
F3

Faculty of Electrical Engineering
Department of Cybernetics

Bachelor Thesis

# Problems of Maximum Likelihood Estimates

**Danil Alshaev**

**May 2024**
**Study program: Open Informatics**
**Specialization: Artificial Intelligence and Computer Science**
**Supervisor: prof. Ing. Mirko Navara, DrSc.**

# BACHELOR‘S THESIS ASSIGNMENT

## I. Personal and study details

Student's name: **Alshaev Danil** Personal ID number: **507272**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Cybernetics**

Study program: **Open Informatics**

Specialisation: **Artificial Intelligence and Computer Science**

## II. Bachelor’s thesis details

Bachelor’s thesis title in English:

**Problems of Maximum Likelihood Estimates**

Bachelor’s thesis title in Czech:

**Problémy maximálně věrohodných odhadů**

Guidelines:

1. Find practical examples where maximum likelihood estimates lead to undesirable results.
2. Investigate the errors occurring when groups of data are replaced by a single value to reduce the computational complexity. Demonstrate their bounds on artificially designed data.
3. In items 1, 2, you may add also positive examples of the use of MLE. A part of the report should be usable as an supplementary material for education.

Bibliography / sources:

[1] Wasserman, L.: All of Statistics. A Concise Course in Statistical Inference. Springer, 2004.
[2] Glickman, Mark E., and Jones, Albyn C., ”Rating the chess rating system” (1999), Chance, 12, 2, 21-28
[3] Glickman, Mark E. (1995), ”A Comprehensive Guide to Chess Ratings”. str. 59–102.

Name and workplace of bachelor’s thesis supervisor:

**prof. Ing. Mirko Navara, DrSc. Machine Learning FEE**

Name and workplace of second bachelor’s thesis supervisor or consultant:

Date of bachelor’s thesis assignment: **31.01.2024** Deadline for bachelor thesis submission: **24.05.2024**

Assignment valid until: **21.09.2025**

| prof. Ing. Mirko Navara, DrSc. | prof. Dr. Ing. Jan Kybic | prof. Mgr. Petr Páta, Ph.D. |
|---|---|---|
| Supervisor’s signature | Head of department’s signature | Dean’s signature |

## III. Assignment receipt

The student acknowledges that the bachelor’s thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the bachelor’s thesis, the author must state the names of consultants and include a list of references.

. _____
Date of assignment receipt

_____
Student’s signature

# Acknowledgement / Declaration

I want to thank professor Mirko Navara for the supervision of this bachelor thesis.

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university theses.

Prague, 24 May 2024

Danil Alshaev

# Abstrakt / Abstract

Odhad maximální věrohodnosti (Maximum Likelihood Estimation, MLE) se stal oblíbeným nástrojem, který nabízí přístup k odhadu parametrů statistického modelu. Oblast MLE však přináší problémy, kterým je třeba porozumět, aby bylo možné pokročit v teorii a aplikaci.

V této práci se zabýváme problémy MLE na souboru dat o šachu. Je použit a zdůvodněn nejvhodnější model pro použití MLE při odhadu parametrů hráčů v souboru šachových partií. Čtenáře seznamuje se všemi potřebnými pojmy pro pochopení šachového hodnotícího systému i modelu.

Práce ukázuje nejen potenciální problémy, které by mohly nastat při použití MLE, ale také některé zajímavé problémy, které mohou nastat v reálném scénáři při pokusu o vytvoření modelu porovnávání dvojic v partiích.

**Klíčová slova:** MLE, Elo, šachy, FIDE, rating, odhad ratingu

Maximum Likelihood Estimation (MLE) has become a popular tool offering an approach to estimating statistical model parameters. However, the realm of MLE presents challenges that must be understood to advance theory and application.

In this thesis, we examine problems of MLE using a chess dataset. The most appropriate model is employed and justified for the use of MLE in estimating the parameters of players in a set of chess games. The reader is provided with all the necessary terms for understanding of both a chess rating system and a model.

The thesis demonstrates to the reader not only the potential issues that could arise with MLE, but also some interesting problems that may occur in a real-world scenario when attempting to create a pair comparison model in games.

**Keywords:** MLE, Elo, chess, FIDE, rating, rating estimation

# Contents /

# Tables / Figures

# Chapter 1
# Introduction

The objective of this thesis is to examine the issues associated with Maximum Likelihood Estimation (MLE). The data used are drawn from chess games. The reader will be provided with an overview of the chess rating system and the model itself, as well as the computational techniques employed for the estimation of players' ratings and the subsequent impact on the estimation process.

The objective of this investigation is to examine various issues and identify potential causes of discrepancies in estimation. Additionally, we will explore potential solutions to enhance the accuracy and precision of these estimates. Finally, we will undertake an experiment to examine further difficulties encountered in rating estimation.

