**Noname manuscript No.**
(will be inserted by the editor)

# Bayesian Score-Based Skill Learning

**Shengbo Guo · Scott Sanner**
**Thore Graepel · Wray Buntine**

Received: date / Accepted: date

**Abstract** In this paper, we extend the Bayesian skill rating system of TrueSkill to accommodate score-based match outcomes. TrueSkill has proven to be a very effective algorithm for matchmaking — the process of pairing competitors based on similar skill-level — in competitive online gaming. However, for the case of two teams/players, TrueSkill only learns from win, lose, or draw outcomes and cannot use additional match outcome information such as scores. To address this deficiency, we propose novel Bayesian graphical models based on the Gaussian and Poisson likelihood as extensions of TrueSkill that (1) model player's offence and defence skills separately and (2) model how these offence and defence skills interact to generate score-based match outcomes. Furthermore, we show that our models can naturally integrate home field advantage in a principled way. We derive efficient Bayesian inference methods for inferring latent skills in the Gaussian models, and fast fixed-point iteration method for the variational inference in the Poisson model, which is more than 1000 times faster than a sampling-based method for the Poisson model. We evaluate the proposed models on three real data sets including Halo 2 XBox Live matches. Empirical evaluations demonstrate that the new score-based models (a) provide more accurate win/loss/draw probability estimates than

Shengbo Guo
Xerox Research Centre Europe
E-mail: shengbo.guo@xrce.xerox.com

Scott Sanner
NICTA and ANU
E-mail: scott.sanner@nicta.com.au

Thore Graepel
Microsoft Research Cambridge
E-mail: thore.graepel@microsoft.com

Wray Buntine
NICTA and ANU
E-mail: wray.buntine@nicta.com.au

TrueSkill (in terms of information gain), (b) provide competitive and often better win/loss/draw (multi-class) classification accuracy, (c) provide reasonably accurate score predictions with an appropriate likelihood — prediction for which TrueSkill was not originally designed but important in many real-world applications. Furthermore, the proposed variational Bayes method for learning and inference in the Poisson model is about 1000 times faster comparing with the sampling method and performs comparably.

**Keywords** variational inference, matchmaking, graphical models

## 1 Introduction

In online gaming, it is important to pair players or teams of players so as to optimize their gaming experience. Game players often expect competitors with comparable skills for the most enjoyable experience; match experience can be compromised if one side consistently outperforms the other. *Matchmaking* attempts to pair players such that match results are close to being even or a draw. Hence, a prerequisite for good matchmaking is the ability to predict future match results correctly from historical match outcomes — a task that is often cast in terms of latent skill learning.

TrueSkill [7] is a state-of-the-art Bayesian skill learning system: it has been deployed in the Microsoft Xbox 360 online gaming system for both matchmaking and player ranking. For the case of two teams/players, TrueSkill, like Elo [6], is restricted to learn skills from match outcomes in terms of win, lose, or draw (WLD). While we conjecture that TrueSkill discards potentially valuable skill information carried by score-based outcomes, there are at least two arguments in favour of TrueSkill's WLD-based skill learning approach:

- WLD-based systems can be applied to any game whose outcome space is WLD, no matter what the underlying scoring system is.
- In many games, the objective is not to win by the highest score differential, but rather simply to win. In this case, it can be said that TrueSkill's skill modeling and learning from WLD outcomes aligns well with the players' underlying objective.

On the other hand, we note that discarding score results ignores two important sources of information:

- High (or low) score differentials can provide insight into relative team strengths.
- Two dimensional score outcomes (i.e., a score for each side) provide a direct basis for inferring separate offense and defense strengths for each team, hence permitting finer-grained modeling of performance against future opponents.

In this article, we augment the TrueSkill model of WLD skill learning to learn from score-based outcomes. We explore single skill models as well as separate offense/defense skill models made possible via score-based modeling. We also investigate both Gaussian and Poisson score likelihood models, deriving a novel variational update for approximate Bayesian inference in the latter case. Comparing with the earlier version of this work appearing in [17], we make a number of significant contributions including:

- Proposing a new sampling alternative for message passing inference of the Poisson score likelihood model,
- Proposing a new home-field advantage model for the Gaussian and Poisson score likelihood models,
- Introducing a new average baseline for predicting score-based outcomes,
- Introducing additional evaluation metrics including the multiclass WLD prediction accuracy and log likelihood, together with the associated computational descriptions,
- Conducting extensive empirical evaluations for comparing all algorithms including previous and new models and inference algorithms with additional discussion.

Our extensive empirical evaluations are conducted on three datasets: 14 years of match outcomes for the UK Premier League, 11 years of match outcomes for the Australian Football (Rugby) League (AFL), and three days covering 6,000+ online match outcomes in the Halo 2 XBox video game. Empirical evaluations demonstrate that the new score-based models (a) provide more accurate win/loss/draw probability estimates than TrueSkill (in terms of information gain) with limited amounts of training data, (b) provide competitive and often better win/loss/draw (multi-class) classification accuracy, (c) provide reasonably accurate score predictions with an appropriate likelihood — prediction for which TrueSkill was not designed but important in cases such as tournaments that rank (or break ties) by points, professional sports betting and bookmaking, and game-play strategy decisions that are dependent on final score projections. In addition, we observe that that proposed variational Bayes method for learning and inference in the Poisson model is about 1000 times faster and performs comparably, thus can achieve real-time belief updating of the skill levels of players/teams, which is essential for large-scale online matchmaking systems.

## 2 Skill Learning using TrueSkill

Since our score-based Bayesian skill learning contributions build on TrueSkill [7], we begin with a review of the TrueSkill Bayesian skill-learning graphical model for two single-player teams. We note that TrueSkill itself allows for matches involving more than two teams and learning team members' individual performances, but these extensions are not needed for the application domains considered in the paper.

Suppose there are $n$ teams available for pairwise matches in a game. Let $M = \{i, j\}$ specify the two teams participating in a match and define the outcome $o \in \{\textit{team-i-win}, \textit{team-j-win}, \textit{draw}\}$. TrueSkill models the probability

**Fig. 1** TrueSkill factor graph for a match between two single-player teams with team i winning. There are three types of variables: $l_i$ for the skills of all players, $p_i$ for the performances of all players and $d$ the performance difference. The first row of factors encode the (product) prior; the product of the remaining factors characterizes the likelihood for the game outcome team $i$ winning team $j$. The arrows show the optimal message passing schedule: (1) messages pass along *gray* arrows from top to bottom, (2) the marginal over $d$ is updated via message 1 followed by message 2 (which requires moment matching), (3) messages pass from bottom to top along *black* arrows.

$p(o|\mathbf{l}, M)$ of $o$ given the skill level vector $\mathbf{l} \in \mathbb{R}^n$ of the teams in $M$, and estimates posterior distributions of skill levels according to Bayes' rule

$$p(\mathbf{l}|o, M) \propto p(o|\mathbf{l}, M)p(\mathbf{l}), \tag{1}$$

where a factorising Gaussian prior is assumed:

$$p(\mathbf{l}) := \prod_{i=1}^{n} \mathcal{N}(l_i; \mu_i, \sigma_i^2). \tag{2}$$

To model the likelihood $p(o|\mathbf{l}, M)$, each team $i$ is assumed to exhibit a stochastic performance variable $p_i \sim \mathcal{N}(p_i; l_i, \beta^2)$ in the game [1]. From this we can model the performance differential $d$ as an indicator function $p(d|\mathbf{p}, M) = \delta(d = p_i - p_j)$ and finally the probability of each outcome $o$ given this differential $d$:

$$p(o|d) = \begin{cases} o = \textit{team-i-win}: & \mathbb{I}[d > \epsilon] \\ o = \textit{team-j-win}: & \mathbb{I}[d < -\epsilon] \\ o = \textit{draw}: & \mathbb{I}[|d| \leq \epsilon], \end{cases} \tag{3}$$

where $\mathbb{I}[\cdot]$ is an indicator function. Then the likelihood $p(o|\mathbf{l}, M)$ in (1) can be written as

$$p(o|\mathbf{l}, M) = \int \cdots \int_{\mathbb{R}^n} \int_{-\infty}^{+\infty} p(o|d)p(d|\mathbf{p}, M) \prod_{i=1}^{n} p(p_i|l_i) \, \mathrm{d}d \, \mathrm{d}\mathbf{p}.$$

The entire TrueSkill model relevant to $M$ is shown in the factor graph of Figure 1 with $P(o|d)$ given for the case of $o$ = *team-i-win*. TrueSkill uses message passing to infer the posterior distribution in (1) — note that the posterior over $l_i$ and $l_j$ will be updated according to the match outcome while the posterior over $l_k$ ($k \notin \{i, j\}$) will remain unchanged from the prior. An optimal message passing schedule in the TrueSkill factor graph (Figure 1) is provided in the caption; the message along arrow 2 is a step function that leads to intractability for exact inference and thus TrueSkill uses message approximation via moment matching.

