are $a_i^{(k-1)}, d_i^{(k-1)}$ and $a_j^{(k-1)}, d_j^{(k-1)}$, we arrive at the following update equations

$$\begin{aligned}a_i^{(k)} &= a_i^{(k-1)} - \gamma \cdot \left[\left(\mu_i^{(k)} - g_i^{(k)} \right) + \lambda \cdot \left(a_i^{(k-1)} - \rho d_i^{(k-1)} \right) \right],\\ d_i^{(k)} &= d_i^{(k-1)} - \gamma \cdot \left[\left(g_j^{(k)} - \mu_j^{(k)} \right) + \lambda \cdot \left(d_i^{(k-1)} - \rho a_i^{(k-1)} \right) \right].\end{aligned} \tag{15}$$

where $\log(\mu_i^{(k)}) = c + h + a_i^{(k-1)} - d_j^{(k-1)}$ and $\log(\mu_j^{(k)}) = c + a_j^{(k-1)} - d_i^{(k-1)}$.

Interpretation. In this case as well, the rating update equations allow for an intuitive interpretation. Recall that $\mu_i^{(k)}$ and $\mu_j^{(k)}$ are the pre-match average goal

scoring rates – the means of the corresponding Poisson variables computed for the rating estimates in the previous iteration $k-1$. Now, for $\lambda = 0$, if team i scored more goals in match k than would be expected based on parameter $\mu_i^{(k)}$ – which also depends on the opponent's defence rating – then its attack rating increases accordingly. Analogously, if this team concedes fewer goals than expected from the overall goal scoring rate of its opponent, $\mu_j^{(k)}$, then its defence rating increases. Moreover, the greater the differences, the larger the updates. Finally, the regularisation component in the update equations push the attack and defence ratings towards zero. Correlation ρ causes the ratings to remain close to each other. Intuitively, strong teams tend to feature both a strong attack and solid defence, while the converse is true for weak teams.

From goals to the final score. As the approach introduced takes into account the number of goals the teams score, we can convert it to a full-time three-way outcome by writing

$$\begin{aligned} \mathbb{P}(O_{ij}^{(k)} = 1) &= \mathbb{P}(Z_{ij}^{(k)} > 0) = 1 - \mathbb{F}_{ij}^{(k)}(0), \\ \mathbb{P}(O_{ij}^{(k)} = 0.5) &= \mathbb{P}(Z_{ij}^{(k)} = 0) = \mathbb{F}_{ij}^{(k)}(0) - \mathbb{F}_{ij}^{(k)}(-1), \\ \mathbb{P}(O_{ij}^{(k)} = 0) &= \mathbb{P}(Z_{ij}^{(k)} < 0) = \mathbb{F}_{ij}^{(k)}(-1) \end{aligned} \tag{16}$$

for the random variable $Z_{ij}^{(k)} = G_i^{(k)} - G_j^{(k)}$, which follows a Skellam distribution with parameters (μ_i, μ_j) and $\mathbb{F}_{ij}^{(k)}$ as its cumulative distribution function. While this function does not lead to as simple an analytical form as that of the Elo or the ordinal logistic regression models, it can be easily computed numerically.

2.4. One-parameter Poisson regression model

Maher (1982) studied various simplifications or extensions of the Poisson model. However, the two-parameter version is the more frequently of those cited in the literature. The last model studied is a version of the above where the attack and defence strengths are reduced to a single parameter, just as in (Ley et al., 2019). As we demonstrate in the next section, it is not only elegantly simple, but also yields quite accurate forecasts.

The parameters of the corresponding Poisson variables in that model are assumed to have the following form

$$\begin{aligned} \log(\mu_i) &= c + h + r_i - r_j, \\ \log(\mu_j) &= c + r_j - r_i. \end{aligned} \tag{17}$$

Here we assume that the attack and defence strengths of each team are equal, $a_i = d_i$. The objective function is defined analogously as in the previous model (Eq. 14). The prediction function remains the same as described in Eq. (16).

