Proceedings of the 2011 Winter Simulation Conference
S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, eds.

MAY THE BEST MAN WIN: SIMULATION OPTIMIZATION FOR MATCH-MAKING IN E-SPORTS

Ilya O. Ryzhov

Robert H. Smith School of Business
University of Maryland
College Park, MD 20742, USA

Awais Tariq
Warren B. Powell

Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08544, USA

ABSTRACT

We consider the problem of automated match-making in a competitive online gaming service. Large numbers of players log on to the service and indicate their availability. The system must then find an opponent for each player, with the objective of creating competitive, challenging games that do not heavily favour either side, for as many players as possible. Existing mathematical models for this problem assume that each player has a skill level that is unknown to the game master. As more games are played, the game master's belief about player skills evolves according to a Bayesian learning model, allowing the game master to adaptively improve the quality of future games as information is being collected. We propose a new decision-making policy in this setting, based on the knowledge gradient concept from the literature on optimal learning. We conduct simulations to demonstrate the potential of this policy.

1 INTRODUCTION

With the appearance of affordable broadband Internet, competitive online gaming or e-sports has become a massive cultural phenomenon, as well as a profitable business venture. A study by Huhh (2008) documents the growth in revenues of South Korean online game company NCSoft, from \$559 million in 2000 to over \$2 billion in 2004. In 2005, Microsoft's Xbox Live online service had over 2 million subscribers (Herbrich et al. 2006). The competitive strategy game Starcraft II, released in 2010, has over 2.5 million players listed in its ranking system (SC2 Rankings 2011) as of this writing.

Competition is at the heart of e-sports. Online game services frequently create and display rankings of players (as seen above for the case of Starcraft II). Outside organizations create their own rankings (KeSPA 2011), used to evaluate professional players. Ranking has a great impact on the experience of even casual players. Large online systems automate the process of match-making: a player logs on to the system and submits a request to play a game, whereupon the system finds an opponent with no further input from the player. The goal is to create fair and challenging games by matching players of similar skill level.

However, a player's skill level is not known exactly to the system, and must be inferred from the player's match history. Methods for statistical modeling of player skills and prediction of game outcomes date back to the work by Elo (1978), which focuses on rating chess players. The emergence of e-sports has sparked a new interest in such methods. Studies by Herbrich et al. (2006) and Dangauthier et al. (2007) adopt a Bayesian perspective for modeling player skills. The game master has a Bayesian belief about the skill of each player, and adjusts this belief as new games are played. Furthermore, the Bayesian beliefs can be used to estimate the quality of a particular match-up, enabling the game master to make match-making decisions. The resulting TrueSkill$^{\mathrm{TM}}$ ranking system has been implemented by Xbox Live.

The match-making problem can be interpreted as a variation on ranking and selection, a fundamental problem in simulation optimization (see e.g., Swisher et al. 2000 or Hong and Nelson 2009 for an introductory overview, or Kim and Nelson 2007 for a view geared more toward recent advances). In

978-1-4577-2109-0/11/\$26.00 ©2011 IEEE

ranking and selection, there is a finite set of alternatives with unknown values (e.g., different settings for a simulator) that can be estimated through expensive simulations. The objective is to efficiently allocate the simulation budget in order to discover the alternative with the highest value. In the match-making problem, individual players can also be viewed as alternatives with unknown values (skill levels). However, the goal is not to discover the single most skilled player, but rather to make each game as even and competitive as possible, thus falling outside the purview of traditional simulation optimization.

The dimension of simulation optimization was not considered in the original work on TrueSkill$^{\text{TM}}$. Herbrich et al. (2006) proposes a Bayesian model for learning, as well as a criterion for evaluating games, but uses a simple greedy algorithm for the actual match-making decisions. The optimal learning literature observes that we can obtain much better results with some experimentation or trial and error; experimental comparisons of such procedures against greedy algorithms can be found, e.g., in Frazier et al. (2008) for ranking and selection, or in Ryzhov and Powell (2009b) for the closely related multi-armed bandit problem. Our paper contributes a new decision rule for match-making.

We apply a class of optimal learning methods known variously as value of information procedures (see Chick and Inoue 2001a, Chick and Inoue 2001b, or Chick 2006) and knowledge gradient policies (Frazier et al. 2008). Our Bayesian belief about player skill induces a probability distribution on the outcome of the next game, and thus, on the future beliefs that we will use to make future decisions. By taking an expectation over this distribution, we can "look into the future," creating an estimate of how much the information we collect now will benefit us in future games. Some recent, relevant applications of look-ahead policies include energy portfolio selection (Ryzhov and Powell 2009a), two-agent newsvendor problems (Ryzhov et al. 2010), and simulation calibration (Frazier et al. 2009, Scott et al. 2010).

In Section 2, we survey TrueSkill$^{\text{TM}}$ and present a derivation of the draw probability criterion for game quality that was not given in the original paper. Section 3 shows how game outcomes can be predicted and derives the KG policy. Section 4 presents simulations demonstrating the potential of KG for improving match-making decisions. Finally, Section 5 concludes.

2 BAYESIAN MODEL FOR PLAYER SKILLS

We approach the problem from the point of view of one fixed player, referred to as "player 0." We are motivated by a situation where a player has logged onto the system and submitted a request to play; we must then do our best to satisfy the request and create a good game specifically for that player. The broader problem of simultaneously optimizing across all players is a subject for future work, but we believe that the concepts put forth in this paper can also be useful in such general settings.