One of the principal difficulties encountered in the estimation of parameters in our model is the replacement of groups of data by a single value in order to reduce the computational complexity. The thesis will concentrate on this specific problem. We will demonstrate the impact of this phenomenon on the final estimation. The datasets relevant to this specific issue will be presented to the reader in Chapter 3. Once the player's rating has been estimated, the difference between the estimated value and the actual player's rating will be examined.

The objective of this examination is to provide insights for researchers, practitioners and statisticians into the nuances of MLE in order to enhance their comprehension of its strengths and limitations. The intention is to address the challenges associated with MLE and to improve approaches that facilitate inference and informed decision-making from data.

**The paper is structured as follows:**

- **Chapter 1** - introduction.
- **Chapter 2** - provides definitions for the terms `MLE` and the chess rating system, as well as the model that will be employed and the method of calculating players' ratings.
- **Chapter 3** - presents a series of problems that can arise based on different chess datasets, and it examines the magnitude of the error based on the confidence interval. It also discusses the underlying causes of these issues and suggests potential solutions for enhancing estimation.
- **Chapter 4** - potential enhancements to the model may be employed to address the nuances that influence the estimation of a player's rating.
- **Chapter 5** - discusses what was not covered in the thesis and also considers other factors that may influence a player's rating.
- **Chapter 6** - conclusion.

# Chapter 2
# Theory

This chapter presents the fundamental concepts related to the subject of this thesis. We introduce the MLE and the associated mathematical formulas, as well as the chess terminology that will be employed in the subsequent model and problem definition.

## 2.1 Maximum Likelihood Estimation

MLE is a method used to estimate the parameters of a statistical model by maximizing the likelihood function, which measures the probability of observing the given data under the assumed model. The following is a definition of MLE:

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with probability density function $f(x;\theta)$, *continuous in the first argument*, where $\theta$ is a parameter vector to be estimated.

The likelihood function $L(\theta)$ is defined as the joint probability density function of the observed data, given the parameter $\theta$:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

The Maximum Likelihood Estimator $\hat{\theta}$ is the value of $\theta$ that maximizes the likelihood function:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ L(\theta)$$

In cases where it is more convenient to work with the logarithm of the likelihood function, the log-likelihood function $l(\theta)$ is defined as:

$$l(\theta) = \log L(\theta)$$

MLE is a widely used method for estimating parameters in statistical models, providing estimates that are asymptotically unbiased and efficient under certain regularity conditions.

## 2.2 Chess rating system

Chess rating systems serve a variety of practical purposes. In tournaments, the use of rating systems allows for the effective pairing of players, thus preventing the most formidable contenders from being matched against each other in the initial rounds.

Additionally, ratings play a crucial role in tournament segmentation and in determining prize eligibility. They are also used as a criterion of quality and importance of tournaments. Furthermore, they serve as a qualifying metric for prestigious events, such

as elite tournaments like the Candidates Tournament, where invitations are extended to the highest-rated players who have not qualified through alternative pathways. The Candidates Tournament is the event that determines the challenger for the reigning World Chess Champion.

This thesis will utilise the Elo Rating System [4], which is currently the most widely used rating system, although it has many variations and improvements. Elo-like rating systems have also been adopted in other contexts, such as other games like Go, online competitive gaming, and dating apps.

### 2.2.1 Elo rating system

The Elo rating system is a method for calculating the relative skill levels of players in two-player games, such as chess. It was developed by Arpad Elo, a Hungarian–American physics professor and chess master, in the late 20th century. The system assumes that the outcome of a game between two players depends solely on their skills and not on other factors.

In the Elo system, each player is assigned a numerical rating. The difference in ratings between two players predicts the outcome of a match. If Player A has a higher rating than Player B, Player A is expected to win more games against Player B. The magnitude of the difference in ratings estimates the probability that each player wins.

### 2.2.2 Mathematical details

Let:
$R_A$ be the current Elo rating of player A.
$R_B$ be the current Elo rating of player B.
The expected score of player A against player B, denoted by $E_A$, is calculated using the logistic function:

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

The expected score of player B denoted $E_B$ is:

$$E_B = \frac{1}{1 + 10^{\frac{R_A - R_B}{400}}} . \tag{2}$$

The probability of drawing, as opposed to having a decisive result, is not specified in the Elo system. Instead, a draw is considered half a win and half a loss ($\frac{1}{2}$).

The difference in Elo points between two players can be used to predict the expected outcome of a game (or series of games) between those two players. The conversion of the difference to probability of winning is shown in Table 2.1.
As the rating difference increases, the probability of the higher-rated player (player A) winning also increases.