We note that this model is, in a way, a marginal case of the correlated Poisson regression model presented above for highly correlated attack and defence strengths. Rewriting the penalty term in Eq. (14) with $\rho = 1$ yields

$$\lambda \cdot \left(\frac{\|\mathbf{r}\|_2^2}{2} - \rho \cdot \mathbf{a}^\top \mathbf{d} \right) = \frac{\lambda}{2} \sum_{i=1}^{n} (a_i - d_i)^2.$$

If the regularisation parameter λ is set to some large-enough value, then setting $a_i = d_i$ cancels the penalty out and the one-parameter model is obtained. However, this is a rather theoretical argument. In practice, setting a large λ and $\rho = 1$ results in convergence problems due to numerical stability issues.

Rating updates. Applying the gradient descent approach, the rating update equation becomes

$$r_i^{(k)} = r_i^{(k-1)} - \gamma \cdot \left[\left(\mu_i^{(k)} - \mu_j^{(k)} \right) - \left(g_i^{(k)} - g_j^{(k)} \right) + \lambda r_i^{(k-1)} \right], \qquad (18)$$

and analogously for team j.

Interpretation. This model can also be interpreted intuitively, based on the expected (pre-match) and observed win margin, $\mu_i^{(k)} - \mu_j^{(k)}$ and $g_i^{(k)} - g_j^{(k)}$, respectively. Assuming $\lambda = 0$, if the expected margin exceeds the observed margin, team i's rating is adjusted downwards. We note that analogous margin-based models inspired by the Elo rating system can be found in the literature, though with scant theoretical foundations behind them (see, e.g., Carbone et al., 2016 or Kovalchik, 2020). This formulation is again a natural consequence of applying the gradient descent algorithm to estimate a model's parameters.

3. Experiments

In this section we consider computational experiments for validating the efficacy of the different models introduced above.

3.1. Data and validation procedure

To evaluate the models quantitatively, we estimate them and generate predictions for top-level divisions for the five strongest football leagues in the world: English, French, German, Italian, and Spanish. The data are available for download at `http://www.football-data.co.uk/`. Following the well-established methodology, see, e.g., (Barrow et al., 2013; Boshnakov et al., 2017; Lasek et al., 2013; Ley et al., 2019), we employ a sliding window procedure for generating the predictions for consecutive match days. More precisely, each prediction is generated using the whole set of preceding matches up to a given date. Once the predictions are obtained for a given set of matches, the training set is extended to keep the ratings up-to-date. For the iterative model versions, this boils down to updating ratings once the new match results are observed.

We follow a standard protocol of training/validation/test data split. The seasons 2009/10–2011/12 are used only as training data for the rating systems and provide an initial sample to build a model. The three next seasons 2012/13–2014/15 are used as a validation set to choose the models' optimal parameter values. Finally, the last four seasons are used as the test set to measure the models' performance – from 2015/16 to 2018/19. This makes up a total of 7304 matches across the five leagues in the test set for evaluation.

3.2. Model setup

The Elo model parameters are again set to their default values from the original formulation (Hvattum & Arntzen, 2010). We only focus on the goal-based version of the model which was shown by the authors to perform best. We denote it with Elo_g.

Further, as for the baseline results, we shall provide the predictions for the ordinal logistic (OLR) and Poisson regression-based models. We also introduce sample weights in the likelihood functions in Eq. (11) and Eq. (14) for the models that estimate ratings using the whole sample of matches. More precisely, the samples (matches) are weighted according to how long ago a given match was played. To this end, we use exponential weighting $w_k = \exp(-b \cdot \tau_k)$, where τ_k stands for how many days ago match k was played (relative to ratings estimation date) and $b \geq 0$ is a decay parameter. This is a standard method applied to account for the recency of matches (Boshnakov et al., 2017; Dixon & Coles, 1997; Koopman & Lit, 2019; Ley et al., 2019). While there exist a variety of methods for accounting for time in prediction models – including autoregressive modelling of team strengths (Crowder et al., 2002), rating teams based on exponential weighted moving average processes (Cattelan et al., 2013), and time varying team strength models based on interpolation (Baker & McHale) – we use exponential weighting here as it is well-founded and widely applied in the recent literature.