We assume that player 0 has a fixed pool of potential opponents, players $1,...,M$, and that these opponents only play against player 0, and only one of them can play per time period. The others simply wait for their next chance to be selected. Of course, in a real gaming environment, the pool of available opponents is constantly changing, and many games occur simultaneously. However, if the total number of players is large enough, we might reasonably assume that, at any given time when player 0 is available, there is a roughly constant number of available opponents to choose from, and they come from the same statistical population, with roughly the same variation in skill level.

We briefly summarize the TrueSkill$^{\text{TM}}$ model of Herbrich et al. (2006). Each player $i = 0,1,...,M$ has an underlying *skill level* s_i. This dimensionless quantity is unknown to the game master. Our uncertainty about s_i is encoded by the Bayesian modeling assumption $s_i \sim \mathcal{N}\left(\mu_i^0, \left(\sigma_i^0\right)^2\right)$.

When player i plays a game, his or her performance is represented as a random variable $p_i \sim \mathcal{N}\left(s_i, \sigma_\varepsilon^2\right)$, a black-box sample of skill level. For simplicity, the variance σ_ε^2 is assumed to be known and constant for all players. Note that player i's performance does not depend on the skill of the opponent. In fact, we assume that the true skill levels s_i are mutually independent for all $i = 0,1,...,M$. Player i is said to win a game against player j if $p_i > p_j$.

If we had the ability to observe the exact value of p_i in a game played by player i, we would have a classic conjugate prior model. Then, given the results of the first n games, the posterior distribution of s_i would be normal, with a standard set of recursive updating equations (DeGroot 1970) for obtaining μ_i^{n+1} and σ_i^{n+1} from μ_i^n, σ_i^n and p_i^{n+1}. However, the match-making problem is complicated by the fact that we have no way of observing the exact values p_i. If player i plays against j, we will only observe who wins, that is, whether $p_i > p_j$ or $p_j < p_i$. (The probability of a draw is zero.) The posterior density

$$P(s_i \in ds \,|\, p_i > p_j) = \frac{P(p_i > p_j \,|\, s_i = s)\, P(s_i \in ds)}{P(p_i > p_j)}$$

is no longer normal. The work by Dangauthier et al. (2007) uses approximate Bayesian inference to derive the updating equations

$$\mu_i^{n+1} = \begin{cases} \mu_i^n + \frac{(\sigma_i^n)^2}{(\sigma_i^n)^2 + (\sigma_j^n)^2 + 2\sigma_\varepsilon^2} \cdot v\left(\frac{\mu_i^n - \mu_j^n}{(\sigma_i^n)^2 + (\sigma_j^n)^2 + 2\sigma_\varepsilon^2}\right) & \text{if } p_i^{n+1} > p_j^{n+1}, \\ \mu_i^n - \frac{(\sigma_i^n)^2}{(\sigma_i^n)^2 + (\sigma_j^n)^2 + 2\sigma_\varepsilon^2} \cdot v\left(\frac{\mu_j^n - \mu_i^n}{(\sigma_i^n)^2 + (\sigma_j^n)^2 + 2\sigma_\varepsilon^2}\right) & \text{if } p_i^{n+1} < p_j^{n+1}, \end{cases} \tag{1}$$

and

$$\left(\sigma_i^{n+1}\right)^2 = \begin{cases} (\sigma_i^n)^2 \left(1 - \frac{(\sigma_i^n)^2}{(\sigma_i^n)^2 + (\sigma_j^n)^2 + 2\sigma_\varepsilon^2} \cdot w\left(\frac{\mu_i^n - \mu_j^n}{(\sigma_i^n)^2 + (\sigma_j^n)^2 + 2\sigma_\varepsilon^2}\right)\right) & \text{if } p_i^{n+1} > p_j^{n+1}, \\ (\sigma_i^n)^2 \left(1 - \frac{(\sigma_i^n)^2}{(\sigma_i^n)^2 + (\sigma_j^n)^2 + 2\sigma_\varepsilon^2} \cdot w\left(\frac{\mu_j^n - \mu_i^n}{(\sigma_i^n)^2 + (\sigma_j^n)^2 + 2\sigma_\varepsilon^2}\right)\right) & \text{if } p_i^{n+1} < p_j^{n+1}, \end{cases} \tag{2}$$

where

$$\begin{aligned} v(x) &= \frac{\phi(x)}{\Phi(x)}, \\ w(x) &= v(x)\,(v(x) + x), \end{aligned}$$

with ϕ, Φ being the standard normal pdf and cdf. We force the posterior distribution to be normal, and choose the parameters of that normal distribution to resemble the actual posterior as much as possible. We use the notation $P^n(\cdot)$ and $\mathbb{E}^n(\cdot)$ to denote the conditional probability of an event, or the conditional expectation of a random variable, given our observations from the first n games ("at time n"). Because we use a normal distribution to approximate the posterior, we also calculate conditional probabilities and expectations approximately, based on this normality assumption.

Due to the normality assumption, our beliefs about all the players at time n can be encoded in the *knowledge state*

$$k^n = \{\mu_i^n, \sigma_i^n \,|\, i = 0, 1, ..., M\}.$$

2.1 Draw Probability for Evaluating Games

Because p_i and p_j are normally distributed, the probability that a draw occurs (that is, the opponents perform equally well) is zero. In fact, draws do not occur in many competitive online games. Microsoft's Halo does not allow them to occur, whereas Starcraft II has the functionality to call a draw in certain cases, but in practice this happens rarely, particularly in a competitive setting. Nonetheless, we use an artificial notion of a draw probability to help us create good games. It is logical to suppose that a game is fair and even if the performances of the two players are more likely to be close together.