### 2.2.3 Updating ratings

The formula for updating ratings [2] involves adjusting a player's rating based on newly observed data. These adjustments are made gradually to avoid the need to recompute a player's estimated rating using his entire tournament history. Instead, a pre-tournament rating serves as a condensed summary of the player's history leading up to the current

| Rating difference | Expected score of player A |
|---|---|
| 0 | 0.50 |
| 100 | 0.64 |
| 200 | 0.76 |
| 300 | 0.85 |
| 400 | 0.91 |
| 500 | 0.95 |
| 600 | 0.97 |
| 700 | 0.99 |

**Table 2.1.** Rating difference and expected score.

tournament. To ensure computational efficiency, an approximation method is used to calculate performance ratings.

$$R_{post} = R_{pre} + K(S - S_{exp})\,, \tag{3}$$

where:
$R_{post}$ is a player's updated post-tournament estimated rating,
$R_{pre}$ is a player's estimated pre-tournament rating,
$S$ is the player's total score in the tournament,
$S_{exp}$ is the expected total score estimated from the player's pre-tournament rating and the player's opponents pre-tournament ratings,
$K$ is an attenuation factor that determines the weight that should be given to a player's performance relative to his or her pre-tournament rating (speed factor of learning).

The expected score can be estimated by calculating the difference between the player's Elo and the average Elo of all his opponents. This method was used by the International Chess Federation [2]. In this thesis, we compute the expected score $E_{exp}$ using the provided formula for computational efficiency:

$$E_{exp} = \frac{1}{1 + 10^{\frac{R_O - R_A}{400}}}\,, \tag{4}$$

$$S_{exp} = n \cdot E_{exp}\,, \tag{5}$$

where
$E_{exp}$ is the expected score,
$n$ is the number of games played in a tournament,
$R_A$ is the current Elo rating of player A,
$R_O$ is the average Elo of all opponents, calculated the following way:

$$R_O = \frac{1}{n}\sum_{i=1}^{n} R_i\,, \tag{6}$$

$R_i$, $i = 1, \ldots, n$, are the individual pre-tournament ratings of the opponents.

## 2.3 Model

In this section, we discuss the model used for parameter estimation and potential issues that may arise.

### 2.3.1 Bernouli distribution

The model will employ the Bernoulli distribution. It is a discrete probability distribution representing a random variable that can take on one of two possible outcomes, typically labeled as success (with probability $p$) or failure (with probability $1-p$). It is named after Swiss mathematician Jacob Bernoulli.

$$f(k;p) = \begin{cases} p & \text{if } k = 1, \\ 1-p & \text{if } k = 0, \end{cases}$$

$$f(k;p) = p^k(1-p)^{1-k}.$$

### 2.3.2 The outcome of a chess game utilized by the model

The game of chess has 3 outcomes – a win, draw, and loss: if player A wins, he gets 1 point, if the game ends in a draw, both players get $\frac{1}{2}$ point, and if player A loses, player A gets 0 points and player B gets 1 point. In the model that will be used, each game is treated as two half-point contests with the same Bernoulli distribution. This allows for the determination of all three probabilities based on a single number.

Denote: $p$ is the probability of winning in a one *half-point contest*. Therefore, $1-p$ is the probability of winning for the second player. The probability of a first player winning the *game* is $p^2$, while the probability of drawing is $2p(1-p)$ and the probability of losing is $(1-p)^2$. A comparable approach was adopted by Mark Glickman in his paper [3], whereby two draws were treated as one win and one loss. Let us examine real-world situations and players' games to determine if the probability we have obtained is reasonable and applicable in a real-world scenario.

The statistics to be examined are those of the top players with a rating of 2700 or higher. It is important to note that statistics can vary depending on the rating range and the player's individual style. In this case, it is important to analyse the statistics of top-rated players as upcoming chapters will focus on the performance of these players. The relevant statistics can be found in a chess database[1].

| Move | # | Most Recent | Av/Perf/Max Rating | White Win/Draw/Black Win | | |
|---|---|---|---|---|---|---|
| 1.e4 | 12907 | 20242757 | 2792/2882 | 29.8% | 49% | 21.3% |
| 1.d4 | 10355 | 20242756 | 2796/2882 | 29.5% | 51.2% | 19.3% |
| 1.Nf3 | 3493 | 20242755 | 2797/2876 | 31.7% | 47.4% | 21% |
| 1.c4 | 2210 | 20242762 | 2796/2882 | 31.3% | 47.6% | 21.1% |

**Figure 2.1.** Statistics of chess game outcomes.

In the first column there is the first move in the game. In the second column there is the number of games played with this move. Based on different opening there can be different statistics. Here are the most common first moves in the top players' games with sufficient amount of recordings for statistics. Now let's examine how it corresponds to our formulas based on different parameter $p$ and how it relates to the real-world scenario.