TrueSkill is an efficient and principled Bayesian skill learning system. However, due to its design goals, it discards score information and does not take

[1] Note that we sometimes abuse notations on the use of $p$, $p_i$ and $\mathbf{p}$. $p$ is a probability measure; $p_i$ and $\mathbf{p}$ represent performance variables. The meaning of them is clear from the context.

into account associated domain knowledge such as offence/defence skill components. Next, we propose extensions of the TrueSkill factor graph and (approximate) inference algorithms for score-based Bayesian skill learning, which address these limitations.

## 3 Score-based Bayesian Skill Models

In this section, we introduce three graphical models as extensions for the TrueSkill factor graph (Figure 1) to incorporate score-based outcomes in skill learning. Our first two graphical models are motivated by modeling score-based outcomes as generated by separate offence and defence skills for each team. The first generative score model uses a Poisson, which is natural model when scores are viewed as counts of scoring events. The second generative model uses a simpler Gaussian model. For both of these two models, we also propose a variant of them that can take into account home field advantages. Our third model is a simplified version of the Gaussian model, which like TrueSkill, only models a single skill per team (not separate offence/defence skills) and places a Gaussian likelihood on the score difference, which may be positive or negative. Next we formulate each model in detail.

### 3.1 Offence and Defence Skill Models

In a match between two teams $i$ and $j$ producing respective scores $s_i \in \mathbb{Z}$ and $s_j \in \mathbb{Z}$ for each team, it is natural to think of $s_i$ as resulting from $i$'s offence skill $o_i \in \mathbb{R}$ and $j$'s defence skill $d_j \in \mathbb{R}$ (as expressed in any given match) and likewise for $j$'s score as a result of $j$'s offence skill $o_j \in \mathbb{R}$ and $i$'s defence skill $d_i \in \mathbb{R}$. This is contrasted with the univariate skill estimates of team $i$'s skill $l_i$ and team $j$'s skill $l_j$ used in TrueSkill, which lump together offence and defence skills for each team.

Given scores $s_i$ and $s_j$ for teams $i$ and $j$, we model the generation of scores from skills using a conditional probability $p(s_i, s_j | o_i, o_j, d_i, d_j)$. We assume that team $i$'s score $s_i$ depends only on $o_i$ and $d_j$ and likewise that team $j$'s score $s_j$ depends only on $o_j$ and $d_i$:

$$p(s_i, s_j | o_i, o_j, d_i, d_j) = p(s_i | o_i, d_j) p(s_j | o_j, d_i). \tag{4}$$

Like TrueSkill, we assume that the joint marginal over skill priors independently factorises:

$$p(o_i, o_j, d_i, d_j) = p(o_i) p(d_j) p(o_j) p(d_i). \tag{5}$$

Given an observation of scores $s_i$ for team $i$ and $s_j$ for team $j$, the problem is to update the posterior distributions over participating teams' offence and defence skills. According to Bayes rule and the previous assumptions, the posterior distribution over $(o_i, o_j, d_i, d_j)$ is given by

$$\begin{aligned} p(o_i, d_i, o_j, d_j | s_i, s_j) &\propto p(s_i, s_j | o_i, d_i, o_j, d_j) p(o_i, d_i, o_j, d_j) \\ &\propto [p(s_i | o_i, d_j) p(o_i) p(d_j)] \; [p(s_j | o_j, d_i) p(o_j) p(d_i)]. \qquad (6) \end{aligned}$$

Here we observe that estimating $p(o_i, d_i, o_j, d_j | s_i, s_j)$ factorises into the two independent inference problems:

$$p(o_i, d_j | s_i) \propto p(s_i | o_i, d_j) p(o_i) p(d_j), \text{and} \qquad (7)$$

$$p(o_j, d_i | s_j) \propto p(s_j | o_j, d_i) p(o_j) p(d_i). \qquad (8)$$

All models considered in this paper (including TrueSkill) assume Gaussian priors on team $i$'s offence and defence skills, i.e., $p(o_i) := \mathcal{N}(o_i; \mu_{oi}, \sigma_{oi}^2)$ and $p(d_i) := \mathcal{N}(d_i; \mu_{di}, \sigma_{di}^2)$. Our objective then is to estimate the means and variances for the posterior distributions of $p(o_i, d_j | s_i)$ and $p(o_j, d_i | s_j)$. So far, the only missing pieces in this skill posterior update are the likelihoods $p(s_i | o_i, d_j)$ and $p(s_j | o_j, d_i)$ that specify how team $i$ and $j$'s offence and defence skills probabilistically generate observed scores. For this we discuss two possible models in the following subsections.

### *3.1.1 Poisson Offence/Defence Skill Model*

Following TrueSkill, we model the generation of match outcomes (in our case, team scores) based on stochastic offence and defence *performances* that account for day-to-day performance fluctuations. Formally, we assume that team $i$ exhibits offence performance $p_{oi} := \mathcal{N}(p_{oi}; o_i, \beta_o^2)$ and defence performance $p_{di} := \mathcal{N}(p_{di}; d_i, \beta_d^2)$. With these performances, we model team $i$'s score $s_i$ as generated from the following process: team $i$'s offence performance $p_{oi}$ promotes the scoring rate while the defence performance $p_{dj}$ inhibits this scoring rate, the difference $p_{oi} - p_{dj}$ being the effective scoring rate of the offence against the defence.

Finally, we model the score by $s_i \sim \text{Poisson}(\lambda)$, where a requirement of a positive rate $\lambda$ for the Poisson distribution requires the use of $\lambda = \exp(p_{oi} - p_{dj})$ since $p_{oi} - p_{dj}$ may be negative.[2] Likewise, one can model $s_j$ by applying the same strategy when given $\lambda = \exp(p_{oj} - p_{di})$. We represent the resulting *Poisson-OD* model in Figure 2 where the joint posterior is

$$\begin{aligned} p(o_i, d_j, p_{oi}, p_{dj} | s_i) &\propto p(s_i | p_{oi}, p_{dj}) p(p_{oi} | o_i) p(p_{dj} | d_j) p(o_i) p(d_j), \\ p(o_j, d_i, p_{oj}, p_{di} | s_j) &\propto p(s_j | p_{oj}, p_{di}) p(p_{oj} | o_j) p(p_{di} | d_i) p(o_j) p(d_i). \end{aligned}$$

[2] This exponentiation of $p_{oi} - p_{dj}$ may seem to be made only to ensure model correctness, but we show experimentally that it has the benefit of allowing the Poisson model to accurately predict scores in high-scoring games even when team skills are very close (and hence $p_{oi} - p_{dj} \approx 0$).

**Fig. 2** The Poisson-OD variants of TrueSkill factor graph for skill update of two teams based on the match score outcome (Left: modeling $s_i$; Right: modeling $s_j$). Note that the score observation factors use the Poisson distribution for the Poisson-OD model. Shaded nodes are observed variables. For each team $i$, it is characterized by offence skill $o_i$ (the offence skill of team $i$) and defence skill $d_i$ (the defence skill of team $i$). Given $s_j$ for team $j$, the posterior distributions over $(o_i, d_j)$ are inferred via message passing.