Let $\delta > 0$. In Dangauthier et al. (2007), a game between players i and j ends in a draw if

$$|p_i - p_j| < \delta.$$

Suppose that we are at time n, that is, n games have occurred. If the next game is between players i and j, the conditional probability of a draw is

$$P^n\left(\left|p_i^{n+1}-p_j^{n+1}\right|<\delta\right)=\mathbb{E}^n P^n\left(\left|p_i^{n+1}-p_j^{n+1}\right|<\delta\,|\,s_i,s_j\right). \tag{3}$$

The outer expectation is over our distribution of belief at time n. The work by Herbrich et al. (2006) suggests letting $\delta \to 0$, whence (3) becomes

$$\begin{aligned}\lim_{\delta\to 0} P^n\left(\left|p_i^{n+1}-p_j^{n+1}\right|<\delta\right) &= \lim_{\delta\to 0}\mathbb{E}^n P^n\left(\left|p_i^{n+1}-p_j^{n+1}\right|<\delta\,|\,s_i,s_j\right)\\ &= \mathbb{E}^n \lim_{\delta\to 0} P^n\left(-\delta<p_i^{n+1}-p_j^{n+1}<\delta\,|\,s_i,s_j\right)\end{aligned}$$

. The probability that $p_i^{n+1}-p_j^{n+1}$ will fall in an infinitesimally small interval centered around zero is equal to the density of $p_i^{n+1}-p_j^{n+1}$, evaluated at zero, multiplied by the infinitesimal interval width. The distribution of $p_i^{n+1}-p_j^{n+1}$ is $\mathcal{N}\left(s_i-s_j,2\sigma_\varepsilon^2\right)$ because performances are normal and independent. Thus,

$$\mathbb{E}^n \lim_{\delta\to 0} P^n\left(-\delta<p_i^{n+1}-p_j^{n+1}<\delta\,|\,s_i,s_j\right)=\mathbb{E}^n\left(\frac{1}{\sqrt{4\pi\sigma_\varepsilon^2}}e^{-\frac{\left(s_i-s_j\right)^2}{4\sigma_\varepsilon^2}}\right)\lim_{\delta\to 0}2\delta. \tag{4}$$

As $\delta \to 0$, the right-hand side of (4) always goes to zero as well. However, different match-ups can be compared by how quickly this occurs. The term inside the expectation in (4) represents the relative likelihood of a draw for a match between i and j. We show in Appendix A that

$$\mathbb{E}^n\left(\frac{1}{\sqrt{4\pi\sigma_\varepsilon^2}}e^{-\frac{\left(s_i-s_j\right)^2}{4\sigma_\varepsilon^2}}\right)=\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\left(\sigma_i^n\right)^2+\left(\sigma_j^n\right)^2+2\sigma_\varepsilon^2}}e^{-\frac{\left(\mu_i^n-\mu_j^n\right)^2}{2\left(\left(\sigma_i^n\right)^2+\left(\sigma_j^n\right)^2+2\sigma_\varepsilon^2\right)}}. \tag{5}$$

Let q_{ij}^n denote the right-hand side of (5). In Herbrich et al. (2006), a normalizing factor is applied, yielding an expression

$$\tilde{q}_{ij}^n=\sqrt{\frac{2\sigma_\varepsilon^2}{\left(\sigma_i^n\right)^2+\left(\sigma_j^n\right)^2+2\sigma_\varepsilon^2}}e^{-\frac{\left(\mu_i^n-\mu_j^n\right)^2}{2\left(\left(\sigma_i^n\right)^2+\left(\sigma_j^n\right)^2+2\sigma_\varepsilon^2\right)}}. \tag{6}$$

The normalizing factor does not depend on i or j, and thus does not affect which pair (i,j) has the highest draw probability. The quantity $\tilde{q}_{ij}^n$ is used by Herbrich et al. (2006) as a criterion for the quality of a game. A higher draw probability indicates that a particular match-up is fairer and more competitive.

2.2 Objective Function

We can now use the draw probability to formulate an objective function for the game master. In e-sports, learning occurs online. That is, we learn about the players' skill levels in real time, while games are being played. We wish to learn about the players' true skills in order to select the best opponent, but we should also try to ensure high-quality match-ups in every game, or at least as many individual games as possible. We will choose an opponent for player 0 using a decision rule X^π, a function mapping a knowledge state k^n to some $X^\pi\left(k^n\right)\in\{1,...,M\}$. The notation π describes the policy represented by the decision rule X^π. Our objective is to choose a policy π for selecting opponents in order to maximize

$$\sup_\pi \mathbb{E}^\pi \sum_{n=0}^{N} q_{0,X^\pi(k^n)}^n, \tag{7}$$

the expected number of "draws" across all N games.

We give two simple examples of policies, before proposing our own policy in Section 3. A simple myopic policy for making decisions ("point-estimate" or PE) might be

$$X^{PE}\left(k^n\right) = \arg\min_j \left|\mu_0^n - \mu_j^n\right|. \tag{8}$$

This policy is based purely on the point estimates μ_i^n of s_i, with no regard for the uncertainty in those estimates. A second policy, which we call DrawChance, takes uncertainty into account by maximizing

$$X^{DC}\left(k^n\right) = \arg\max_j \tilde{q}^n_{0,j}. \tag{9}$$

where $\tilde{q}^n_{0,j}$ is the normalized draw probability from (6). This is also the policy used by Herbrich et al. (2006) to make decisions.

3 A KNOWLEDGE GRADIENT POLICY FOR MATCH-MAKING

We propose to improve on the DrawChance policy by using the potential change in our estimate of the draw probability as a factor in our decision-making. For each possible opponent, we will look ahead to the expected results of the next game and consider how our new estimate of the draw probability will affect the quality of future games. This approach is known as a knowledge gradient policy, previously considered by Gupta and Miescke (1996) and Frazier et al. (2008) for the ranking and selection problem, and by Ryzhov et al. (2011) for its online analog, the multi-armed bandit problem.