Tables 2.2, 2.3, 2.4 demonstrate that the probabilities closely resemble those found in real-world situations. Our assumption holds true and works well for both opening moves and individual players. The database of individual chess games was derived from the chess.com[2] database.

[1] https://old.chesstempo.com/game-database.html
[2] https://www.chess.com/games

| $p$ | Winning | Drawing | Losing |
|---|---|---|---|
| 0.55 | 0.30 | 0.50 | 0.20 |
| 0.56 | 0.31 | 0.49 | 0.20 |
| 0.57 | 0.32 | 0.49 | 0.19 |

**Table 2.2.** Parameter $p$ and probabilities of outcome of the game.

| Player Name | Rating | Total Games | Wins | Draws | Losses |
|---|---|---|---|---|---|
| Anatoly Karpov | 2617 | 4377 | 0.41 | 0.46 | 0.13 |
| Garry Kasparov | 2812 | 2419 | 0.49 | 0.42 | 0.09 |
| Viswanathan Anand | 2751 | 4253 | 0.36 | 0.50 | 0.14 |
| Vladimir Kramnik | 2753 | 4022 | 0.39 | 0.44 | 0.17 |

**Table 2.3.** Individual players statistics.

| Player Name | $p$ | Wins | Draws | Losses |
|---|---|---|---|---|
| Anatoly Karpov | 0.64 | 0.41 | 0.46 | 0.13 |
| Garry Kasparov | 0.7 | 0.49 | 0.42 | 0.09 |
| Viswanathan Anand | 0.6 | 0.36 | 0.48 | 0.16 |
| Vladimir Kramnik | 0.62 | 0.39 | 0.47 | 0.14 |

**Table 2.4.** Calculated $p$ for individual players and expected outcomes from it.

Based on statistics, it is reasonable to assume that two half-point contests are equivalent to a chess game. This assumption holds true in the long run and for a sufficient number of games. Disregarding ties simplifies the model by avoiding the need for additional parameters to account for a third outcome.

### 2.3.3 Estimating the player's expected score

The estimate of the expected score can be achieved through MLE.

Let:
$w_n$ be the number of wins,
$d_n$ be the number of draws,
$l_n$ be the number of losses,

The likelihood function for a Bernoulli distribution with parameter $p$ is given by:

$$L(p) = p^{2w_n}(2p - 2p^2)^{d_n}(1-p)^{2l_n}\,.$$

Taking the natural logarithm of the likelihood function (log-likelihood), we get:

$$\log L(p) = 2w_n \log p + d_n \log(2p - 2p^2) + 2l_n \log(1-p)\,.$$

Now, we will determine the MLE for $p$ by taking the derivative of the log-likelihood function with respect to $p$.

$$\frac{d}{dp}\log L(p) = \frac{2w_n}{p} + \frac{d_n(1-2p)}{p(1-p)} - \frac{2l_n}{1-p}\,.$$

Setting the derivative equal to zero:

$$2w_n(1-p) + d_n(1-2p) - 2l_n p = 0\,.$$

The MLE for the mean of the Bernoulli distribution can be obtained by solving for $p$:

$$p = \frac{w_n + \frac{d_n}{2}}{w_n + d_n + l_n}\,. \tag{7}$$

The estimate of a player's expected score:

$$\hat{E} = p\,.$$

This expression represents the sample mean, which is also the maximum likelihood estimate for the parameter $p$ in the Bernoulli distribution.

### 2.3.4 Equation

Given an estimate of a player's probability of winning which is $\hat{E}$, we can estimate a player's rating if we know the ratings of his opponents in a sequence of games by solving the following equation that we discussed above:

$$\hat{E}_A = \frac{1}{1 + 10^{\frac{R_A - R_O}{400}}}\,.$$

Using MLE estimation we can estimate not only the expected score but also a rating of a player.

The likelihood function for a Bernoulli distribution with parameter $R_A$ can be derived from our estimate:

$$\begin{aligned}
p = \frac{w_n + \frac{1}{2}d_n}{w_n + d_n + l_n} &= \frac{1}{1 + 10^{\frac{-R_A + R_O}{400}}}\,,\\
10^{\frac{-R_A + R_O}{400}} &= \frac{1-p}{p}\,,\\
\frac{-R_A + R_O}{400} &= \log_{10}\frac{1-p}{p}\,,\\
\hat{R}_A = R_O - 400\log_{10}\frac{1-p}{p} &= R_O - 400\log_{10}\left(\frac{l_n + \frac{1}{2}d_n}{w_n + \frac{1}{2}d_n}\right)\,.
\end{aligned} \tag{8}$$

In this formula, which we were able to estimate using the MLE, we can interpret $400\log_{10}\left(\frac{l_n+\frac{1}{2}d_n}{w_n+\frac{1}{2}d_n}\right)$ as the difference between the opponent's player rating and the actual player rating based on the results of the matches. The expression $\frac{l_n+\frac{1}{2}d_n}{w_n+\frac{1}{2}d_n}$ can be interpreted as the ratio of points scored by the opponent to points scored by the player. This formula is the solution to the equation above. Now we are ready for our model to estimate a player's actual rating, knowing his opponent's rating and the result in a sequence of games.

