Beyond the Scoreboard: Advancing Fairness in Athlete Selection with Simulation-Based Tournament Strategies

E. Zhixuan Zeng[1] **Yuhong Zeng** [2]
[1]Vision and Image Processing Group, Systems Design Engineering, University of Waterloo
[2]Combinatorics and Optimization, University of Waterloo
{ezzeng,ivan.zeng1}@uwaterloo.ca

Abstract

The process of selecting athletes for competitive sports teams is often undermined by the limitations of traditional tournament formats, which can misrepresent the true skill levels of participants. This issue is exemplified by a scenario observed in a table tennis team tryout, where a moderately skilled player advanced to the final round due to consistently facing weaker opponents, while more adept players were eliminated early against stronger competitors. Such occurrences cast doubt on the fairness and effectiveness of single elimination tournaments for player assessment.

Addressing these concerns, our study conducts a thorough analysis of various tournament selection strategies, including single elimination, Swiss tournaments, and novel graph and sorting-based methods. By modeling players as Gaussian distributions with established mean skill levels, we simulate match outcomes to quantitatively evaluate the efficiency and accuracy of each strategy. Our evaluation employs two loss functions: Strict Loss, to gauge ranking precision, and Binary Loss, to assess the accuracy in identifying top performers.

The experimental results reveal significant insights. Strategies integrating Elo ratings with circular graph approaches show enhanced performance, particularly in larger player groups, while TrueSkill and single elimination exhibit limitations in scalability and nuanced player ranking. The Swiss tournament, although consistent, experiences fluctuations in loss, suggesting areas for refinement. Notably, a novel graph-based strategy emerges as a stable and efficient alternative, underscoring its potential for future research. These findings aim to guide the development of more equitable and precise selection processes in sports team composition.

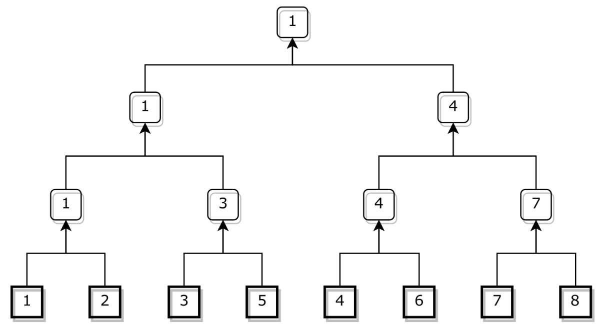

Fig. 1: An example of a single elimination tournament result, where each number represents each player's true skill level ranking (1 is most skilled). While the most skilled player is able to win the tournament, the actual second most skilled player was eliminated in the first round, and the 4th skilled player was able to move into the semifinals.

1 Introduction

Competitive sports demand a rigorous and meticulous approach to selecting athletes who will represent a team or organization. Unlike other tournament types where individuals compete solely for personal glory, team selection carries the unique challenge of identifying the top "x" individuals in terms of skill. This distinction underscores the gravity of the team selection process, making it markedly distinct from traditional tournaments.

In the context of player selection, many sports teams veer away from traditional tournament formats due to the inherent randomness they introduce. Instead, they rely on coaches and evaluators to subjectively assess player skill levels and make selections based on their judgment [1]. However, this approach comes with its own set of challenges. The subjectivity in evaluations can be influenced by various biases, including personal preferences, and may not always yield the most equitable outcomes. Moreover, the burden of selection lies squarely on the shoulders of the evaluators, who may themselves vary in expertise and impartiality.

The dynamics of team selection are further complicated by the need to select not just a single champion but a cadre of top performers. However, a granular ranking of individual skills is less important compared to a more binary outcome of accept or reject. This dichotomy adds another layer of complexity to the selection process.

Moreover, the logistical constraints of time and resources impose a practical ceiling on the number of matches or evaluations that can be feasibly conducted, especially when dealing with a large pool of candidates. It is imperative, therefore, to strike a delicate balance between the thoroughness of the selection process and the efficient allocation of available resources.