**Fig. 3** The Poisson-OD-AH variants of the Poisson-OD factor graph for skill update of two teams based on the match score outcome (Left: modeling $s_i$; Right: modeling $s_j$), with team $i$ playing home field. Note that $h_i$ is the latent variable representing home field advantages associated with team $i$, and note also that the score observation factors use the Poisson distribution for the Poisson-OD-AH model. Shaded nodes are observed variables. For each team $i$, it is characterized by offence skill $o_i$ (the offence skill of team $i$), defence skill $d_i$ (the defence skill of team $i$), together with home field advantage variable $h_i$. Given $s_i$ for team $i$, the posterior distributions over $(o_i, h_i, d_j)$ are inferred via message passing). Likewise, given team $j$'s score $s_j$, the posterior distributions over $(o_i, d_j)$ are inferred via message passing.

We are only interested in the posterior distributions of $o_i, d_j$ and $o_j, d_i$ given $s_i$ and $s_j$, respectively. Thus, we integrate out the latent performance variables to obtain the desired posteriors

$$p(o_i, d_j|s_i) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(o_i, d_j, p_{oi}, p_{dj}|s_i) \mathrm{d}p_{oi} \mathrm{d}p_{dj},$$
$$p(o_j, d_i|s_j) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} p(o_j, d_i, p_{oj}, p_{di}|s_j) \mathrm{d}p_{oj} \mathrm{d}p_{di}.$$

Like TrueSkill, we use Bayesian updating to update beliefs in the skill levels of both teams in a pairwise match based on the score outcome, thus leading to an online learning scheme. Posterior distributions are approximated to be Gaussian and used as the priors in order to learn each team's skill for the next match. Approximate belief updates via variational Bayesian inference in this model will be covered in Section 4.2.

We pause at this point to introduce a variant of the Poisson-OD model 2 for modeling home field advantages, and name this model Poisson-OD-AH (advantages for home playing teams) shown in Figure 3. The Poisson-OD-AH factor graph represents the generative process for a match between team $i$ and $j$ scored $s_i$ and $s_j$, respectively, and team $i$ plays in the home field in this match. To model the home field advantages, we associate each team $i$ with an additional latent variable $h_i$ to represent the possible gain in performance for a home field playing team, and propose to learn the distribution for $h_i$ in the same fashion as the offence and defence skill variables. We can omit the details of the inference in this model as one can apply the same inference method for the Poisson-OD model for this Poisson-OD-AH model.

### *3.1.2 Gaussian Offence/Defence Skill Model*

An alternative to the previous Poisson model is to model $s_i \in \mathbb{R}$ and assume it is generated as $s_i \sim \mathcal{N}(\mu, \gamma^2)$, where $\mu = p_{oi} - p_{dj}$. One can similarly model $s_j$

**Fig. 4** The Gaussian-OD variant of the TrueSkill factor graph for skill update of two teams based on the match score outcome (Left: modeling $s_i$; Right: modeling $s_j$). Note that the score observation factors use the Gaussian distribution for the Gaussian-OD model. Shaded nodes are observed variables. For each team $i$, it is characterized by offence skill $o_i$ (the offence skill of team $i$) and defence skill $d_i$ (the defence skill of team $i$). Given $s_j$ for team $j$, the posterior distributions over $(o_i, d_j)$ are inferred via message passing.

**Fig. 5** The Gaussian-OD-AH variant of the Gaussian-OD factor graph for belief updating of two teams based on (1) the match score outcome (Left: modeling $s_i$; Right: modeling $s_j$) and (2) home/away fields. Note that the score observation factors use the Gaussian distribution for the Gaussian-OD model. Shaded nodes are observed variables. For each team $i(j)$, it is characterized by offence skill $o_i$ (the offence skill of team $i$), defence skill $d_i$ (the defence skill of team $i$), and home field advantage variable $h_i$. Given $s_i$ for team $i$, the posterior distributions over $(o_i, h_i, d_j)$ are inferred via message passing, and same for $(o_j, d_i)$ given $s_j$.

by applying the same strategy when given $\mu = p_{oj} - p_{di}$. We note that unlike the Poisson model, $\mu$ can be negative here so we need not exponentiate it. While this allows us to directly model match outcomes that allow negative team scores (c.f., Halo2 as discussed in Section 5.1), it is problematic for other match outcomes that only allow non-negative team scores. One workaround would be to introduce a truncated Gaussian model to avoid the problem of assigning non-zero probability to negative scores, but we avoid this complication in exchange for the simple and exact updates offered by a purely Gaussian model.

We show the resulting *Gaussian-OD* model in Figure 4, which differs from our proposed Poisson model only in modeling the observed score $s_i$ ($s_j$) for team $i$ ($j$) given the univariate performance difference variable $x$ ($y$). In this model, all messages passed during inference are Gaussian, allowing for efficient and exact belief updates.

For the Gaussian-OD model, we also introduce a variant to encode the home advantage variable $h_i$ for team $i$ playing in the home field, and we name this model the Gaussian-OD-AH model shown in Figure 5. The left hand side of the factor graphs is the generative model for team $i$'s score $s_i$ given that team $i$ is the home-field playing team, team $i$'s offence skill $o_i$, and team $j$'s defence skill $d_j$. Likewise for the right hand side for modeling team $j$'s score $s_j$ except the home advantage variable. We can omit the detailed derivation for the exact Bayesian inference in this model, because it is the same as in the Gaussian-OD model.

### 3.2 Gaussian Score Difference (SD) Model

Again assuming $s_i \in \mathbb{R}$ and $s_j \in \mathbb{R}$, algebra for the performance means in Figure 4 gives:

$$s_i = p_{oi} - p_{dj}, \qquad s_j = p_{oj} - p_{di}. \tag{9}$$

**Fig. 6** The Gaussian-SD variant of the TrueSkill factor graph model for skill update of two teams based on the score difference. Both team $i$ and team $j$ are characterized by skill level $l_i$ and $l_j$, respectively. The shaded node $s$ ($s = s_i - s_j$) denotes the score difference between $s_i$ and $s_j$. Bayesian inference for the posterior skill level distributions has a closed-form solution.

This implies

$$\begin{aligned} s_i - s_j &= (p_{oi} - p_{dj}) - (p_{oj} - p_{di}) \\ &= \underbrace{(p_{oi} + p_{di})}_{p_{li}} - \underbrace{(p_{oj} + p_{dj})}_{p_{lj}}, \end{aligned} \tag{10}$$

which is like modeling the score difference with performance expressions $p_{li}$ and $p_{lj}$ of respective univariate skill levels, $l_i$ and $l_j$. Motivated by (9), we propose a score difference (SD) Gaussian model that uses a likelihood model for the observed difference $s := s_i - s_j$ specified as $s \sim \mathcal{N}(p_{li} - p_{lj}, \gamma^2)$ as shown in Figure 6.

## 4 Skill and Win/Lose/Draw Probability Inference

We infer skill distributions in all proposed models via online Bayesian updating. While exact inference in the purely Gaussian models can be achieved by solving linear systems, Bayesian updating provides an efficient (also exact) incremental learning alternative. Equations for Bayesian updates and the WLD probability of matches are model-dependent and presented below.

### 4.1 Inference in TrueSkill

**Bayesian update:** The Bayesian update equations in the TrueSkill model (Figure 1) are presented in [7].
**WLD probability:** Given skill levels of team $i$ and $j$, $l_i \sim \mathcal{N}(l_i; \mu_i, \sigma_i^2)$ and $l_j \sim \mathcal{N}(l_j; \mu_j, \sigma_j^2)$, we first compute the distribution over performance difference variable $d$, and get $d \sim \mathcal{N}(d; \mu_d, \sigma_d^2)$ with $\mu_d = \mu_i - \mu_j$ and $\sigma_d^2 = \sigma_i^2 + \sigma_j^2 + 2\beta^2$. The winning probability of team $i$ is given by the probability $p(d > \epsilon)$ defined as

$$p(d > \epsilon) = 1 - \Phi\left(\frac{\epsilon - \mu_d}{\sigma_d}\right), \tag{11}$$

where $\Phi(\cdot)$ is the normal CDF and the $\epsilon$ is the draw margin computed on the training data. Likewise, one can define the lose probability for team $i$ as $p(d < \epsilon)$, and the draw probability as $p(|d| < \epsilon)$. The setting of draw margin is essential for WLD probability calculation. For the TrueSkill model, there are two draw margins involved: one is used in the TrueSkill factor graph (Figure 1) computed by taking the proportion of matches with draws out of all the matches; the other is used here to compute the WLD probability, which is obtained by maximizing the WLD prediction accuracy (defined in Section 5.2.2) on the training data set.