Note that this formula can only estimate a player's rating if their expected score is not 0 or 1. Otherwise, the player would be considered the weakest or strongest player, with a rating of -$\infty$ or $\infty$ respectively. One draw also has the potential to lead to a degenerate solution, as it implies that the strength of the players after one game is equal. This estimation is not meaningful and does not accurately reflect the true strength of a player. In this instance, our MLE estimate fails.

## 2.4 Conclusion

This chapter introduced key concepts relevant to the topic of the thesis. The chapter gives us a definition of MLE and concepts that are important for understanding the model we are going to use. Later in the thesis we will look at possible problems with these estimates of a chess player's rating and possible improvements.

# Chapter 3
# Maximum likelihood estimation problems

## 3.1 Estimating a chess player's rating

In the previous chapter, we introduced a model for estimating a player's rating. Here, we have collected data on the games played by a single player in the Quatar Masters 2023[1], including the scores of the games and the ratings of his opponents. Using this data 3.1, we will estimate the player's rating and examine how the MLE works in this scenario.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 7 | Srihari L R | | 2438 | 1 w | +0.8 | Qatar Masters (1) | 11.10.2023 |
| 8 | Suleymenov, Alisher | | 2512 | 0 b | −8.7 | Qatar Masters (2) | 12.10.2023 |
| 9 | Muthaiah AL | | 2470 | 1 w | +1.0 | Qatar Masters (3) | 13.10.2023 |
| 10 | Pranesh M | | 2515 | ½ b | −3.7 | Qatar Masters (4) | 14.10.2023 |
| 11 | Bharath Subramaniyam H | | 2500 | 1 w | +1.2 | Qatar Masters (5) | 15.10.2023 |
| 12 | Samant Aditya S | | 2511 | 1 b | +1.3 | Qatar Masters (6) | 17.10.2023 |
| 13 | Karthikeyan, Murali | | 2611 | 0 w | −7.9 | Qatar Masters (7) | 18.10.2023 |

**Figure 3.1.** Chess player's games.

Let:
$R_A$ be the player's rating to be estimated.
$R_O$ be the average rating of the player's opponents.
$w_n$ be the number of games won by the player.
$d_n$ be the number of games drew by the player.
$l_n$ be the number of games lost by the player.

Average Elo of all opponents of the player:

$$R_O = \frac{2438 + 2512 + 2470 + 2515 + 2500 + 2511 + 2611}{7} = 2508\,.$$

Estimating the player's rating:

$$\hat{R}_A = R_O - 400 \log_{10} \frac{l_n + \frac{d_n}{2}}{w_n + \frac{d_n}{2}} = 2508 + 400 \log_{10} \frac{2.5}{4.5}$$

Estimated player's rating:

$$\hat{R}_A = 2610\,.$$

A player rating has been estimated, which can be used to compare his strength to other players in the future. However, it is necessary to ascertain the actual rating of the person whose games were provided (see Figure 3.2), and to determine how it differs from the estimate.

[1] `https://www.chess.com/events/2023-qatar-masters/results`


**Figure 3.2.** Chess player's actual rating.

The real player is **Magnus Carlsen** with a rating of 2839 at the time. He is considered one of the greatest chess players of all time. During the Quatar Masters 2023 tournament, he was the highest-rated player in the world.

Using this method, we obtained a rating of 2610, which does not even place the player among the top 100 strongest players[2], despite the original player having the highest rating out of all players.

In accordance with the Elo formula, the player's anticipated expected score is:

$$E_{exp} = \frac{1}{1 + 10^{\frac{R_O - R_A}{400}}} = 0.87\,.$$

The actual expected score is from the dataset:

$$E = \frac{w_n + \frac{d_n}{2}}{w_n + d_n + l_n} \doteq 0.64$$

In the next sections, we will explore the possible reasons for this discrepancy between the actual result and the estimate, as well as potential improvements for our model.

## 3.2 Confidence interval

**Confidence interval** is an interval which is expected to typically contain the parameter being estimated. Given the small size of our dataset, it is advisable to examine confidence interval to account for the significant difference between the estimated and actual player ratings at that time.