In response to these challenges, a body of literature has emerged, exploring innovative strategies that aim to refine the design of tournaments and the estimation of player rankings [2–6], as well as providing in-depth analyses of existing methods [7–11]. This study will aim to do both, scrutinizing established methods, such as single-elimination, Swiss tournaments, as well introducing and testing innovative graph and sorting-based selection strategies. By modeling players as Gaussian distributions with known mean "skill levels," we not only quantify the efficiency and accuracy of each strategy but also illuminate their strengths and weaknesses through a comprehensive simulation-based approach.

2 Methodology

2.1 Simulation

Our methodology revolves around a simulation-based approach to evaluate different tournament selection strategies. To accurately assess player skill levels and the fairness of each strategy, we model participants as Gaussian distributions $N(\mu, \sigma)$ with mean μ and standard deviation σ. These values are randomly assigned through sampling a distribution of $N_{\mu}(\mu_{\mu}, \sigma_{\mu})$ for the means and $N_{\sigma}(\mu_{\sigma}, \sigma_{\sigma})$ for the standard deviations.

For each match between two players, values are sampled from their skills distribution repeatedly. A higher value between the two players would result a single point won by that player. Following table tennis rules, the player to first reach a total of 11 points while leading by at least 2 points win. If 11 points is reached when the opponent has 10 points, the game continues until a player leads by 2 points.

2.2 Evaluation

Our study aims to assess the effectiveness of each tournament selection strategy in two distinct dimensions: ranking accuracy and top performer identification. To evaluate these aspects, we employ two distinct loss functions:

1. **Strict Loss:** This loss quantifies the squared difference between the predicted and actual player rankings. It assesses the accuracy of the strategies in correctly ordering the players.

2. **Binary Loss:** This loss measures the number of false positives and false negatives in identifying top performers, treating it as a binary classification problem ("top performer" vs. "bottom performer").