In the problem under discussion, the look-ahead is relatively simple because the observation is binary. The following result computes our belief about the probability of each outcome for a game played by players i and j.

Proposition 1 Under the normality assumption, the conditional probability that player i wins against player j, given the knowledge state k^n, is given by

$$P^n\left(p_i^{n+1} > p_j^{n+1}\right) = \Phi\left(\frac{\mu_i^n - \mu_j^n}{\left(\sigma_i^n\right)^2 + \left(\sigma_j^n\right)^2 + 2\sigma_\varepsilon^2}\right). \tag{10}$$

Proof: We compute the winning probability as

$$\begin{aligned} P^n\left(p_i^{n+1} > p_j^{n+1}\right) &= \mathbb{E}^n P^n\left(p_i^{n+1} - p_j^{n+1} > 0 \,|\, s_i, s_j\right) \\ &= \mathbb{E}^n \Phi\left(\frac{s_i - s_j}{\sqrt{2\sigma_\varepsilon^2}}\right) \\ &= \int_{-\infty}^{\infty} \Phi\left(\frac{x}{\sqrt{2\sigma_\varepsilon^2}}\right) \frac{1}{\sqrt{2\pi\left(\left(\sigma_i^n\right)^2 + \left(\sigma_j^n\right)^2\right)}} e^{-\frac{\left(x - \left(\mu_i - \mu_j\right)\right)^2}{2\left(\left(\sigma_i^n\right)^2 + \left(\sigma_j^n\right)^2\right)}} dx. \end{aligned} \tag{11}$$

We can rewrite (11) as $P(X \leq Y)$, where $X \sim \mathcal{N}\left(0, 2\sigma_\varepsilon^2\right)$ and $Y \sim \mathcal{N}\left(\mu_i - \mu_j, \left(\sigma_i^n\right)^2 + \left(\sigma_j^n\right)^2\right)$ are independent. Since both X and Y are normal, this probability is

$$P(X - Y \leq 0) = \Phi\left(\frac{\mu_i - \mu_j}{\left(\sigma_i^n\right)^2 + \left(\sigma_j^n\right)^2 + 2\sigma_\varepsilon^2}\right),$$

as required. □

As a consequence, the losing probability has the simple form

$$P^n\left(p_i^{n+1} < p_j^{n+1}\right) = 1 - \Phi\left(\frac{\mu_i^n - \mu_j^n}{\left(\sigma_i^n\right)^2 + \left(\sigma_j^n\right)^2 + 2\sigma_\varepsilon^2}\right) = \Phi\left(\frac{\mu_j^n - \mu_i^n}{\left(\sigma_i^n\right)^2 + \left(\sigma_j^n\right)^2 + 2\sigma_\varepsilon^2}\right).$$

For each possible outcome of the game, our new beliefs can easily be computed using (1) and (2). Suppose that, at time n, we choose player j to be the opponent in the $(n+1)$st game. Let $k^{w,n+1}$ denote the knowledge state that we would have at time $n+1$ if player 0 were to win the game. Similarly, let $k^{l,n+1}$ denote the new knowledge state in the event that player 1 loses. We can also define $q_{ij}^{w,n+1}$ and $q_{ij}^{l,n+1}$ to be the new estimates of the draw probability between i and j that are based on $k^{w,n+1}$ and $k^{l,n+1}$, respectively.

Let

$$F^{w,n+1} = \max_j q_{0,j}^{w,n+1}, \qquad F^{l,n+1} = \max_j q_{0,j}^{l,n+1}$$

be our time-$(n+1)$ estimates of the best possible draw probability, for each possible outcome of the game at time n. Then, the conditional expectation of our future estimate of the biggest draw probability, given our time-n beliefs k^n and our choice of player j at time n, is given by

$$F_j^n = \Phi\left(\frac{\mu_0^n - \mu_j^n}{\left(\sigma_0^n\right)^2 + \left(\sigma_j^n\right)^2 + 2\sigma_\varepsilon^2}\right) F^{w,n+1} + \Phi\left(\frac{\mu_j^n - \mu_0^n}{\left(\sigma_0^n\right)^2 + \left(\sigma_j^n\right)^2 + 2\sigma_\varepsilon^2}\right) F^{l,n+1}.$$

We weigh the value for each outcome by the probability of that outcome, as computed in Proposition 1.

Finally, the knowledge gradient (KG) policy chooses an opponent at time n according to the rule

$$X^{KG,n}\left(k^n\right) = \arg\max_j q_{0,j}^n + (N-n)\,F_j^n. \tag{12}$$

The policy only considers the change in our beliefs that results from the very next game. In other words, the knowledge gradient method assumes that our knowledge state will change only one more time, from k^n to k^{n+1}, and from that point on, we will stop learning and $k^{n'} = k^{n+1}$ for all $n' \geq n+1$. Under this assumption, the very next decision at time n is the last decision that will provide any new information. The KG method allocates this decision optimally.

The value of each game in (7) is represented by the draw probability of that game. If we stop learning at time $n+1$, the best possible decision for each game from time $n+1$ onward is to choose the opponent that maximizes $q_{0,j}^{n+1}$. The estimated value of that decision is, accordingly, $\max_j q_{0,j}^{n+1}$. The time-n expectation of that value is precisely F^n. The multiplier $N-n$ is the number of games remaining after time n. This reflects the online nature of the problem: in (7), we collect a value for each individual game.