## 4.2 Inference in Poisson-OD Model

### *4.2.1 Bayes Update*

Some of the update equations in the Poisson-OD model (Figure 2) have been presented in [7], with the exception of the marginal distribution over $x$ and the message passing from the Poisson factor to $x$. Given a prior Gaussian distribution over $x$, $\mathcal{N}(x; \mu, \sigma^2)$, we next demonstrate how to update the belief on $x$ when observing team $i$'s score $s_i$.

By the sum-product algorithm [8], the marginal distribution of $x$ is given by a product of messages

$$p(x|s_i) = m_{\delta \to x}(x) m_{s_i \to x}(x). \tag{12}$$

To avoid cluttered notation, let us use $m_1(x)$ to represent $m_{\delta \to x}(x) = \mathcal{N}(x; \mu, \sigma^2)$, i.e., the message passing from the factor $\delta(\cdot)$ to $x$, and $m_2(x)$ for $m_{s_i \to x}(x) = Poisson(s_i; \exp(x))$, i.e., the message passing from the Poisson factor to $x$ (c.f., messages labeled 1 and 2 in Figure 2). Due to the multiplication of $m_1(x)$ and $m_2(x)$, the exact marginal distribution of $p(x|s_i)$ is not Gaussian, which makes exact inference intractable. To maintain a compact representation of offence and defence skills, one can approximate $p(x|s_i)$ with a variational Bayes framework or a sampling-based approach considering its being a univariate distribution.

**Bayesian update with VB** In a variational Bayes framework, the problem is to choose a Gaussian distribution $q(x)^* : \mathcal{N}(x; \mu_{\text{new}}, \sigma^2_{\text{new}})$ that minimizes the KL divergence between $p(x|s_i)$ and $q(x)$, i.e.,

$$q(x)^* = \underset{q(x)}{\arg\min} \, \mathrm{KL}\left[q(x) || p(x|s_i)\right]. \tag{13}$$

We derive a fixed-point approach for optimizing $q(x)$ [12] and describe this approach below.

**Minimizer $q(x)$ for KL$(q(x)||p(x|s_i))$:** We first expand the KL-divergence into its definition:

$$\begin{aligned} \mathrm{KL}\left(q(x)||p(x|s_i)\right) &= \int q(x) \log \left( \frac{q(x)}{p(x|s_i)} \right) dx \\ &= -\log \sqrt{2\pi e \sigma^2_{new}} - E_{x \sim q(x)} \log \left(p(x|s_i)\right), \end{aligned} \tag{14}$$

where $p(x|s_i)$ is the posterior probability of $x$ when observing the score $s_i$. Since $q(x)$ is Gaussian and the posterior has convenient Gaussian parts, manipulation of this yields an equation for $\mu_{new}$ and $\sigma^2_{new}$ that can be solved using an iterative fixed-point approach:

**Lemma 1** *Values for $\mu_{new}$ and $\sigma^2_{new}$ minimizing KL $(q(x)||p(x|s_i))$ satisfy*

$$\begin{aligned}\mu_{new} &= \sigma^2 \left(s_i - e^{\kappa}\right) + \mu, \\ \sigma^2_{new} &= \frac{\sigma^2}{1+\sigma^2 e^{\kappa}},\end{aligned} \tag{15}$$

*where*

$$\kappa = \log\left(\frac{\mu + s_i\sigma^2 - 1 - \kappa + \sqrt{(\kappa - \mu - s_i\sigma^2 - 1)^2 + 2\sigma^2}}{2\sigma^2}\right). \tag{16}$$

*Ths fixed point equation in $\kappa$ converges linearly with factor upper bounded by $\left(2/3\sigma^2 e^{\kappa}\right)^{-1}$.*

*Proof* The second term in (14) is evaluated using Bayes Theorem, $p(x|s_i) = p(s_i|x)p(x)/p(s_i)$. The term in $\log p(s_i)$ can be dropped because it is constant with respect to $\mu_{\text{new}}$ and $\sigma^2_{\text{new}}$. The term $E_{x\sim q(x)}[\log p(s_i|x)]$ is found by expanding the Poisson distribution. Note the expected value[3] of $\exp(y)$ for $\mathcal{N}(y;\mu,\sigma^2)$ is $\exp(\mu + \sigma^2/2)$. Thus

$$E_{x\sim q(x)}[\log p(s_i|x)] \;=\; s_i\mu_{\text{new}} - \exp(\mu_{\text{new}} + \sigma^2_{\text{new}}/2) - \log(s_i!)\ . \tag{17}$$

The term $E_{x\sim q(x)}[\log p(x)]$ is readily derived because $\log p(x)$ is a quadratic. This becomes

$$-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\sigma^2_{new} + \mu^2_{new} - 2\mu\mu_{new} + \mu^2\right)\ . \tag{18}$$

Plugging (17) and (18) into (14) gives

$$\begin{aligned}&\text{KL}\left(q(x)||p(x|s_i)\right) \equiv -\log\sqrt{2\pi e\sigma^2_{new}} - \\ &\Bigg(\underbrace{s_i\mu_{\text{new}} - \exp(\mu_{\text{new}} + \sigma^2_{\text{new}}/2) - \log(s_i!)}_{E_{x\sim q(x)}(\log p(s_i|x))} \\ &\underbrace{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\sigma^2_{new} + \mu^2_{new} - 2\mu\mu_{new} + \mu^2\right)}_{E_{x\sim q(x)}(\log p(x))}\Bigg).\end{aligned}$$

To find the minimizer $q(x)$, we calculate the partial derivatives of KL $(q(x)||p(x|s_i))$ w.r.t. $\mu_{\text{new}}$ and $\sigma_{new}$, and set them to zero, leading to

$$\begin{aligned}\mu_{\text{new}} &= \sigma^2\left(s_i - \exp\left(\mu_{\text{new}} + \frac{\sigma^2_{\text{new}}}{2}\right)\right) + \mu, \\ \sigma^2_{\text{new}} &= \frac{\sigma^2}{1 + \sigma^2\exp(\mu_{\text{new}} + \frac{\sigma^2_{new}}{2})}\ .\end{aligned}$$

[3] Shown by manipulating the Gaussian integral.

Define $\kappa = \mu_{\text{new}} + \sigma^2_{\text{new}}/2$ and these equations yield (15). Moreover, summing the first plus half the second of these equations yields the equation for $\kappa$ of

$$\kappa = \mu + \sigma^2(s_i - e^\kappa) + \frac{\sigma^2}{2(1+\sigma^2 e^\kappa)}. \tag{19}$$

We convert (19) by solving for $e^\kappa$ as it appears on the right-hand side. This yields a quadratic equation in $e^k$:

$$\sigma^2 e^{2\kappa} + (\kappa - \mu + 1 - s_i\sigma^2)e^\kappa + \big((\kappa-\mu)/\sigma^2 - s_i - 1/2\big) = 0 \ .$$

For the quadratic in the form $Ae^{2\kappa} + Be^\kappa + C$ we note by appropriate manipulation of (15)

$$\begin{aligned} AC = \kappa - \mu - s_i\sigma^2 - \sigma^2/2 &= -\sigma^2 e^\kappa + \frac{\sigma^2}{2\,(1+\sigma^2 e^\kappa)} - \sigma^2/2 \\ &= -\sigma^2 e^\kappa \left(1 + \frac{\sigma^2}{2\,(1+\sigma^2 e^\kappa)}\right) \ . \end{aligned}$$

This must be negative. Thus it follows that the quadratic has a positive and negative solution. We take the positive solution since $e^\kappa$ must be non-negative. Simplifying the term inside the square root,

$$(\kappa - \mu + 1 - s_i\sigma^2)^2 - 4\left(\kappa - \mu - s_i\sigma^2 - \sigma^2/2\right) = (\kappa - \mu - 1 - s_i\sigma^2)^2 + 2\sigma^2 \ ,$$

gives us (16).