The optimal parameter values are determined using the grid search. More precisely, possible values for parameter b are between 0 and 0.006 with a step of 0.001 and for regularisation from 0 to 15 with a step of 0.25. Additionally, correlation parameter for ρ between 0 and 0.95 with a step of 0.05 is considered (the case of 0.99 is also studied). The models jointly optimising the likelihood function for the whole sample of matches are estimated using the BFGS algorithm (Nocedal & Wright, 2006).

As for the iterative models, different approaches require different parameters to be specified, including c, h, λ and, specifically for the attack-defence version of the Poisson model, correlation parameter ρ. Moreover, all the models require the learning rate to be set. In the case of the Poisson regression-based models, c is searched for among values 0, 0.0001, 0.0002, 0.001, 0.002, 0.004, 0.01, 0.02, 0.1, and 0.2. Parameter h is searched for from 0.2 to 0.4 with a step of 0.05. Possible values of regularisation λ are 0.00001, 0.00002, 0.00005, 0.0001, 0.0002, 0.0005, 0.001, 0.002, and 0.01. As for correlation ρ, its values are searched for from 0 to 1 with a step of 0.05. In the case of iterative OLR model, h is searched for from 0.2 to 0.5 with a step 0.05, c from 0.4 to 0.7 with a step of 0.05 and learning rate γ is chosen among 0.001, 0.005, 0.01, 0.02, 0.04, 0.06, 0.08, and 0.1. Finally, the regularisation parameter is selected from among 0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, and 0.02.

These parameter combinations produce a large grid of parameter settings. To keep the computation time at a reasonable level, we randomly choose only 250 parameter settings sampled without replacement from all the combinations from the parameter grid determined by their possible values. Parameters are optimised globally for all the leagues.

To assign the rating parameters initial values, a simple strategy is employed: assign the initial value of the rating parameters across all teams to zero. For the Elo model, this value can be chosen arbitrarily, as only the relative differences in

the ratings matter. For the other iterative approaches, zero is a natural choice that allows the impact of any regularisation to be diminished in the initial estimation phase. Finally, all league newcomers are also assigned the rating of zero. For the other methods, we assigned ratings equal to the averages across all the teams. In practice, zero is close to the ratings average since all iterative model versions yield the updates – ignoring the regularisation component – that sum up to zero.

To optimise a model's parameters, a single target (validation) metric is needed. For this purpose we choose the logarithmic loss described in the next section.

3.3. Evaluation metrics

The prediction results presented in the next section are evaluated according to logarithmic loss, ranked probability score, Brier score and accuracy. These metrics are popular for evaluating a model's predictive power (Constantinou & Fenton, 2012; Goddard, 2005; Hvattum & Arntzen, 2010; Koopman & Lit, 2019; Ley et al., 2019; Peeters, 2018). We now recall how they are defined. Let $\mathbf{p} = (p_1, p_2, p_3)$ denote the three-way match outcome probabilities obtained from a given model with $p_1 + p_2 + p_3 = 1$, $p_i \geq 0$. Further, let $\mathbf{q} = (q_1, q_2, q_3)$ denote the vector indicating the true outcome of a match with $q_1 + q_2 + q_3 = 1$, $q_i \in \{0, 1\}$. For example, if the home team wins a match, $q_1 = 1$ and $q_2 = q_3 = 0$. For simplicity, the metrics are described for a single match. These metrics are computed and averaged to obtain aggregate performance over a dataset of matches.

Logarithmic loss. Logarithmic loss, or *logloss* for short, is computed as

$$-\sum_{i=1}^{3} q_i \cdot \log(p_i). \tag{19}$$

In the maximum likelihood setting, logloss (subject to regularisation) is the criterion directly optimised by many prediction models – including the ordinal logistic regression considered here. We choose this metric as the target one when optimising a model's parameters.

Ranked probability score. The ranked probability score, or *RPS* for short, is computed as

$$\frac{1}{2}\sum_{i=1}^{2}\left(\sum_{j=1}^{i} p_j - \sum_{j=1}^{i} q_j\right)^2. \tag{20}$$

Again, the normalisation accounts for the number of outcomes (minus one). Unlike the other metrics, this one accounts for the ordinal nature of the results.