In a real-world setting, the available opponents at time $n+1$ may be different from the ones at time n. However, we might still use KG as a heuristic, and make the one-step look-ahead decision under the simplifying assumption that the same opponents would be present in the next time step. To expedite computation, we may wish to narrow down the set of possible opponents before computing (12), perhaps with a technique like the one discussed by Ryzhov and Powell (2009a).

Additionally, a real gaming service would have no way of knowing the total number N of games that player 0 intends to play. We propose a simple heuristic modification, based on the infinite-horizon version of the online KG method from Ryzhov et al. (2011). We fix a value $0 < \gamma < 1$ and use the decision rule

$$X^{KG}\left(k^n\right) = \arg\max_j q_{0,j}^n + \frac{\gamma}{1-\gamma} F^n. \tag{13}$$

The parameter γ is meant to be a discount factor representing the rate at which future rewards diminish in value. There is no natural choice of discount factor in our problem, and so we can view γ as a tunable parameter. Higher values of γ place more emphasis on exploration and information collection, whereas lower values will make the policy behave similarly to DrawChance.

4 SIMULATION STUDY

Our objective function in (7) uses the draw probability to measure the quality of a game. However, draws never actually occur in any game, making it somewhat difficult to compare policies. In our simulation study, we used several intuitive performance measures to compare the point estimate policy of (8), the DrawChance policy of (9), and the heuristic KG policy of (13) with γ tuned to 0.99. Chief among these was the *true* draw probability of each match-up, defined as

$$q^{true}_{0,X^{\pi}(k^n)} = \frac{1}{\sqrt{4\pi\sigma^2_{\varepsilon}}} e^{-\frac{\left(s_0 - s_{X^{\pi}(k^n)}\right)^2}{4\sigma^2_{\varepsilon}}}. \tag{14}$$

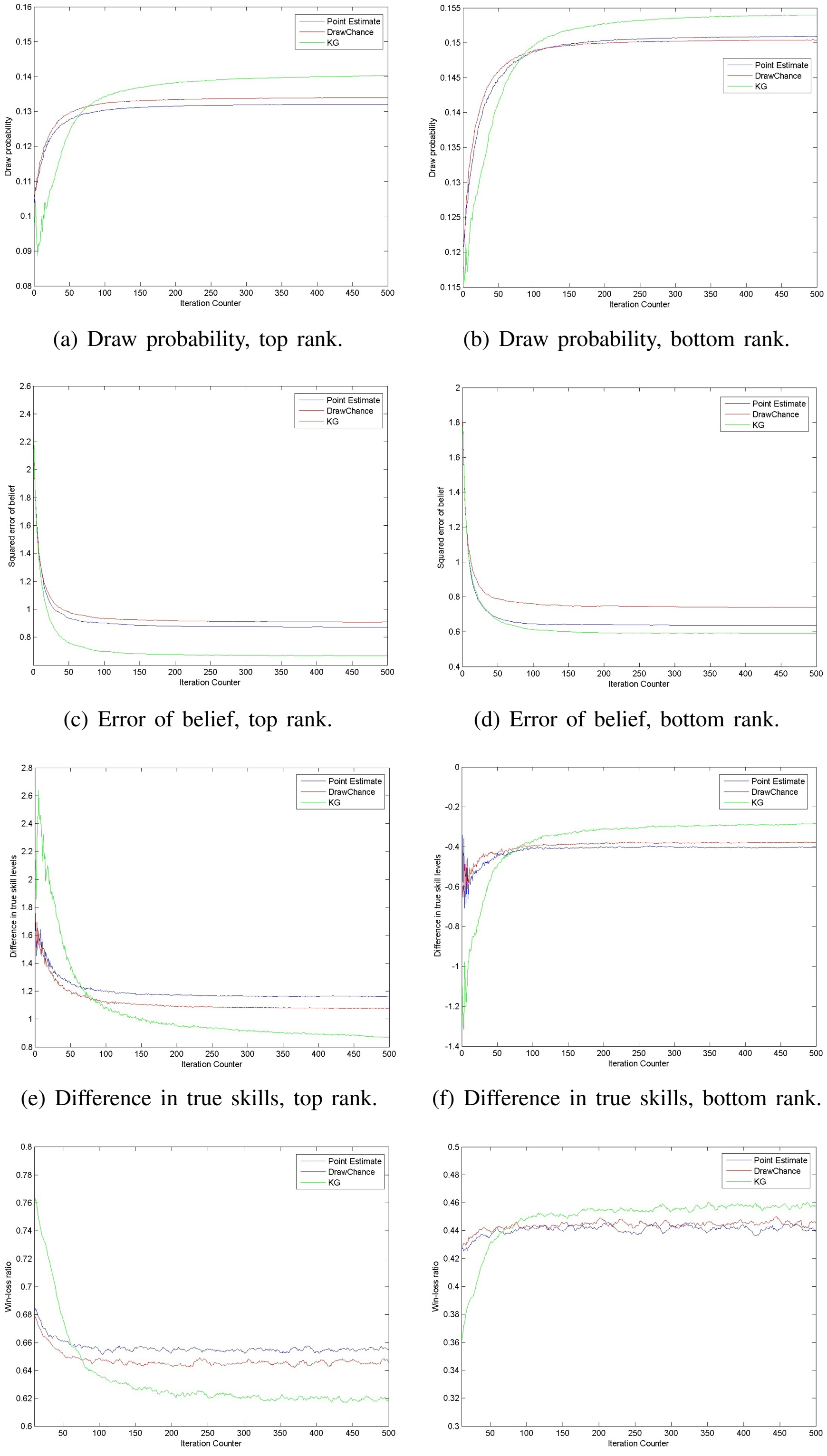
In (14), $X^{\pi}(k^n)$ is the opponent chosen by policy π based on the knowledge state at time n. Note that (14) requires us to know the exact values of the true skills s_i for $i = 0,1,...,M$. We set these values at the start of each set of games and used them to generate performance values p^n_i for $n = 1,...,N$. The policies were not allowed to see the true values when making decisions. However, we used the true values to compare the policies after the decisions had been made.