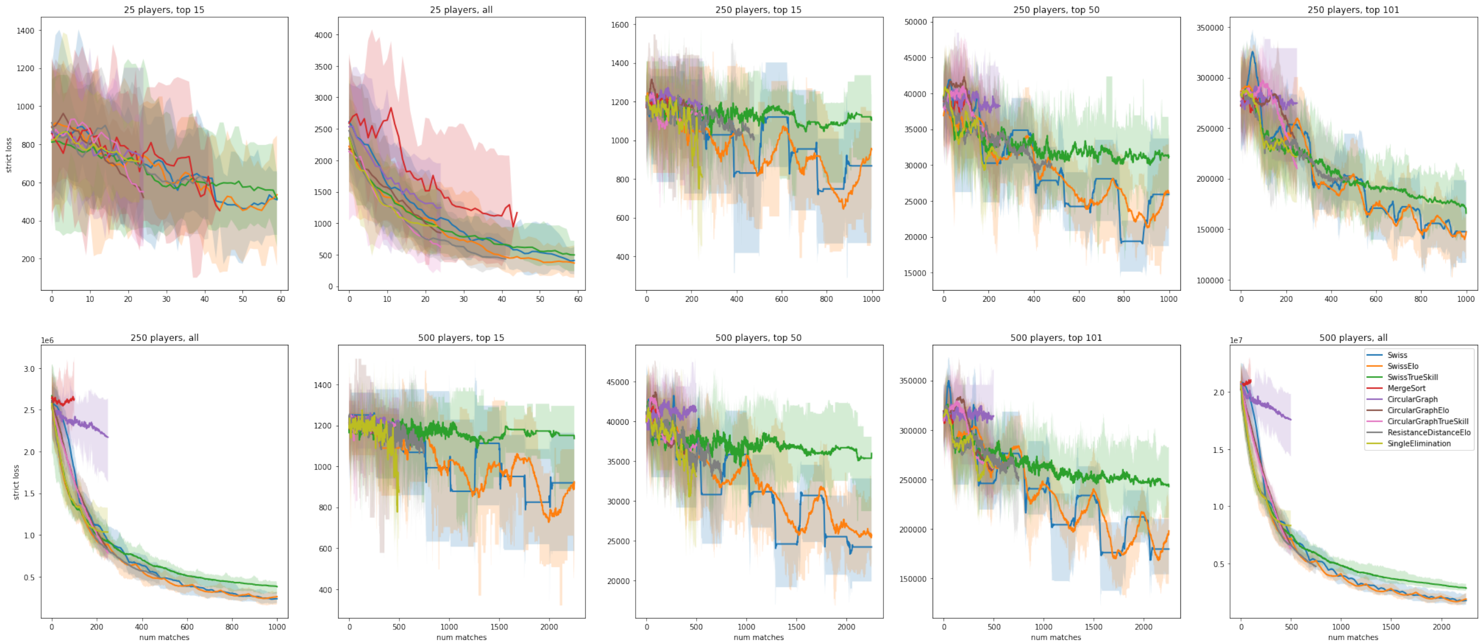

Fig. 2: Tournament selection algorithms evaluated on strict ranking loss

2.3 Selection Strategies

2.3.1 Swiss-system Tournament

The Swiss-system tournament is a traditional format designed to accommodate a large number of players through a series of rounds, pairing players with similar records. It is structured as follows:

- Players are initially paired randomly.
- In each round, players with similar records face each other.
- Players accumulate points based on wins and draws.

In the case of this simulation, the player's ranks are based on their accumulated scores.

2.3.2 Single Elimination

Single elimination is a classic format where players are eliminated after a single loss, leading to a final showdown between the last two remaining competitors. The player's ranks are based on how far they manage to proceed in the tournament.

2.3.3 Merge Sort

our modified merge sort ranking method optimizes comparisons among players by assuming transitivity, meaning that if $A > B > C$, it follows that $A > C$ as well. Players are assigned to groups, which are then continuously merged with adjacent branches by comparing the members of each group in order. The time complexity for a full ranking of n players is approximately $n\lceil\log(n)\rceil$.

2.3.4 Circular Graph

The Circular Graph strategy is designed for situations with very limited time. Each player is only compared to their 2 adjacent players. After each player plays exactly 2 games, we are able to construct a circular graph. Through propagating the performance of each player through their connected edges, we estimate their performance relative to the entire group.

2.3.5 Graph Resistance Distance

The Graph Resistance Distance strategy employs resistance distance calculations within a graph to identify opponents which are further away [12]. For example, if A has played B, and B played C, and C played D, then D would have a higher resistance distance calculation, assuming no other edges. Unlike a Swiss tournament, this selection algorithm does not consider potential relative skill or performance.

2.4 Ranking Strategies

Some of the above mentioned selection strategies can be further complemented by ranking or skill estimation algorithms. The Circular graph strategy, for example, simply uses the win/loss rate between two players to estimate their relative skill, but this calculation can be replaced by other popular rating systems like the ones mentioned below. Similarly, they can also be used in Swiss tournaments to estimate good pairings, or for the Graph Resistance strategy to estimate the final ranking.

2.4.1 ELO

The ELO rating system [13] is a widely used method for ranking players. It calculates the expected outcome of a match based on the difference in players' ratings and updates the ratings after each match. The formula for calculating expected outcome is:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{1}$$

where E_A is the expected outcome for player A, R_A and R_B are the ratings of players A and B, respectively. The actual outcome S_A (1 for a win, 0 for a loss, and 0.5 for a draw) is used to update the ratings:

$$R'_A = R_A + K(S_A - E_A) \tag{2}$$

where R'_A is the new rating for player A, K is a constant, and S_A is the actual outcome.

2.4.2 TrueSkill

In the TrueSkill algorithm [14, 15], player ratings are represented as Gaussian distributions rather than a scalar, with each player having both a mean skill level (μ) and an uncertainty (σ) associated with their rating. Draw probability is then calculated as a "quality estimate". This rank estimate can again be used in both a Swiss style tournament as well as the Circular Graph strategy.

By combining these selection and ranking strategies, our study aims to provide comprehensive insights into the most effective methods for accurately identifying top-performing athletes in the context of team selection, while considering the constraints of time and resources.