Now consider (16) as a fixed point equation in $e^\kappa$. Linear convergence will hold with rate $\left|\frac{\delta \exp(\kappa')}{\delta \exp(\kappa)}\right|$, so we need an upper bound on this value. From (16) we get

$$\frac{\delta \exp(\kappa')}{\delta \exp(\kappa)} = \frac{1}{2\sigma^2 e^\kappa}\left(-1 + 2\frac{(\kappa - \mu - s_i\sigma^2 - 1)}{\left((\kappa - \mu - s_i\sigma^2 - 1)^2 + 2\sigma^2\right)^{-1/2}}\right)$$

Note the term in brackets is in the range $-1 + -2$ to $-1 + 2$. The bound on the rate follows.

We can use (16) as a fixed-point rewrite rule. For a given $\mu$ and $\sigma^2$ together with an initial value of $\kappa$, one iterates (16) until convergence. Empirically, this happens within 2-3 iterations because typically $2\sigma^2 e^\kappa > 10^3$. With convergence, we substitute the fixed-point solution into (15) to get the optimal mean and variance for $q(x)^*$.

**Bayes Update with Slice Sampling** One caveat associated with variational Bayes is that it may yield local minimum; and thus one tends to use sampling methods such as Markov chain Monte Carlo (MCMC) to obtain exact solutions. In our Poisson-OD model, recall that the problem is to compute the outgoing message for the variable $x$ as it is not tractable. To approximate $p(x|s_i)$ in the message passing framework, we propose an MCMC based sampling method named slice sampling, which can draw samples from any analytical univariate distribution [1].

During message passing inference framework for the Poisson-OD model factor graph (Figure 2), we first compute the outgoing message for the variable $x$ by defining its analytical function as $m_{\delta \to x}(x) m_{s_i \to x}(x)$ (Refer to Equation 12 for details of these notations), and then call the Matlab embedded function *slicesample* to draw samples from this function. With these samples, we can compute a normal approximation of this outgoing message.

### *4.2.2 WLD Probability*

Suppose we are given the offence and defence skills for team $i$ and $j$, we can estimate the distributions over performance difference variables of $x$ and $y$ (c.f., Figure 2), and compute the Poisson parameters for $s_i$ and $s_j$ by using $\lambda_i = \exp(x)$ and $\lambda_j = \exp(y)$. To compute the winning probability of team $i$, i.e., $p(s_i > s_j)$, we first construct a new variable $s = s_i - s_j$, the difference variable between two Poisson distributions, which proves to be a Skellam distribution in [10]. Thus, we can compute the win probability of $P(s > 0)$ of team $i$, according to the probability mass function for the Skellam distribution

$$P(s = k; \lambda_i, \lambda_j) = e^{-(\lambda_i + \lambda_j)} \left( \frac{\lambda_i}{\lambda_j} \right)^{k/2} I_{|k|} \left( 2\sqrt{\lambda_i \lambda_j} \right),$$

where $I_k(z)$ is the modified Bessel function of the first kind given in [3]:

$$I_k(z) = \left( \frac{z}{2} \right)^k \sum_{i=0}^{+\infty} \frac{(z^2/4)^i}{i! \Gamma(k + i + 1)}. \tag{20}$$

We approximated $P(s > 0, \lambda_i, \lambda_j)$ with $\sum_{k=1}^{n} P(s = k; \lambda_i, \lambda_j)$ using $n = 100$ since $P(s = k; \lambda_i, \lambda_j) \approx 0$ for all of our experiments when $k > 100$. Likewise, we can compute the winning probability of team $j$ by constructing $s = s_j - s_i$ and computing $P(s > 0)$, and the draw probability by computing $P(s = 0)$.

## 4.3 Inference in Gaussian-OD Model

### *4.3.1 Bayesian update*

In the Gaussian-OD model (Figure 4), all messages are Gaussian so one can compute the belief update in closed-form as follows

$$\begin{aligned}
\pi_{o_i} &= \frac{1}{\sigma_{o_i}^2} + \frac{1}{\beta_1^2 + \beta_2^2 + \gamma^2 + \sigma_{d_j}^2}, \\
\tau_{o_i} &= \frac{\mu_{o_i}}{\sigma_{o_i}^2} + \frac{s_i + \mu_{d_j}}{\beta_1^2 + \beta_2^2 + \gamma^2 + \sigma_{d_j}^2}, \\
\pi_{d_j} &= \frac{1}{\sigma_{d_j}^2} + \frac{1}{\beta_1^2 + \beta_2^2 + \gamma^2 + \sigma_{o_i}^2}, \\
\tau_{d_j} &= \frac{\mu_{d_j}}{\sigma_{d_j}^2} + \frac{\mu_{o_i} - s_i}{\beta_1^2 + \beta_2^2 + \gamma^2 + \sigma_{o_i}^2}, \qquad (21)
\end{aligned}$$

where $\mu_{o_i}$ and $\sigma_{o_i}$ are the mean and standard deviation of the prior offence skill distribution of team $i$, $\pi_{o_i}(\pi_{d_j}) = \frac{1}{\sigma_{post}^2}$ and $\tau_{o_i}(\tau_{d_j}) = \frac{\mu_{post}}{\sigma_{post}^2}$ are the precision and precision-adjusted mean for the posterior offence (defence) skill distribution of team $i$ $(j)$. Likewise, one can derive the update equations for team $j$'s offence skill $o_j$ and team $i$'s defence skill $d_i$.

### *4.3.2 WLD Probability*

To compute the probability of team $i$ winning vs team $j$, we first use message passing to estimate the normally distributed distributions for score variables $s_i$ and $s_j$, and then compute the probability that $s_i - s_j > \epsilon$, i.e., team $i$'s score is larger than team $j$'s. Given $s_i \sim \mathcal{N}(s_i; \mu_{si}, \sigma_{si}^2)$ and $s_j \sim \mathcal{N}(s_j; \mu_{sj}, \sigma_{sj}^2)$, we can compute the winning probability of team $i$ by

$$p(s > \epsilon) = 1 - \Phi\left(\frac{\epsilon - (\mu_{si} - \mu_{sj})}{\sigma_{si}^2 + \sigma_{sj}^2}\right). \qquad (22)$$

where $\epsilon$ is a parameter to represent draw margin. Similarly, the lose probability of team $i$ can be computed by $p(s < -\epsilon)$, and the draw probability by $p(|s| < \epsilon$, which are both easy to compute as $s$ is distributed normally. Note that the parameter draw margin $\epsilon$ is optimized by maximizing the WLD accuracy (defined in Section 5.2.2) on the training data set.

## 4.4 Inference in Gaussian-SD Model

### *4.4.1 Bayes Update*

In the Gaussian-SD model (Figure 6), all messages are Gaussian so we can again derive the update for the single team skills $l_i$ and $l_j$ in closed-form as follows:

$$\begin{aligned}
\pi_{l_i} &= \frac{1}{\sigma_{l_i}^2} + \frac{1}{\beta_1^2 + \beta_2^2 + \gamma^2 + \sigma_{l_j}^2}, \\
\tau_{l_i} &= \frac{\mu_{l_i}}{\sigma_{l_i}^2} + \frac{(s_i - s_j) + \mu_{l_j}}{\beta_1^2 + \beta_2^2 + \gamma^2 + \sigma_{l_j}^2}, \\
\pi_{l_j} &= \frac{1}{\sigma_{l_j}^2} + \frac{1}{\beta_1^2 + \beta_2^2 + \gamma^2 + \sigma_{l_i}^2}, \\
\tau_{l_j} &= \frac{\mu_{l_j}}{\sigma_{l_j}^2} + \frac{\mu_{l_i} - (s_i - s_j)}{\beta_1^2 + \beta_2^2 + \gamma^2 + \sigma_{l_i}^2}, \qquad (23)
\end{aligned}$$

where $\mu_{l_i}$ ($\mu_{l_j}$) and $\sigma_{l_i}$ ($\sigma_{l_j}$) are the mean and standard deviation of team $i$'s (team $j$'s) prior skill distribution, $\pi_{l_i}$ ($\pi_{l_j}$) and $\tau_{l_i}$ ($\tau_{l_j}$) are the precision and precision adjusted mean for team $i$'s (team $j$'s) posterior skill distribution.