Brier score. Brier score, or *quadratic loss*, is computed with the following formula

$$\frac{1}{3}\sum_{i=1}^{3}(p_i - q_i)^2. \tag{21}$$

Accuracy. This metric is defined as the proportion of correctly predicted results. In the case of a single match it equals

$$\mathbb{1}\left(\arg\max(\mathbf{p}) = \arg\max(\mathbf{q})\right). \tag{22}$$

Here, the arg max function for a vector returns the index with the highest value among all coordinates of the vector. However, as it is not a strictly proper metric (Gneiting & Raftery, 2007), it is included in the comparison only for the intuitive interpretation it offers.

To provide a context for the metrics considered, we also report the results of two extra baselines: the observed results' frequency in a particular league and the predictions derived from bookmaker odds in a decimal format by inverting and normalising them. Analogously to the previous studies for comparing the differences in the forecasting ability of the match prediction models (Hvattum & Arntzen, 2010; Peeters, 2018), the significance of the differences between the methods is determined by using the paired t-test with a significance level of 0.05. Due to the special role of logloss in experimental setup as a validation criterion, the tests are applied and p-values are reported for this metric while the other metrics are used for extra evaluation and to provide background.

3.4. Results

Table 1 presents the aggregate statistics for the test season predictions for all the metrics considered. The base models are denoted with OLR for the ordinal logistic regression, PR_1 and PR_2 for one- and two-parameter Poisson regression-based models, respectively. Their iterative versions are denoted with superscript I. By analysing the pairwise differences in the predictions, we found that the most accurate predictions are obtained by the iterative version of the one-parameter Poisson model – PR_1^I. Applying the t-test for the differences in logloss values for $\mathrm{Elo_g}$ and PR_1^I models we conclude that the latter is significantly more accurate in predicting match results (p-value = 0.036).

Table 1: Aggregate prediction statistics for the test seasons. The iterative models outperform their static variants.

Model	**Section**	**Logloss**	**RPS**	**Brier**	**Accuracy**
$\mathrm{Elo_g}$	Sec. 2.1	0.9726	0.1973	0.1928	0.5307
OLR	Sec. 2.2	0.9759	0.1983	0.1936	0.5272
OLR^I	Sec. 2.2	0.9739	0.1978	0.1932	0.5297
PR_1	Sec. 2.4	0.9763	0.1983	0.1936	0.5334
PR_1^I	Sec. 2.4	**0.9708**	**0.1968**	**0.1925**	**0.5352**
PR_2	Sec. 2.3	0.9764	0.1983	0.1936	0.5326
PR_2^I	Sec. 2.3	0.9721	0.1969	0.1928	0.5344
Betting odds	–	**0.9568**	**0.1924**	**0.1893**	**0.5427**
Class frequency	–	1.0631	0.2278	0.2141	0.4577

Interestingly, the iterative versions of the regression-based rating models – OLR^I, PR_1^I, and PR_2^I – perform better than their respective counterparts –

OLR, PR_1, and PR_2 estimated using the BFGS algorithm on a whole sample of matches, even when the result weighting is applied. In the case of the Poisson regression-based approaches, we do observe a statistically significant improvement (p-values < 0.001 and 0.004 for the one- and two-parameter model versions, respectively). For the OLR models, the difference is however not significant ($p = 0.071$).

Finally, the iterative Poisson regression-based models turn out to be more accurate than both the Elo- and the ordinal logistic regression-based approaches. We conclude that goal-based modelling provides better predictions than that based solely on the three-way match outcome. This confirms the results from previous studies on this issue (see, e.g., Hvattum & Arntzen, 2010; Koopman & Lit, 2019; Ley et al., 2019).

The prediction quality in terms of logloss for individual leagues is presented in Table 2. The predictions based on the iterative Poisson models are unanimously more accurate for all the leagues. This additionally confirms the superior performance of these models.