The standard error is:

$$err = \left(\frac{\hat{E}(1-\hat{E})}{n}\right)^{\frac{1}{2}}.$$

$$err = \left(\frac{9}{14} \cdot \frac{\frac{5}{9}}{7}\right)^{\frac{1}{2}} \doteq 0.18\,.$$

The confidence interval is calculated by adding and subtracting the standard error from the estimated parameter.

$$Confidence\ interval = \hat{E} \pm err\,.$$

The standard error indicates that the estimated value is expected to fall within the interval.

$$(0.46, 0.82)\,.$$

The expected points gain is expected to fall within the following interval:

$$(3.22, 5.74)\,.$$

[2] `https://ratings.fide.com/top.phtml?list=men`

The rating is expected to fall within the following range:

$$(2480, 2771)\,.$$

However, although we calculated the expected score for the player, our calculated actual score does not fall within the upper bound. Additionally, the player's expected gain should have been 6 when the upper bound for the expected points gain is 5.74, which can be rounded to 5.5.

## 3.3 Analysis of an encountered problem

This section examines potential reasons for the discrepancy in our estimates and identifies areas for improvement to enhance our performance.

### 3.3.1 Average Elo

Using the average Elo rating of all players' opponents can be misleading. Consider the following fictional example. An open chess tournament, open to players of any rating. Estimated rating of a chess player who won a game against a player with a rating of 1958 lost the game to a Stockfish computer with a rating[3] of 3644 on April 2024.
Estimate of the expected score:

$$\hat{E} = 0.5\,.$$

The average Elo of all the player's opponents:

$$R_O = \frac{1958 + 3642}{2} = 2800\,.$$

Based on Table 2.1, our estimate of the rating is 2800, as the expected score is 0.5. By winning a player with a rating of 1958 and losing to Stockfish, the player has the same strength as a player with a rating of 2800, which is the rating of the strongest chess players in history.

This estimate does not provide information about the actual strength of the player. Using the average Elo rating of all players' opponents is a flawed assumption that can lead to complications and inaccurate estimates. However, calculating each game separately can be computationally expensive. Therefore, assuming that all players have roughly equal ratings in one tournament, the method works well.

### 3.3.2 Small dataset

We analysed the player's games in a particular tournament to estimate his rating. However, this estimate may be misleading as we did not consider the player's previous games. Estimating the rating based on all previous games, which could be thousands, may be computationally demanding. Nonetheless, the described method provides a relatively quick estimate.

## 3.4 Possible improvements

One possible improvement for estimating the chess player's rating is to update it, which can increase accuracy. Additionally, we will examine our calculation without using the average Elo of all opponents.

[3] https://computerchess.org.uk/ccrl/4040/

### 3.4.1 Elo actualization

The Fédération Internationale des Échecs (FIDE)[4], which is the international governing body for chess, updates its ratings list at the beginning of each month. In contrast, the unofficial `Live ratings`[5] calculate the change in players' ratings after every game. These Live ratings are based on the previously published FIDE ratings, so a player's Live rating is intended to correspond to what the FIDE rating would be if FIDE were to issue a new list that day.

All top players have a $K$-factor of 10, which means that the maximum ratings change from a single game is a little less than 10 points.

### 3.4.2 Not using the average Elo

The equation we need to solve, which does not involve averaging Elo, is as follows:

$$\hat{E} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + 10^{\frac{R_i - R_A}{400}}} .$$

It is not possible to solve this equation analytically. However, it can be solved numerically using known optimization methods. For this problem, Newton's method [1] will be used, an iterative method for finding the roots of a differentiable function.

The following steps outline the methodology employed by the algorithm to estimate players' ratings:

1.**Initialization:** Start with an initial guess $R_p^{(0)}$ for the player's rating.

2.**Iterative Update:** Iterate through the following steps until a stopping criterion is met:

3.**Calculate Expected Score:** Using the current estimate of $R_p$, calculate the expected score $E$ of the player against the sequence of opponents using the Elo rating formula.

4.**Calculate Derivative:** Compute the derivative of the expected score function with respect to $R_p$ to determine the slope of the function at the current estimate of $R_p$.

5.**Update Estimate:** Update the estimate of $R_p$ using the formula:

$$R_p^{(k+1)} = R_p^{(k)} + \alpha \frac{E - real\ expected\ score}{f'(R_p^{(k)})} ,$$

where $\alpha$ is the learning rate (step size) controlling the speed of convergence. This step ensures that we move towards the root of the expected score function $E$.

6.**Convergence Criterion:** Check if the absolute difference between the estimated expected score and the real expected score falls below a predefined tolerance. If it does, stop iterating, as we have found a satisfactory estimate of the player's rating.

7.**Output:** Once the convergence criterion is met, output the final estimate of the player's rating.