### *4.4.2 WLD probability*

To estimate the winning probability of team $i$ for a match with team $j$, one can first use message passing to estimate the normally distributed score difference variable $s$, and then compute the winning probability of team $i$ by

$$p(s > \epsilon) = 1 - \Phi\left(\frac{\epsilon - l_i - l_j}{\sigma_i^2 + \sigma_j^2 + 2\beta^2}\right), \qquad (24)$$

where $l_i$ and $\sigma_i$ are the mean and standard deviation for team $i$'s skill level, and $\beta$ the standard deviation of the performance variable. Likewise, the lose probability of team $i$ is $p(s < -\epsilon)$, and the draw probability $p(|s| < \epsilon)$. Similar with the Gaussian-OD model, $\epsilon$ is computed by optimizing the WLD predict accuracy defined in Section 5.2.2 on the training data set.

## 5 Empirical Evaluation

### 5.1 Data Sets

Experimental evaluations are conducted on three data sets: Halo 2 XBox Live matches, Australian Football (Rugby) League (AFL)and UK Premier League (UK-PL)[4]. The Halo 2 data consists of a set of match outcomes comprising 6227 games for 1672 players. We note there are negative scores for this data, so we add the absolute value of the minimal score to all scores to use the data with all proposed models.

The training and testing settings are described as follows. For Halo 2 [5], the last 10% of matches are used for testing, and we use different proportions

[4] `http://www.football-data.co.uk/englandm.php`

[5] Credit for the use of the Halo 2 Beta Data set is given to Microsoft Research Ltd. and Bungie.

of the first 90% of data for training. There are 8 proportions used for training, ranging from 10% to 80% with an increment of 10%, and 90% is not used for training due to cross validation. To cross validate, we randomly sample the data and run the learning 10 times at each proportion level to get standard error bars. Note that there are some players in the testing games who are not involved in any training data sets, particularly when small proportion of training data set is selected (e.g., the first 10 percent games); we remove these games in the testing set when reporting performances for all models.

For both UK-PL and AFL datasets, cross validation is performed by training and testing for each year separately (14 years for UK-PL, and 11 years for AFL). For these two datasets, we test the last 20% percent of matches in each year, with the training data increasing incrementally from 10% to 80% of the initial matches.

## 5.2 Evaluation Criteria

We evaluate performances using four criteria: *information gain* (Section 5.2.1), *win/lose/draw prediction accuracy* (Section 5.2.2), *log-likelihood* (Section 5.2.3), and *score prediction errors* (Section 5.2.4). While the first three criteria focus on evaluating win/lose/draw prediction accuracy, the fourth criterion measures how good a model is at predicting scores, for which TrueSkill does not apply since it is restricted to WLD only. Let us introduce each criterion in detail.

### *5.2.1 Information Gain*

The first criterion we use to evaluate different approaches is *information gain*, which is proposed in the *Probabilistic Footy Tipping Competition*[6]: if a predictor assigns probability $p$ to team $i$ winning, then the score (in "bits") gained is $1+\log_2(p)$ if team $i$ wins, $1+\log_2(1-p)$ if team $i$ loses, $1+(1/2)\log_2(p(1-p))$ if draw happens. This evaluation metric can be viewed as an information gain interpretable variant of a log likelihood score where an uninformed prediction of $p = 0.5$ leads to a score of 0 and a definite prediction of $p = 1$ ($p = 0$) leads to a score of $-\infty$ if predicting incorrectly and 1 if predicting correctly. In Section 4, we showed how to compute the win probability $p$ for each model.

### *5.2.2 WLD Prediction Accuracy*

One important task for matchmaking is to predict the match outcomes in terms of WLD, which is a multi-class classification problem involving 3 classes in this case. Contrasted with our earlier work in [17] reporting the results for predicting Win/noWin, a binary task, we consider this multi-class setting in this paper.

[6] Refer to `http://www.csse.monash.edu.au/~footy/`

We use the accuracy of the WLD prediction as performance measurement. The accuracy is computed by first generating the three by three confusion matrix based on the WLD probability calculated for each model according to Section 4, with which we take the number of correctly predicting matches divided by the total number of testing matches.

### *5.2.3 Log Likelihood*

Log likelihood is often one of the criteria used to evaluate Bayesian models, so we report the log likelihood for all proposed models together with TrueSkill. In order to compare with TrueSkill that is limited to predict WLD only, we convert our score-based predictions into WLD prediction. Details of computing the WLD probability calculation for each model can be found in Section 4.

### *5.2.4 Score Prediction Error*

We evaluate the score prediction accuracy for the two full score prediction models: Poisson-OD and Gaussian-OD, and their variants that model home advantages. For the Poisson-OD model (Figure 2) and its home advantage variant model, the expected score for team $i$ ($j$) is $\exp(x)$ ($\exp(y)$) and for the Gaussian-OD model (Figure 4), the expected score for team $i$ ($j$) is $x$ ($y$), where $x$ ($y$) is the difference in mean performances giving $s_i$ ($s_j$).

The measure we use for evaluating score prediction accuracy is the mean absolute error (MAE) defined below

$$\frac{1}{2N}\sum_{i=1}^{2N}(|\hat{s}_i - s_i| + \hat{s}_j - s_j|) \tag{25}$$

where $\hat{s}_i$ ($\hat{s}_j$) is the predicted score for team $i$ ($j$), $s_i$ ($s_j$) the ground truth, and $N$ the number of two-team matches for with teams indexed by $i$ and $j$.

Note that we must omit the Gaussian-SD model since it can only predict score differences rather than scores. To benchmark the score prediction performance of the Poisson-OD and Gaussian-OD models, we compare with an average score prediction methods. This average score methods simply use the average scores for a team computed from the training games as predictions for testing games.

## 5.3 Results on Four Criteria

Experimental results are reported according to the parameter configurations shown in Table 1. The draw margins used to compute WLD probability for the TrueSkill, Gaussian-OD, and Gaussian-SD are chosen in the way such that the WLD classification accuracy on the training data is maximized, and we omit the details of these optimized parameters, because we optimize these parameters for each training/testing splitting for each training proportion.

Parameters for the slice sampling used in the Poisson-OD model include the burn-in, thinning, and the number of samples required, which we set to 500, 2, and 500, respectively. Before we discuss the results against these four criteria on three real data sets, we emphasize that the WLD probabilities for the TrueSkill, Gaussian-OD, and Gaussian-SD models are computed in the way that they can maximize the WLD prediction accuracies, contrasted with the binary prediction Win/no-Win probabilities as reported in [17], leading to noticeable differences in results for the information gain criterion.

**Table 1** Parameter settings. Priors on offence/defence skills: $\mathcal{N}(\mu_0, \sigma_0^2)$ with $\mu_0 = 25$ and $\sigma_0 = 25/3$. Performance variance: $\beta$, $\beta_o$, $\beta_d$.

| Model | Parameter ($\epsilon, \gamma$ empirically estimated) |
|---|---|
| TrueSkill | $\beta = \sigma_0/2$, $\epsilon$: draw probability |
| Poisson-OD(VB) | $\beta_o = \beta_d = \sigma_0/2$ |
| Poisson-OD(Sampling) | $\beta_o = \beta_d = \sigma_0/2$ |
| Gaussian-OD | $\beta_o = \beta_d = \sigma_0/2$, $\gamma$: score variance |
| Gaussian-SD | $\beta = \sigma_0/2$, $\gamma$: score difference variance |

### *5.3.1 Information Gain*

Information gain for four models on the UK data set is shown in Figure 7 (Top Panel). The results indicate that the proposed Gaussian-OD, Gaussian-SD, and Gaussian-OD-AH performed significantly better than TrueSkill for all the training proportions. These results indicated that score information was indeed useful for learning teams' skill levels when training data is limited. As shows in the results, the Gaussian-OD-AH slightly outperformed Gaussian-OD when 30% of the data was used for training, indicating the meaningful information carried by the home/away feild for a football game. We also observed that the Poisson-OD(VB) and Poisson-OD(VB)-AH did not perform as good as other models. This is because that the scores for the UK data set are relatively small; thus the amplification due to the exponential term in the Poisson-OD model led to more extreme probaiblities, thus may hurt the performance if the predictions are wrong.