Table 2: Model performance according to logloss for individual leagues for the test seasons.

league	$\mathbf{Elo_g}$	**OLR**	$\mathbf{OLR}^I$	$\mathbf{PR}_1$	$\mathbf{PR}_1^I$	$\mathbf{PR}_2$	$\mathbf{PR}_2^I$
England	0.9574	0.9633	0.9615	0.9644	**0.9545**	0.9642	0.9551
France	1.0003	1.0040	1.0013	1.0029	1.0000	1.0026	**0.9980**
Germany	0.9974	0.9988	0.9980	0.9966	**0.9952**	0.9970	0.9979
Italy	0.9455	0.9496	0.9475	0.9541	**0.9446**	0.9550	0.9491
Spain	0.9674	0.9682	0.9659	0.9672	**0.9645**	0.9674	0.9652

3.5. Optimal parameter values

We now focus on the technical details of the optimal parameter setting found by optimising the predictions based on logloss. The detailed parameter specification is presented in Table 3. In the case of the iterative Poisson-based models, parameters $c = 0.02$ and $h = 0.3$ mean that for equally rated teams the prediction function yields ca. $(0.45, 0.27, 0.28)$. In the case of the OLR^I model, for $c = 0.6$ and $h = 0.4$ we obtain $(0.45, 0.28, 0.27)$. This is in line with the result frequencies observed in real-world data.

The correlation parameter in the model regularisation component ρ equals 0.9 and 1.0 for the base two-parameter Poisson model and its iterative version, respectively. This is a relatively large value which partially explains why the model performance is close to the one-parameter variant as the attack and defence ratings are highly correlated. In a way, the model converges to its simpler version with no loss in accuracy. We argue that predicting match results is a difficult task and any form of model regularisation (e.g., by restricting the parameter space) proves useful in reducing the error metrics.

Finally, the optimised value for parameter b equals 0.002 and is the same for all the three base models that employ it. For these models, parameters c and h are not reported in Table 3 as they are recalculated for each day for which the predictions are to be generated. On average, they are close to the values reported for their iterative variants.

Table 3: Optimised parameter values for different models.

Model	c	h	γ	λ	b	ρ
$\mathbf{Elo_g}$	10	–	10	–	–	–
OLR	–	–	–	1.25	0.002	–
$\mathbf{OLR}^I$	0.6	0.4	0.08	0.0002	–	–
$\mathbf{PR}_1$	–	–	–	2.5	0.002	–
$\mathbf{PR}_1^I$	0.02	0.3	0.01	0.0001	–	–
$\mathbf{PR}_2$	–	–	–	13	0.002	0.9
$\mathbf{PR}_2^I$	0.02	0.3	0.02	0.0005	–	1.0

3.6. Iterative model updates with momentum

In the iterative model formulations we considered a basic version of the gradient descent algorithm. There are many extensions of this method including momentum updates that accumulate the gradients using exponential smoothing (see, e.g., Goh 2017). Intuitively, momentum results in higher rating updates for the teams in consecutive runs of wins and, conversely, larger decreases for the teams with a record of consecutive losses. This has a natural interpretation in sports.

In the follow-up experiments, we have extended the iterative model versions so that they are based on gradient descent with momentum. However, there are no significant improvements over the base models with no momentum parameter. Moreover, the optimised values of the momentum term weights were close to zero, so these model versions effectively reduce to the standard gradient descent updates studied here. We have therefore not included the detailed results here as the performance of the models is virtually identical.

These results are in line with the conclusions from other empirical studies. In particular, the question whether the runs of results might reveal anything of substance about the next game has been considered, for example, by Dobson & Goddard (2003) or Goddard (2006) who all concluded that there is a zero or even negative momentum effect. On the other hand, Heuer & Rubner (2009) found that teams on a losing streak are more likely to lose the next match. As for a winning streak, the authors observed either no effect or even a slight decrease in the probability of winning the next match.

3.7. Team ratings in time

It could be interesting to see how the team ratings evolve in time for the best performing model, PR_1^I. Figure 1 presents the ratings for the "Big Six" Premier League teams: Arsenal, Chelsea, Liverpool, Manchester City, Manchester United, and Tottenham.

We observe that the highest rated team at the end of the 2016/17 season was not Chelsea (the actual champions) but Tottenham (the runners-up). It is noteworthy that their rating improved substantially after two large-margin away wins at the end of the season – 6-1 and 7-1 – respectively, against Leicester and Hull. This shows some vulnerability of the model to outliers. However, clipping the goals scored (as in Rue & Salvesen 2000) at different values did not improve the predictions.