The simulations were set up as follows. First, we created the prior means μ^0_i for $i = 0,1,...,49$ by generating samples from a normal distribution with mean 0 and variance 4. The prior variances $\left(\sigma^0_i\right)^2$ were generated from a uniform distribution on the interval $[2,3]$. We then generated 10^4 sets of true skills $s_i \sim \mathcal{N}\left(\mu^0_i, \left(\sigma^0_i\right)^2\right)$, allowing for a fair amount of variation in the true skills, while keeping the priors accurate on average. For each set of true skills, we simulated a time horizon of $N = 500$. The noise in the observations was chosen to be $\sigma^2_{\varepsilon} = 2.5$.

In addition to the true draw probability from (14), we also compared each match-up with respect to the squared error $\left(\mu^n_0 - s_0\right)^2$ of the beliefs, the difference $s_0 - s_{X^{\pi}(k^n)}$ in the true skills of the players in the match-up, and the win/loss ratio of player 0 for that game. We report the difference $s_0 - s_{X^{\pi}(k^n)}$ as a signed quantity, not as an absolute value. Hence, positive values of the difference mean that player 0 was matched up against a less skillful player, whereas negative values mean that the opponent was favoured.

We discovered that KG was most effective when player 0 was ranked either at the top (highest μ^0_0), or at the bottom (lowest μ^0_0). As we see in Figure 1, KG explores more than other policies early on, choosing opponents that are farther away from player 0 in terms of skill. For the first hundred games, match-ups have lower draw probability under KG, and a greater difference in true skills. We learn the true skill level of player 0 faster, with the error $\left(\mu^n_0 - s_0\right)^2$ decreasing faster for KG in Figures 1(c) and 1(d). However, after the first hundred games, KG closes the gap and visibly pulls out ahead of the competition. In Figures 1(g) and 1(h), win/loss ratios are closer to 0.5 under KG. We also see the differences in true skills in Figures 1(e) and 1(f) come closer to zero under KG. Draw probabilities become higher, as well. KG consistently maintains its lead for the rest of the time horizon.

In the middle cases, when player 0 starts with a rank closer to the middle (e.g., rank 12, 25 or 37 out of 50), all three policies perform similarly, with no clear winner. Figure 2 provides an illustration for two performance measures. We see that the win/loss ratios, as well as the differences in true skills, are quite noisy and the results overlap for all the policies and all the starting ranks under consideration. We can see that KG still does more exploration in the early stages, but this does not result in an appreciable improvement over the other policies.



(a) Draw probability, top rank.

(b) Draw probability, bottom rank.

(c) Error of belief, top rank.

(d) Error of belief, bottom rank.

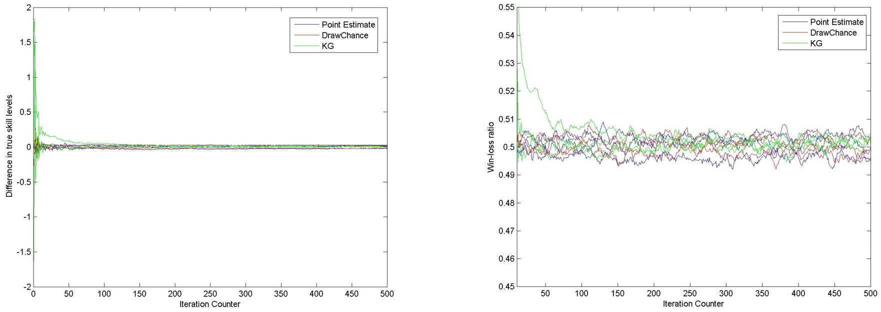
(e) Difference in true skills, top rank.

(f) Difference in true skills, bottom rank.

(g) Win/loss ratio, top rank.

(h) Win/loss ratio, bottom rank.

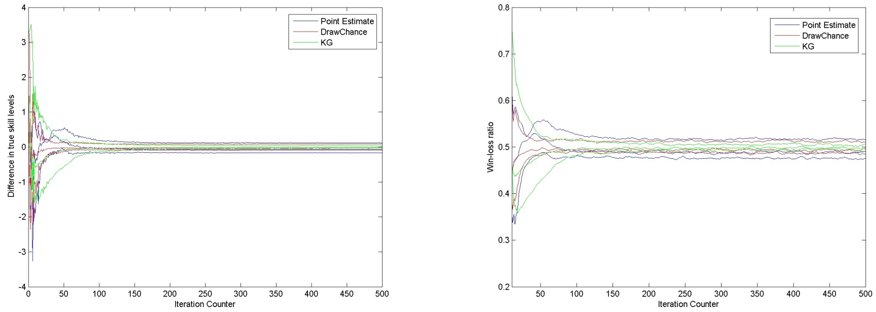
Figure 1: Performance measures, averaged over 10^4 sample paths, with player 0 starting in the top and bottom ranks (largest/smallest μ_0^0).



(a) Difference in true skills, middle ranks.

(b) Win/loss ratio, middle ranks.

Figure 2: Performance measures, averaged over 10^4 sample paths, with player 0 starting closer to the middle (25th, 50th and 75th percentile).



(a) Difference in true skills, middle skill levels.