Results on the Halo data set are shown in Figure 7 (Middle Panel), again the Gaussian-OD and the Gaussian-SD models significantly outperformed the TrueSkill. Note that there is no notion of home/away field for the Halo data set, thus we omit the Gaussian-OD-AH and the Poisson-OD-AH models for this data set. We also omit reporting performance for the Poisson-OD model with slice sampling as it is comptutationally demanding to obtain the results.

For the AFL data set, the information gain for different models shown in Figure 7 (Bottom Panel) indicated that the Gaussian likelihood based models significantly perform the rest for all training settings, outperforming TrueSkill and the Poisson likelihood based on models. Regarding the Poisson models, the home advantage Poisson model Poisson-OD(VB)-AH performed the worst for small training data, and started to slightly edge out the Poisson-OD(VB) model for more training data. This may be caused by the fact that the Poisson-OD(VB)-AH model is equipped with the additional home advantage variable for each team, thus requiring more data to refine the uncertainty associated with this latent variable. Finally, we observed that the variational Bayes and the sampling method for the Poisson-OD model achieved comparable performance.

**Fig. 7** Results on the UK-PL (upper), Halo (middle), and AFL (bottom), evaluated using information gain. Error bars indicate 95% confidence intervals.

### *5.3.2 Win/Lose/Draw Prediction Accuracy*

We further studied the performances of various models under a multi-class classification setting of predicting WLD for a match between two teams $(i, j)$. For the results on the UK data set shown in Figure 8 (Top panel), the best performing model is the Gaussian-OD-AH, which achieved significantly better performance than that for the Gaussian-OD, indicating the importance of introducing the home field advantage variable. Together with the Gaussian-SD model, these three models significantly outperform TrueSkill for all the training settings. For limited training data, the proposed Poisson-OD and Poisson-OD-VB-AH models outperform TrueSkill as well, indicating the importance of modeling scores.

Results on the Halo data set Figure 8 (Middle panel) again demonstrated that the proposed models obtained much better WLD predictions comparing with TrueSkill. For this data set, the Gaussian models outperform the Poisson-OD model, which was not surprising as the average score for the Halo data set is relatively small numbers. Note that the performances for the Gaussian models were comparable with each other, which suggested that it may not be beneficial to model offence/defence skills separately for applications where the notions of offence and defence are not clear, which is the case for the Halo data but not for the AFL and UK data sets.

For the results on the AFL data set in Figure 8 (Bottom panel), we observed that the Poisson-OD, Gaussian-SD, Gaussian-OD, and Poisson-OD(VB)-AH all significantly outperformed TrueSkill, which further supported the importance of learning team skills by taking into account the score information. For the Poisson likelihood based models, we observed that when training data was limited, the introduction of the home field advantage variable actually hurt the performance, as shown in Figure 8 (Bottom panel) when less or equal to 40% of data was used for training. But with the increasing amount of training data, the refined home advantage variable can help to achieve better performance compared with the Poisson-OD(VB), indicating the marginal benefits of modeling home field advantages. Another observation was that the Possion-OD(VB) achieved the comparable performance with that of the Poisson-OD(Sampling).

When comparing the best performance across the three data sets for all the models, we note that the mean accuracies for all the models on the AFL data set were the largest than that for the UK data sets, perhaps indicating that football games may be much harder than that the rugby games for AFL. The difficulty in predicting win/lose/draw for football matches is perhaps caused by the unexpected player injuries, teams trading players, etc. One may argue that these factors apply to the AFL data set, but it seems that the AFL has somehow stricter rules in regulating the trading of players.

**Fig. 8** Results on the UK-PL (Top), Halo (Middle), and AFL (Bottom), evaluated using the multiclass WLD prediction accuracy. Error bars indicate 95% confidence intervals.

### *5.3.3 Log likelihood*

For the log likelihood on UK data set (Figure 9 Top Panel), we observed that the TrueSkill performed the best across all models for limited data, however, the Poisson models edge out all other models when more than 30% of data was used for training. For the Gaussian models, the Gaussian-OD and Gaussian-SD models performed comparably with each other.

For the log-likelihood of the Halo data set show in Figure 9 (Middle Panel), we observed that the Poisson-OD(VB), Gaussian-OD, and Gaussian-SD, outperformed the TrueSkill. Another observation was that the Gaussian-SD model significantly outperformed the Gaussian-OD model, which was not surprising because there were no explicit connections with a player's offence and defence skills for the Halo data set, contrasted with that for the AFL and UK data sets.

Results on the AFL data set was shown in Figure 9 (Bottom Panel). We observed that the Gaussian-OD-AH, Gaussian-SD, and Gaussian-OD models were the best performing ones comparing with the rest. Amongst these three models, they were comparable with each other. For the Poisson-OD(VB)-AH, it initially performed worse than the Poisson-OD(VB) model, but obtained slightly better performance when more data was used for training. Finally, we observed that there was no distinctions between the Poisson-OD(VB) and Poisson-OD(Sampling).

**Fig. 9** Results on the UK-PL (upper), Halo (middle), and AFL (bottom), evaluated using log likelihood. Error bars indicate 95% confidence intervals.

### *5.3.4 Score Prediction Errors*

We reported the score prediction errors for different data sets in Figure 10. For the UK data set (Figure 10 Top Panel), the Poisson-OD(VB) and Poisson-OD(VB)-AH model clearly did not work well in predicting this type of matches, which are characterized by low match scores (average scores being 1.3 for the UK data set). But the Gaussian-OD and Gaussian-OD-AH models significantly outperformed the baseline for most of the training settings.

For the results on the Halo dataset (Figure 10 Middle Panel), we observed that the proposed Poisson-OD and Gaussian-OD models significantly outperformed the baseline. This demonstrated that our proposed model can make sensible score predictions for online games as well. When comparing between the Poisson-OD and Gaussian models, it is clear that the Gaussian-OD model achieves much better performance.

For the AFL data set in Figure 10 (Bottom Panel), we observed that the Gaussian-OD and Gaussian-OD-AH model clearly failed in predicting the scores of the matches, because the differences in the learned skill levels between teams are relatively small. Thus, a Gaussian likelihood function with the small difference has low probability of generating huge match scores for the AFL data set. Given the same difference, the Poisson-OD and Poisson-OD-AH models with an exponentiated scoring rate would seem to amplify these small performance differences in learned AFL skills to make more accurate score predictions on AFL data comparing with the baseline. This amplification appears to hurt the Poisson-OD model on the lower-scoring UK-PL (the mean score for the AFL data is 95.4 vs 42.7 and 1.3 respectively for the Halo 2 and UK-PL data) as shown in Figure 10 (Top Panel). Finally, we note that the Poisson-OD(VB) and the Poisson-OD(Sampling) performed comparably with each other.

We want to emphasize that for the AFL data set, we observe that the baseline method based on average scores of each team makes better predictions than the Poisson model for the single training setting where 10% of data was used for training. This was not surprising because the belief associated with the Poisson-OD(VB), Poisson-OD(VB)-AH, and Poisson-OD(Sampling) model on team skills exhibited large uncertainty when only a few observations were used for training. As we can see when more data was used for training, all of the Poisson likelihood based models performed significantly better than the baseline.

**Fig. 10** Results on the UK-PL (upper), Halo (middle), and AFL (bottom), evaluated using the score prediction error in term of MAE. Error bars indicate 95% confidence intervals.

### 5.4 Variational Inference Vs. Sampling for the Poisson-OD Model

For the proposed Poisson-OD model, we studied the performance and efficiency for Bayesian inference based on the proposed variational Bayes, against one sampling method named slice sampling. For both the UK-PL and AFL data sets, we show the performance of the slice sampling for all evaluation criteria (see the Top and Bottom Panels of Figure 7 8 9 10). We observed that the sampling approach only slightly performed better than the variational Bayes for a few training settings for the UK-PL data set, but the differences were not significant; for most of the cases, these two approaches achieved comparable performances.