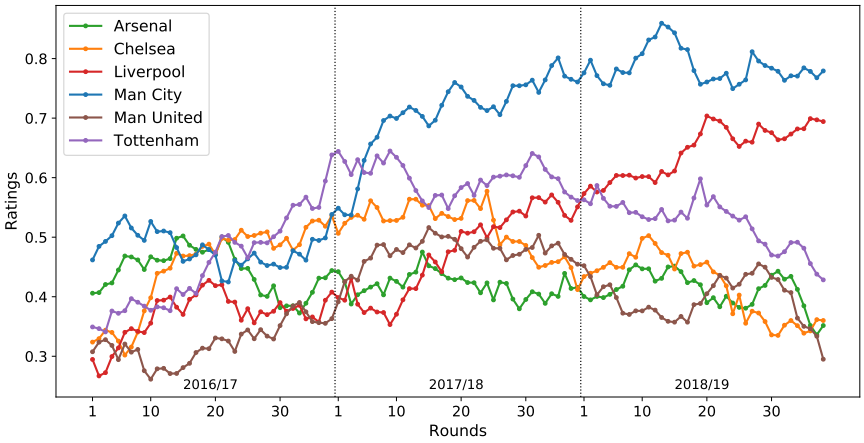

Figure 1: Ratings for selected English Premier League teams over the 2016/17–2018/19 seasons.

In the 2017/18 season, Manchester City dominated the league with an impressive 100 points tally. Their dominance is reflected in their unanimously superior rating at the end of this season. We also observe a steady growth in time of the strength scores of Liverpool and Manchester City while the ratings of other teams appear to decline from the middle of 2017/18 season. These two teams were also locked in very close ratings at the end of the 2018/19 season, which saw them vie for the title until the very last round. Ultimately, Manchester City took the title.

4. Conclusions

In this paper, the plausible features of the Elo model as a fast, analytically tractable and interpretable rating system have been successfully translated to other prominent rating systems. In recognising it as a special case of the gradient descent algorithm, we have discussed its theoretical underpinnings. The proposed ordinal logistic and Poisson regression-based rating systems are derived in an analogous way and also boast the appealing features that made the Elo model popular in various application scenarios in sports and beyond.

In the computational experiments presented here, the iterative rating systems based on the Poisson regression proved significantly more accurate. In particular, the iterative version of the one-parameter Poisson model (Ley et al., 2019) proposed here turns out to have more predictive power than the Elo rating system. Hence, this model rates teams more accurately. Interestingly, extensive parameter search allows us to conclude that the two-parameter Poisson model (which uses attack and defence ratings to describe a team) converges to its simpler version via high correlation ρ between the ratings. While the former approach has been extensively studied, we believe that these results should contribute to a greater popularity of the latter model as it turns out to be more accurate. In the case of the ordinal logistic regression, the iterative model version provides better forecasting accuracy, albeit not significantly.

The proposed framework based on the (stochastic) gradient descent algorithm for maximising the log-likelihood function is a simple and effective method for devising team rating systems in an online fashion. The Poisson model directly uses the information on the number of goals scored, rather than, as in the Elo model (subject to some modifications included in the K-factor), only the final match result. In our application, the goals-based Poisson model produced better results than the variations of the Elo rating system presented here.

We argue that there are two main reasons why the iterative approaches work better than optimising jointly the whole sample of matches. First, iterative approaches consider temporal adjustments naturally. Unlike their whole-batch counterparts, they do not require result weighting, and the recency of the matches is reflected by the recency of the updates. Hence, they may be better at adapting to the true (latent) teams' shape. Second, a model which jointly optimises a sample of matches may be prone to overfitting to the whole sample, even if temporal weighting and regularisation is applied. Overall, the iterative approaches appear to be better at estimating the up-to-date team ratings.

The rating systems based on the gradient descent algorithm can be computed quickly. In terms of a dataset of k matches played by n teams, the iterative models presented in this work are of $O(k)$ time complexity. They also have minimal memory requirements of $O(n)$ as only the team ratings need to be stored. By construction, it is easy to revise the ratings using a new set of matches. On the other hand, the complexity of estimating the ratings using, e.g., the BFGS algorithm (or any other mathematical programming solver) is, in general, of a higher order of magnitude. Moreover, updating the ratings when new data arrive requires that such an algorithm recompute all ratings from scratch. The benefits of building a rating system using the gradient descent algorithm are evident, particularly in large-scale scenarios (in e-sports, for example). Yet another advantage is that this method provides transparent and interpretable update rules.