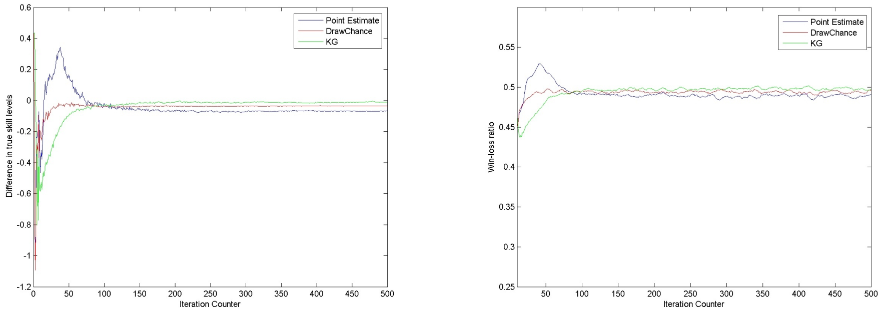
(b) Win/loss ratio, middle skill levels.

Figure 3: Performance measures, averaged over 10^4 sample paths, with equal priors.

We ran a second set of experiments where the starting prior was set to $\mu_i^0 = 0$ for all i, while the truths were generated from the same distributions as before, meaning that the prior provides little useful information about the players, and no way to distinguish them from one another. Figure 3 repeats the analysis of Figure 2 in this setting. The results are now less noisy, and KG now has a more pronounced tendency to come closer to the desired win/loss ratio of 0.5 and true skill difference of 0. This tendency is clearer in Figure 4, which presents a portion of the same results for the case where the true skill of player 0 is exactly in the middle. KG also maintained its good performance in the extreme cases where player 0 had the highest or lowest skill level of any player.

5 CONCLUSION

We have proposed a knowledge gradient policy for improving match-making decisions in competitive online gaming. We use the draw probability of Herbrich et al. (2006) as a criterion for the quality of a game, and demonstrate through simulations that KG can outperform other policy with respect to this and other performance measures. KG does particularly well when the player under consideration is ranked close to the top or bottom of the player pool, and performs competitively in other cases. In the e-sports setting, it is particularly important to do well in the extreme cases. The game designers would like to minimize frustration among new players, while also providing a greater challenge to professional players.



(a) Difference in true skills, middle skill level.

(b) Win/loss ratio, middle skill level.

Figure 4: Performance measures, averaged over 10^4 sample paths, with equal priors.

More importantly, however, we believe that the issues discussed in this paper have impact far outside the e-sports setting. The match-making problem is an example of what we might call "targeting and selection," a ranking and selection problem where the goal is not to find the alternative with the largest value, but rather to find an alternative whose value matches a target (e.g., a player's skill level). This problem arises in settings such as simulation calibration and design (see e.g., Frazier et al. 2009), where we are designing a simulator of a real-world system, and the goal is to make the simulation as true to life as possible, before we can begin to optimize it. We believe that the methodology examined in this paper can be highly useful in this setting. Future work in this direction is underway.

A DERIVATION OF DRAW PROBABILITY

The time-n distribution of $s_i - s_j$ is $\mathcal{N}\left(\mu_i^n - \mu_j^n, (\sigma_i^n)^2 + \left(\sigma_j^n\right)^2\right)$. We write

$$\mathbb{E}^n\left(\frac{1}{\sqrt{4\pi\sigma_\varepsilon^2}}e^{-\frac{\left(s_i-s_j\right)^2}{4\sigma_\varepsilon^2}}\right) = \int_{-\infty}^{\infty}\frac{1}{\sqrt{4\pi\sigma_\varepsilon^2}}e^{-\frac{x^2}{4\sigma_\varepsilon^2}}\frac{1}{\sqrt{2\pi\left((\sigma_i^n)^2+\left(\sigma_j^n\right)^2\right)}}e^{-\frac{\left(x-\left(\mu_i^n-\mu_j^n\right)\right)^2}{2\left((\sigma_i^n)^2+\left(\sigma_j^n\right)^2\right)}}. \tag{15}$$

Completing the square in the exponential terms in (15), the numerator inside the exponent becomes

$$\begin{aligned}
& x^2 - 2x\left(\mu_i^n - \mu_j^n\right) + \left(\mu_i^n - \mu_j^n\right)^2 + x^2\frac{(\sigma_i^n)^2 + \left(\sigma_j^n\right)^2}{2\sigma_\varepsilon^2} \\
= & \left[x\sqrt{1+\frac{(\sigma_i^n)^2+\left(\sigma_j^n\right)^2}{2\sigma_\varepsilon^2}} - \frac{\mu_i^n-\mu_j^n}{\sqrt{1+\frac{(\sigma_i^n)^2+\left(\sigma_j^n\right)^2}{2\sigma_\varepsilon^2}}}\right]^2 + \left(\mu_i^n-\mu_j^n\right)^2\left[1-\frac{1}{1+\frac{(\sigma_i^n)^2+\left(\sigma_j^n\right)^2}{2\sigma_\varepsilon^2}}\right] \\
= & \left(1+\frac{(\sigma_i^n)^2+\left(\sigma_j^n\right)^2}{2\sigma_\varepsilon^2}\right)\left[x-\frac{\mu_i^n-\mu_j^n}{1+\frac{(\sigma_i^n)^2+\left(\sigma_j^n\right)^2}{2\sigma_\varepsilon^2}}\right]^2 + \left(\mu_i^n-\mu_j^n\right)^2\left[1-\frac{1}{1+\frac{(\sigma_i^n)^2+\left(\sigma_j^n\right)^2}{2\sigma_\varepsilon^2}}\right] \qquad (16)
\end{aligned}$$

Observe that

$$1-\frac{1}{1+\frac{\left(\sigma_i^n\right)^2+\left(\sigma_j^n\right)^2}{2\sigma_\varepsilon^2}}=\frac{\left(\sigma_i^n\right)^2+\left(\sigma_j^n\right)^2}{\left(\sigma_i^n\right)^2+\left(\sigma_j^n\right)^2+2\sigma_\varepsilon^2}.$$

Simplifying and combining (15) with (16), then canceling out a normal integral over the real line, yields (5), as required.