The following Python code illustrates the implementation of the method. The functions expected_score and derivative_expected_score have already been implemented for calculating the expected score and its derivative.

[4] `https://www.fide.com/`
[5] `https://2700chess.com/`

```
def estimate_rating(R_o, E_real, R_p, tol=1e-6, max_iter=100, alpha=0.1):
    for _ in range(max_iter):
        E = expected_score(R_p, R_opponents)

        if abs(E - E_real) < tol:
            return R_p

        dE = derivative_expected_score(R_p, R_o)
        R_p -= alpha*((E_real - E) / dE)

    raise ValueError("Newton's method did not converge")
```

The implementation yielded a rating of 2612 for the initial problem, which is nearly identical to the initial estimate. This indicates that the assumption was effective, even when calculating independently, as the estimate did not improve significantly. The ratings were nearly equal, suggesting that the problems were not related to the average Elo of opponents.

## 3.5 A problem where the player's performance exceeded expectations

Consider another issue to enhance our model analysis. The dataset in Figure 3.3 was obtained from the FIDE webpage[6].

| | | | | | |
|---|---|---|---|---|---|
| ☐ | Malli, Suvirr | | 2126 | SUI | 1.00 |
| ■ | Guyer, Marlon | | 1830 | SUI | 1.00 |
| ☐ | Stopa, Jacek | g | 2351 | POL | 1.00 |
| ■ | Grieve, Harry | m | 2462 | ENG | 0.00 |

**Figure 3.3.** Other Chess player's games.

Let:
$R_A$ be the player's rating to be estimated.
$R_O$ be the average rating of the player's opponents.
$w_n$ be the number of games won by the player.
$d_n$ be the number of games drew by the player.
$l_n$ be the number of games lost by the player.

Average Elo of all opponents of the player:

$$R_O = \frac{2126 + 1830 + 2351 + 2462}{4} = 2192\,.$$

Estimating the player's rating using derived formula:

$$\hat{R}_A = R_O - 400 \log_{10}\left(\frac{1}{3}\right).$$

Estimated player's rating:

[6] https://ratings.fide.com/calculations.phtml?id_number=33387230&period=2024-03-01&rating=0

$$\hat{R}_A = 2383\,.$$

The actual player's rating:

$$R_A = 1892\,.$$

The player is Ashwath Kaushik, in February 2024 he was the youngest player to defeat a grandmaster, at the age of 8 years and 6 months. The difference between our estimated rating and the actual rating is significant, almost 500 points. However, we must consider the small size of our dataset. Therefore, before proposing further ideas for improvement, we should check the confidence interval, as we did in the previous example.
The anticipated expected score is:

$$E_{exp} = \frac{1}{1 + 10^{\frac{R_O - R_A}{400}}} = 0.15\,,$$

The actual expected score is:

$$E = \frac{w_n + \frac{d_n}{2}}{w_n + d_n + l_n} = 0.75\,.$$

The standard error is:

$$err = \left( \frac{\hat{E}(1-\hat{E})}{n} \right)^{1/2},$$

$$err = \left( \frac{0.75 \cdot 0.25}{4} \right)^{1/2} \doteq 0.22\,.$$

The standard error indicates that the estimated value is expected to fall within the interval

$$(0.53, 0.97)\,.$$

The magnitude of the error is considerable, yet the standard error is not informative, as the true expected score does not fall within the specified interval. We can calculate the 99% confidence interval for the true proportion using the Wilson score interval method [5]. The Wilson score interval was developed by E.B. Wilson in 1927. It represents an improvement over the normal approximation interval in multiple respects. Unlike the symmetric normal approximation interval, the Wilson score interval is asymmetric, and it does not suffer from the same problems of overshoot and zero-width intervals that afflict the normal interval. It can be safely employed with small samples and skewed observations. As with the standard interval, the interval can be calculated directly from a formula:

$$Lower\ Bound = \frac{\hat{E} + \frac{z^2}{2n} - \frac{z}{2n} \cdot (4n \cdot \hat{E}(1-\hat{E}) + z^2)^{1/2}}{1 + \frac{z^2}{n}},$$

$$Upper\ Bound = \frac{\hat{E} + \frac{z^2}{2n} + \frac{z}{2n} \cdot (4n \cdot \hat{E}(1-\hat{E}) + z^2)^{1/2}}{1 + \frac{z^2}{n}}.$$

In this context, $z$ represents the standard normal interval half-width, which is equivalent to the desired confidence with a value of 2.576.