As for using other variants of the gradient descent algorithm for optimising a given loss function (e.g., RPS or Brier score), it is a worthwhile research area which may lead to new and interesting models. Moreover, the theoretical links between the base optimisation routines such as the gradient descent and the score-driven models proposed in the econometric literature (Creal et al., 2013; Koopman & Lit, 2019) merit deeper investigation. This will lead to better understanding of the advances in both areas and to a unification of their concepts.

The models presented here are evaluated for the club competitions for the domestic football league championships. They can also be applied for rating teams at the international level. This opens up interesting research issues such as the study of effects of setting weights for different matches (e.g., a friendly or a world cup match). Such a model may constitute an interesting alternative to the Elo rating system implementations for international football. Poisson regression or ordinal logistic regression-based methods seem more appropriate for modelling football scores as these approaches explicitly include draws and, in the former case, directly model the number of goals scored.

References

Aitchison, J., & Silvey, S. (1957). The generalization of probit analysis to the case of multiple responses. *Biometrika*, *44*, 131–140.

Baker, R. D., & McHale, I. G. (). Time varying ratings in association football: the all-time greatest team is.. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*, 481–492.

Barrow, D., Drayer, I., Elliott, P., Gaut, G., & Osting, B. (2013). Ranking rankings: An empirical comparison of the predictive power of sports ranking methods. *Journal of Quantitative Analysis in Sports*, *9*, 187–202.

Boshnakov, G., Kharrat, T., & McHale, I. G. (2017). A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting*, *33*, 458–466.

Boyd, S., & Vandenberghe, L. (2004). *Convex Optimization*. New York, NY, USA: Cambridge University Press.

Carbone, J., Corke, T., & Moisiadis, F. (2016). The rugby league prediction model: Using an Elo-based approach to predict the outcome of National Rugby League (NRL) matches. *International Educational Scientific Research Journal*, *2*, 26–30.

Cattelan, M., Varin, C., & Firth, D. (2013). Dynamic Bradley–Terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *62*, 135–150.

Constantinou, A., & Fenton, N. E. (2012). Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *Journal of Quantitative Analysis in Sports*, *8*, –.

Constantinou, A. C., & Fenton, N. E. (2013). Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports*, *9*, 37–50.

Constantinou, A. C., Fenton, N. E., & Neil, M. (2012). pi-football: A Bayesian network model for forecasting association football match outcomes. *Knowledge-Based Systems*, *36*, 322–339.

Creal, D., Koopman, S. J., & Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, *28*, 777–795.

Crowder, M., Dixon, M., Ledford, A., & Robinson, M. (2002). Dynamic modelling and prediction of English football league matches for betting. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *51*, 157–168.

Dixon, M. J., & Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *46*, 265–280.

Dobson, S., & Goddard, J. (2003). Persistence in sequences of football match results: A Monte Carlo analysis. *European Journal of Operational Research*, *148*, 247–256.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Resesearch*, *12*, 2121–2159.

Elo, A. (1961). The new U.S.C.F. rating system. *Chess Life*, *16*, 160–161.

Elo, A. (1978). *The Rating of Chessplayers, Past and Present*. New York, NY, USA: Arco Pub.

EloRatings.net (2020). The World Football Elo Rating System. `http://www.eloratings.net`. Last access date: 5 August 2020.

FIFA.com (2020). Ranking procedure. `https://www.fifa.com/fifa-world-ranking/procedure`. Last access date: 5 August 2020.

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics*, *48*, 377–394.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*, 359–378.

Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, *21*, 331–340.

Goddard, J. (2006). Who wins the football? *Significance*, *3*, 16–19.

Goh, G. (2017). Why momentum really works. *Distill*, .

Groll, A., Schauberger, G., & Tutz, G. (2015). Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: An application to the FIFA World Cup 2014. *Journal of Quantitative Analysis in Sports*, *11*, 97–115.