Despite the similar performance between the variational Bayes and the slice sampling method for the Poisson-OD model, the slice sampling method required much longer time to obtain stationary samples for computing the statistics. Our empirical experiments indicated that the slice sampling took about 10 seconds when performing belief updating for an observation; however, the proposed fixed-point solution for the variational Bayes often converged after two or three iterations, which can be finished within less than 0.01 seconds, leading to 1000 times gain in computational efficiency. The gain in speed is extremely important for conducting online Bayesian skill learning for real-world matchmaking systems. Note that all the experiments were conducted on a laptop with an Intel i5 CPU, 4G memory, and codes are implemented in Matlab. Due to the high computational requirements for the sampling method, we omit the report of its performance for other data sets. Note that one can perhaps explore the advanced sampling approaches such as [18] to improve computational efficiency for the sampling method.

### 5.5 Model Home Field Improving Performance

Home team advantages for many games particularly the football games has been encoded in different ways to improve match outcome prediction accuracy. To validate this known fact, we took the UK and AFL data sets where there is the notion of home field advantages. For the UK data set, we computed the probability that the home teams win the away teams in the following way. For each of the 41 teams of the UK data set, we first computed their win probability when playing on home fields, took the average win probability, and obtained its mean being 0.5665 with the standard deviation is 0.1511. Given this probability marginally larger than 0.5 with the large standard deviation, it is not clear if the home advantage can indeed help to predict WLD based match outcomes. This is reflected in the empirical evaluations for the four criteria reported in the previous sections, where we observed that the models with the home field advantages performed better for limited cases.

For the AFL data set, we conducted the same analysis, and obtained an average win probability across the 16 teams being 0.5907 (with the standard deviation being 0.1381), larger than that for the UK data set. Referring back to our empirical evaluations for the AFL data set in the bottom panels of Figure 7 8 9, we indeed observed that the modeling of home field advantages led to the slightly better performance, particularly for the Poisson-OD model.

## 6 Related Work

**Skill rating** dates at least as far back as the Elo system [6], the idea of which is to model the probability of the possible game outcome as a function of the two players' skill levels. Players' skill levels are updated after each game in a way such that the observed game outcome becomes more likely and the summation of players' ratings remains unchanged.

The Elo system cannot handle the case when three or more teams participate in one match, a disadvantage addressed by TrueSkill [7]. Further extensions of TrueSkill incorporate time-dependent skill modeling for historical data [5].

In [4], the authors model and learn the correlation between all players' skills when updating skill beliefs, and develop a method called "EP-Correlated", contrasted with the independent assumption on players' skills (EP-Independent). Empirically, EP-Correlated outperforms EP-Independent on professional tennis match results; this suggests modeling correlations in extensions of the score-based learning presented here.

These skill learning methods all share a common feature that they are restricted to model WLD only and have to discard meaningful information carried with scores. While we proposed score-based extensions of TrueSkill in this work; it remains to incorporate other extensions motivated by this related work.

**Score modeling** has been studied since the 1950s [15] [16] [11] [14] [13]; one of the most popular score models is the Poisson model, first presented in [15], and this work continues to the present [13]. Other commonly used score models are based on normal distributions [11]. However, it appears that most score-based models do not distinguish offence and defence skills of each team and the results here indicate that such separate offence/defence skill models can perform better than univariate models with limited data.

More recently, [2] introduced a log-linear random effect model to model the number of goals for a football match, which takes into account home field advantages and distinguish teams' attack and defense skills, and proposed a Bayesian hierarchical model to generate the match outcomes in terms of scores. Inference in the model is conducted by MCMC, which can be slow as discussed in Section 5.4.

## 7 Conclusion

We proposed novel score-based, online Bayesian skill learning extensions of TrueSkill that modeled (1) player's offence and defence skills separately, (2) how these offence and defence skills interact to generate scores, and (3) home field advantages. Overall these new models — and Gaussian-OD (using a separate offence/defence skill model) in particular — show an often improved ability to model winning probability and win/loss prediction accuracy over TrueSkill, especially when the amount of training data is limited. This indicates that there is indeed useful information in score-based outcomes that is ignored by TrueSkill and that separate offence/defence skill modeling does help (c.f. the performance of Gaussian-OD vs. Gaussian-SD). The introduction of home field advantages to the models also show improved performance for some of the settings. Furthermore, these new models allow the prediction of scores (unlike TrueSkill), with the Poisson-OD model and its fast variational Bayesian update derived in Section 4.2 performing best on the high-scoring AFL data for predicting scores. For the Poisson-OD model, we also provided a sampling based Bayesian inference approach, but results indicated that the proposed Poisson-OD variational Bayes was 1000x faster than sampling but almost always performed just as well. Altogether, these results suggested the potential advantages of score-based Bayesian skill learning over state-of-the-art WLD-based skill learning approaches like TrueSkill.

Future research could combine the proposed models with related work that models time-dependent skills, multi-team games, and correlated skills to utilise score-based outcomes.

**Acknowledgements** We thank Marconi Barbosa, Guillaume Bouchard, David Stern and Onno Zoeter for interesting discussions, and we also thank the anonymous reviewers at ECML-PKDD'12 for their constructive comments, which help to improve the paper. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

## References

1. R. M. Neal *Slice sampling.* The Annals of Statistics, 31(3):705-767, 2003.
2. G. Baio and M. A. Blangiardo *Bayesian hierarchical model for the prediction of football results.* JOURNAL OF APPLIED STATISTICS, 37(2):253-264, 2010.
3. M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables.* Dover Publications, New York, 1974.
4. A. Birlutiu and T. Heskes. Expectation propagation for rating players in sports competitions. In *ECML-PKDD*, volume 4702 of *LNCS*, pages 374–381. Springer, 2007.
5. P. Dangauthier, R. Herbrich, T. Minka, and T. Graepel. Trueskill through time: Revisiting the history of chess. In *NIPS*, pages 337–344. MIT Press, Cambridge, MA, 2008.
6. A. E. Elo. *The rating of chess players: past and present.* Arco Publishing, New York, 1978.
7. R. Herbrich, T. Minka, and T. Graepel. Trueskill$^{TM}$: A Bayesian skill rating system. In *NIPS*, pages 569–576, 2006.

8. F. R. Kschischang, B. J. Frey, and H. A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, February 2001.
9. T. Minka. Expectation propagation for approximate bayesian inference. In *UAI*, pages 362–369. Morgan Kaufmann, 2001.
10. J. G. Skellam. The frequency distribution of the difference between two Poisson variates belonging to different populations. *Journal of the Royal Statistical Society: Series A*, 109(3):296, 1946.
11. M. E. Glickman, and H. S. Stern. A state-space model for football league scores. *Journal of the American Statistical Association*, 93(441):25–35, 1998.
12. M. J. Beal, and Z. Ghahramani. The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures. *Proceedings of the Seventh Valencia International Meeting*: 453–464, 2002.
13. D. Karlis and I. Ntzoufras, Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference, *IMA Journal of Management Mathematics*, 20(2):133–145, 2009.
14. D. Karlis and I. Ntzoufras, Analysis of Sports Data by Using Bivariate Poisson Models, *Journal of the Royal Statistical Society: Series D*, 52(3):381-393, 2003.
15. M. J. Moroney Facts from figures, Penguin Press Science, 3rd, 1956.
16. M. J. Dixon and S. G. Coles, Modelling Association Football Scores and Inefficiencies in the Football Betting Market, *Journal of the Royal Statistical Society: Series C*, 46(2):265–280, 1997.
17. S. Guo and S. Sanner and T. Graepel and W. Buntine *Score-based Bayesian Skill Learning*. Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 106-121, 2012
18. I. Murray and R. P. Adams and D. J.C. MacKay *Elliptical slice sampling*. Proceedings of the 13th International Conference on Artificial Intelligence and Statistics, 541-548, 2010