REFERENCES

Chick, S. 2006. “Subjective Probability and Bayesian Methodology”. In *Handbooks of Operations Research and Management Science, vol. 13: Simulation*, edited by S. Henderson and B. Nelson, 225–258. North-Holland Publishing, Amsterdam.

Chick, S., and K. Inoue. 2001a. “New Procedures to Select the Best Simulated System Using Common Random Numbers”. *Management Science* 47 (8): 1133–1149.

Chick, S., and K. Inoue. 2001b. “New Two-Stage and Sequential Procedures for Selecting the Best Simulated System”. *Operations Research* 49 (5): 732–743.

Dangauthier, P., R. Herbrich, T. Minka, and T. Graepel. 2007. “TrueSkill Through Time: Revisiting the History of Chess”. In *Advances in Neural Information Processing Systems*, edited by J. C. Platt, D. Koller, Y. Singer, and S. Roweis, Volume 20, 337–344.

DeGroot, M. H. 1970. *Optimal Statistical Decisions*. John Wiley and Sons.

Elo, A. 1978. *The rating of chessplayers, past and present*. Arco Pub., NY.

Frazier, P. I., W. B. Powell, and S. Dayanik. 2008. “A knowledge gradient policy for sequential information collection”. *SIAM Journal on Control and Optimization* 47 (5): 2410–2439.

Frazier, P. I., W. B. Powell, and H. P. Simão. 2009, December. “Simulation Model Calibration With Correlated Knowledge-Gradients”. In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 339–351. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Gupta, S., and K. Miescke. 1996. “Bayesian look ahead one-stage sampling allocations for selection of the best population”. *Journal of Statistical Planning and Inference* 54 (2): 229–244.

Herbrich, R., T. Minka, and T. Graepel. 2006. “TrueSkill$^{\text{TM}}$: A Bayesian Skill Rating System”. In *Advances in Neural Information Processing Systems*, edited by B. Schölkopf, J. C. Platt, and T. Hoffman, Volume 19, 569–576.

Hong, L. J., and B. L. Nelson. 2009, December. “A Brief Introduction To Optimization Via Simulation”. In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 75–85. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Huhh, J. 2008. “Culture and business of PC bangs in Korea”. *Games and Culture* 3 (1): 26–37.

KeSPA 2011. “Korea e-Sports Association: What is e-Sports?”. Accessed Mar. 31, 2011. http://www.e-sports.or.kr/esports/Eng/esports_intro_10.kea?m_code=espor_10.

Kim, S., and B. Nelson. 2007, December. “Recent advances in ranking and selection”. In *Proceedings of the 2007 Winter Simulation Conference*, edited by S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, 162–172. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ryzhov, I. O., and W. B. Powell. 2009a, December. “A Monte Carlo Knowledge Gradient Method For Learning Abatement Potential Of Emissions Reduction Technologies”. In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 1492–1502. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ryzhov, I. O., and W. B. Powell. 2009b. "The knowledge gradient algorithm for online subset selection". In *Proceedings of the 2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning*, 137–144.

Ryzhov, I. O., W. B. Powell, and P. I. Frazier. 2011. "The knowledge gradient algorithm for a general class of online learning problems". *Operations Research (to appear)*.

Ryzhov, I. O., M. R. Valdez-Vivas, and W. B. Powell. 2010, December. "Optimal Learning of Transition Probabilities in the Two-Agent Newsvendor Problem". In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 1088–1098. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

SC2 Rankings 2011. "Starcraft II Rankings". Accessed Mar. 31, 2011. http://sc2ranks.com.

Scott, W. R., W. B. Powell, and H. P. Simão. 2010, December. "Calibrating simulation models using the knowledge gradient with continuous parameters". In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 1099–1109. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Swisher, J. R., P. D. Hyden, S. H. Jacobson, and L. W. Schruben. 2000, December. "A survey of simulation optimization techniques and procedures". In *Proceedings of the 2000 Winter Simulation Conference*, edited by J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, Volume 1, 119–128. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

ILYA O. RYZHOV is an Assistant Professor in the Robert H. Smith School of Business at the University of Maryland. He received a Ph.D. in Operations Research and Financial Engineering from Princeton University in 2011. His research seeks to bridge the gap between optimal learning and stochastic optimization by developing efficient decision-making strategies for many broad classes of optimization problems, and incorporating optimal learning concepts into fundamental operations research models such as network problems, linear programs, and Markov decision processes. His email address for these proceedings is iryzhov@rhsmith.umd.edu.

AWAIS TARIQ received a B.Sc. from the Department of Electrical Engineering at Princeton University. He also obtained certificates in Applications of Computing, Robotics and Engineering and Management Systems. He plans to continue his higher studies in the field of Operations and Industrial Engineering. His email address for these proceedings is atariq@princeton.edu.

WARREN B. POWELL is a Professor in the Department of Operations Research and Financial Engineering at Princeton University, and director of CASTLE Laboratory (http://www.castlelab.princeton.edu). He has coauthored over 150 refereed publications in stochastic optimization, stochastic resource allocation and related applications. He is the author of the book *Approximate Dynamic Programming: Solving the curses of dimensionality*, published by John Wiley & Sons. Currently, he is involved in applications in energy, transportation, finance and homeland security. His email address for these proceedings is powell@princeton.edu.