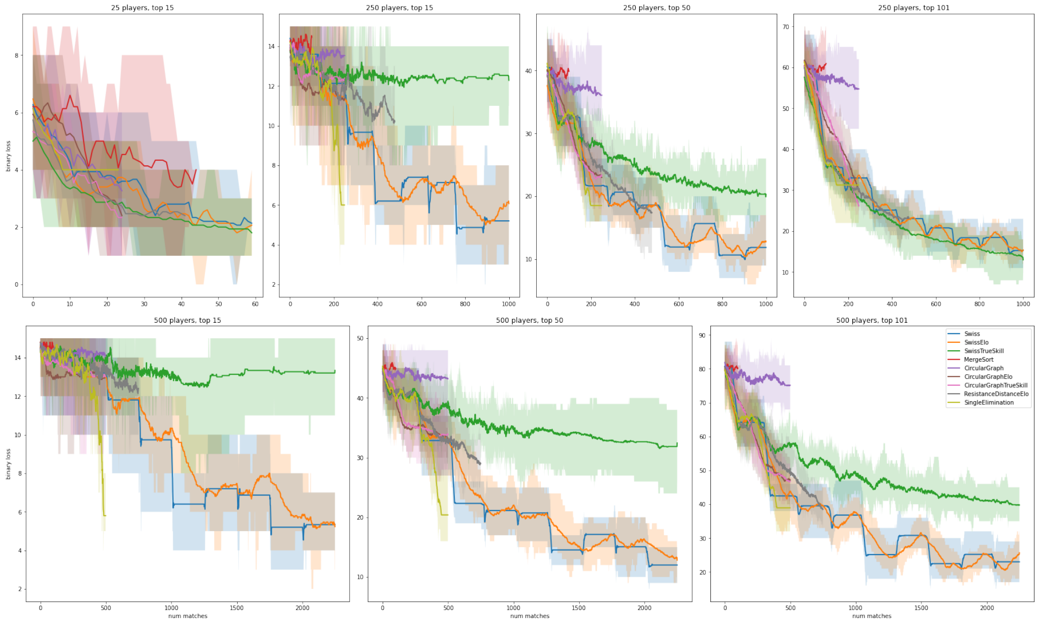

Fig. 3: Tournament selection algorithms evaluated on binary selection loss

3 Experimental Results

In this section, we present the outcomes of our extensive simulations, where we assessed the performance of various tournament selection strategies in terms of strict and binary losses, aiming to evaluate their efficiency and accuracy in selecting top-performing athletes.

3.1 Strict and Binary Loss Convergence Analysis:

To monitor the efficiency of each tournament selection strategy, we tracked the evolution of both strict and binary losses after each simulated match. These plots, which illustrate the strategies' convergence towards accurate player rankings and the identification of top performers, provide valuable insights into their respective strengths and weaknesses.

3.2 Performance Observations:

Our experiments yielded several notable observations:

1. **TrueSkill's limited Performance:** TrueSkill shows decent performance for evaluating the true rank of the entire player population. However, it exhibited suboptimal performance when specifically identifying the top performers, suggesting potential limitations in this regard.
2. **Single Elimination's Efficiency and Limitations:** Single elimination demonstrated exceptional efficiency when quickly identifying top performers in a large pool of players. However, its early stopping condition curtails its potential for further convergence to establish more refined rankings. Additionally, its efficiency is less notable in smaller group settings.
3. **Merge Sort and Circular Graph Strategy Dynamics:** In contrast to single elimination, the merge sort and the base circular graph strategy displayed decent competency in smaller player groups but faced significant challenges when scaling to larger ones, ultimately faltering in this context.
4. **Circular Graph Strategies with Elo or TrueSkill Integration:** Substituting the base win rate estimation calculation in the circular graph strategies with Elo or TrueSkill proved to be a transformative enhancement. This modification enabled these strategies to sustain competitive performance even in larger player groups. This underscores the potential of integrating well-established rating systems to alleviate scalability concerns encountered by other strategies.
5. **Swiss Tournament's Consistency:** The Swiss tournament, when employing its base ranking strategy as well as Elo for skill estimation, consistently demonstrated top-tier performance in both strict and binary losses. However, it exhibited cyclic fluctuations in loss over time, which may require further investigation.
6. **Resistance Distance Strategy's Efficiency and Stability:** While not eclipsing Swiss tournaments in overall performance, the resistance distance strategy proved dependable in terms of loss reduction rate. Importantly, it sidestepped the cyclic loss fluctuations observed in Swiss and SwissElo. This suggests that such graph-based selection approaches may be an interesting point for further research.