Herbrich, R., Minka, T., & Graepel, T. (2006). Trueskill™: A Bayesian skill rating system. In *Proceedings of the 19th International Conference on Neural Information Processing Systems* NIPS'06 (pp. 569–576). Cambridge, MA, USA: MIT Press.

Heuer, A., & Rubner, O. (2009). Fitness, chance, and myths: an objective view on soccer results. *European Physical Journal B*, *67*, 445–458.

Hvattum, L. M., & Arntzen, H. (2010). Using Elo ratings for match result prediction in association football. *International Journal of Forecasting*, *26*, 460–470.

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *52*, 381–393.

Kharrat, T. (2016). *A Journey Across Football Modelling with Application to Algorithmic Trading*. Ph.D. thesis University of Manchester.

Koning, R. H. (2000). Balance in competition in Dutch soccer. *Journal of the Royal Statistical Society: Series C (The Statistician)*, *49*, 419–431.

Koopman, S. J., & Lit, R. (2015). A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*, 167–186.

Koopman, S. J., & Lit, R. (2019). Forecasting football match results in national league competitions using score-driven time series models. *International Journal of Forecasting*, *35*, 797–809.

Kovalchik, S. (2020). Extension of the Elo rating system to margin of victory. *International Journal of Forecasting*, *36*, 1329–1341. URL: `http://www.sciencedirect.com/science/article/pii/S0169207020300157`. doi:`https://doi.org/10.1016/j.ijforecast.2020.01.006`.

Lasek, J., & Gagolewski, M. (2018). The efficacy of league formats in ranking teams. *Statistical Modelling*, *18*, 411–435.

Lasek, J., Szlávik, Z., & Bhulai, S. (2013). The predictive power of ranking systems in association football. *International Journal of Applied Pattern Recognition*, *1*, 27–46.

Ley, C., de Wiele, T. V., & Eetvelde, H. V. (2019). Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches. *Statistical Modelling*, *19*, 55–77.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, *36*, 109–118.

Moulton, R. (2014). A skill ranking system for Natural Selection 2. `https://moultano.wordpress.com/2014/08/04/a-skill-ranking-system-for-natural-selection-2`. Last access date: 5 August 2020.

Neave, N., & Wolfson, S. (2003). Testosterone, territoriality, and the "home advantage". *Physiology & Behavior*, *78*, 269–275.

Nocedal, J., & Wright, S. J. (2006). *Numerical Optimization*. (2nd ed.). New York, NY, USA: Springer.

Peeters, T. (2018). Testing the wisdom of crowds in the field: Transfermarkt valuations and international soccer results. *International Journal of Forecasting*, *34*, 17–29.

Pelánek, R. (2016). Applications of the Elo rating system in adaptive educational systems. *Computers & Education*, *98*, 169–179.

Pieters, W., van der Ven, S. H., & Probst, C. W. (2012). A move in the security measurement stalemate: Elo-style ratings to quantify vulnerability. In *Proceedings of the 2012 New Security Paradigms Workshop* NSPW '12 (pp. 1–14). New York, NY, USA: ACM.

Pörschmann, U., Trillmich, F., Müller, B., & Wolf, J. B. W. (2010). Male reproductive success and its behavioural correlates in a polygynous mammal, the Galápagos sea lion (Zalophus wollebaeki). *Molecular Ecology*, *19*, 2574–2586.

Rue, H., & Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *49*, 399–418.

Sismanis, Y. (2010). How I won the "Chess Ratings - Elo vs the Rest of the World" competition. *CoRR*, *abs/1012.4571*. `arXiv:1012.4571`.

Stefani, R. T. (2011). The methodology of officially recognized international sports rating systems. *Journal of Quantitative Analysis in Sports*, *7*.

Swartz, T. B., & Arce, A. (2014). New insights involving the home team advantage. *International Journal of Sports Science & Coaching*, *9*, 681–692.

Tutz, G., & Gertheiss, J. (2016). Regularized regression for categorical data. *Statistical Modelling*, *16*, 161–200.

Please cite this paper as:
J. Lasek, M. Gagolewski, Interpretable sport team rating models based on the gradient descent algorithm, International Journal of Forecasting 37(3), 1061–1071, 2021, doi:10.1016/j.ijforecast.2020.11.008