4 Conclusion

The quest for a fair and objective method of selecting top athletes for team sports is fraught with challenges, from the unpredictability of tournament outcomes to the subjective biases of human judgment. Our simulation-based study has critically evaluated various tournament selection strategies, revealing a complex landscape where no single method is without its drawbacks. TrueSkill and single elimination, while efficient in certain contexts, fall short in consistently identifying the true top performers, particularly in larger pools of athletes. On the other hand, strategies that incorporate Elo ratings into circular graph approaches show remarkable adaptability and robustness, suggesting that the integration of established rating systems can significantly enhance selection processes. The Swiss tournament, with its consistent performance, remains a reliable method, though its periodic fluctuations in ranking accuracy point to areas for potential refinement. Notably, the resistance distance strategy stands out for its stability and consistent loss reduction, marking it as a promising direction for future research. Ultimately, this study contributes to the ongoing dialogue on athlete selection, advocating for a more data-driven, analytical approach that can serve to improve the fairness and objectivity of team composition in sports.

References

[1] "College recruiting process: How do colleges recruit athletes?" Sep 2023. [Online]. Available: https://www.ncsasports.org/recruiting/how-to-get-recruited/college-recruiting-process

[2] X. Yan, Y. Du, B. Ru, J. Wang, H. Zhang, and X. Chen, "Learning to identify top elo ratings: A dueling bandits approach," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 8, 2022, pp. 8797–8805.

[3] S. Motegi and N. Masuda, "A network-based dynamical ranking system for competitive sports," *Scientific reports*, vol. 2, no. 1, p. 904, 2012.

[4] L. Csató, "On the ranking of a swiss system chess team tournament," *Annals of Operations Research*, vol. 254, no. 1-2, pp. 17–36, 2017.

[5] P. Führlich, Á. Cseh, and P. Lenzner, "Improving ranking quality and fairness in swiss-system chess tournaments," in *Proceedings of the 23rd ACM Conference on Economics and Computation*, 2022, pp. 1101–1102.

[6] Á. Cseh, P. Führlich, and P. Lenzner, "The swiss gambit," *arXiv preprint arXiv:2302.10595*, 2023.

[7] B. R. Sziklai, P. Biró, and L. Csató, "The efficacy of tournament designs," *Computers & Operations Research*, vol. 144, p. 105821, 2022.

[8] T. McGarry and R. Schutz, "Efficacy of traditional sport tournament structures," *Journal of the Operational Research Society*, vol. 48, no. 1, pp. 65–74, 1997.

[9] B. Vaziri, S. Dabadghao, Y. Yih, and T. L. Morin, "Properties of sports ranking methods," *Journal of the Operational Research Society*, vol. 69, no. 5, pp. 776–787, 2018.

[10] D. G. d. P. Zanco, L. Szczecinski, E. V. Kuhn, and R. Seara, "A comprehensive analysis of the elo rating algorithm: Stochastic model, convergence characteristics, design guidelines, and experimental results," *arXiv preprint arXiv:2212.12015*, 2022.

[11] A. Berg, "Statistical analysis of the elo rating system in chess," *Chance*, vol. 33, no. 3, pp. 31–38, 2020.

[12] A. K. Chandra, P. Raghavan, W. L. Ruzzo, and R. Smolensky, "The electrical resistance of a graph captures its commute and cover times," in *Proceedings of the twenty-first annual ACM symposium on Theory of computing*, 1989, pp. 574–586.

[13] A. E. Elo and S. Sloan, "The rating of chessplayers: Past and present," *(No Title)*, 1978.

[14] R. Herbrich, T. Minka, and T. Graepel, "Trueskill™: a bayesian skill rating system," *Advances in neural information processing systems*, vol. 19, 2006.

[15] T. Minka, R. Cleven, and Y. Zaykov, "Trueskill 2: An improved bayesian skill rating system," *Technical Report*, 2018.