Based on the error for 99% confidence interval, our estimated value is expected to fall within the interval

$$(0.22, 0.97)\,.$$

The expected points gain is expected to fall within the following interval:

$$(0.88, 3.84)\,.$$

The rating is expected to fall within the range:

$$(1972, 2744)\,.$$

The error is of a considerable magnitude and has not yielded the desired outcome. Furthermore, the actual expected score does not fall within the expected interval. Even a 99% confidence interval did not assist with our estimate. The expected score's lower bound can be rounded to 1, indicating that the player expected to receive 1 point, while the upper bound is 4 points. However, the actual rating shows that the player's expected score is 0.6, which can be rounded to half a point.

## 3.6 A problem where opponents' ratings differ a lot

This section will examine the open tournament, which involved a diverse range of players with ratings spanning from 900 to 2700. In the tournament sequence, players were paired with opponents with significant rating discrepancies. This may lead to inaccurate estimate in our model, as the average rating of all players is used. A player who wins against a lower-rated opponent and then loses to a higher-rated opponent does not provide an accurate representation of that player's strength when the average opponent's rating is used as a reference point. The tournament under examination is Candidate Blitz 2024[7], and the dataset to be analysed is in Figure 3.4:

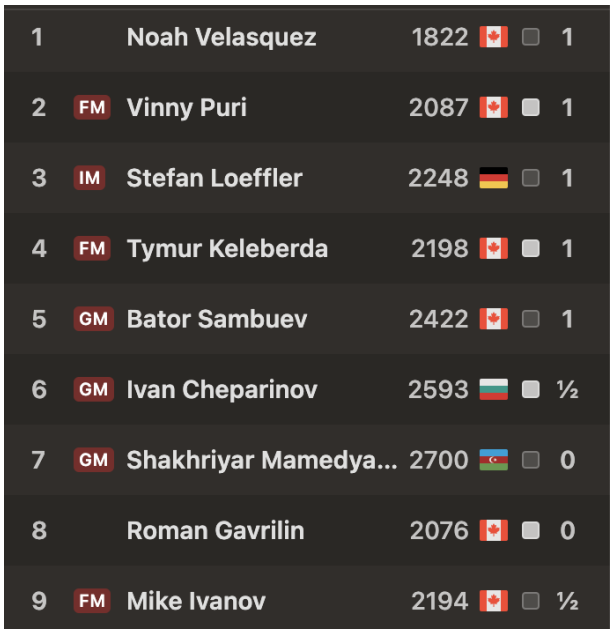


**Figure 3.4.** A dataset of games in which the ratings of the opponents differ significantly.

[7] https://www.chess.com/events/2024-candidates-blitz/results

The following table presents the ratings of the players, with three players exhibiting notably higher ratings than the others. The lowest-rated player has a rating of 1822 points, while the highest-rated player has a rating of 2700 points. The examined player achieved a score of 1.5 out of 3 with the highest-rated players.

Let us now compute the average rating.

$$R_O = \frac{1822 + 2087 + 2248 + 2198 + 2422 + 2593 + 2700 + 2076 + 2194}{9} = 2260\,.$$

Estimating the player's rating using derived formula:

$$\hat{R}_A = R_O - 400 \log_{10}\left(\frac{3}{6}\right).$$

The estimate of his rating is:

$$\hat{R}_A = 2380\,,$$

The actual expected score is:

$$E = \frac{w_n + \frac{d_n}{2}}{w_n + d_n + l_n} \doteq 0.66\,.$$

Let us now examine the standard error,

$$err \doteq 0.157\,.$$

The standard error indicates that the estimated value is expected to fall within the interval:

$$(0.51, 0.82)\,.$$

The estimated rating is expected to fall within the interval:

$$(2267, 2523)\,.$$

The real rating of the player is $R_A = 2660$, which does not fall into the interval. The actual player is Evgeny Bareev.

It is evident that there is a considerable discrepancy between the estimated and actual ratings. The estimated ratings are approximately 300 points less than the actual rating. The estimate indicated a weaker performance than was observed in reality. It can be demonstrated that the inclusion of players with slightly lower ratings in the average rating significantly alters the overall result.

In section 3.4.2, we implemented a numerical method that does not require the average rating of a player's opponents. In contrast, the approach taken is to consider each game separately, which in this case should yield a more accurate estimate. Let us now examine the outcome of this calculation.

The calculated rating, as determined by the Newton's method, is as follows:

$$\hat{R}_A = 2426\,.$$

This method yielded a rating that was more closely aligned with that of the actual player.

The estimate of a player's rating in a tournament where the players have a wide range of ratings can be more challenging when using the average rating. This is due to the fact that in such tournaments, the pairing of players can often result in an unequal distribution of ratings. To illustrate, it will be demonstrated that another competitor in the tournament achieved the same result, but had paired with players having slightly less ratings. This implies that the estimated strength will be slightly different.

Let us now examine another dataset in Figure 3.5.

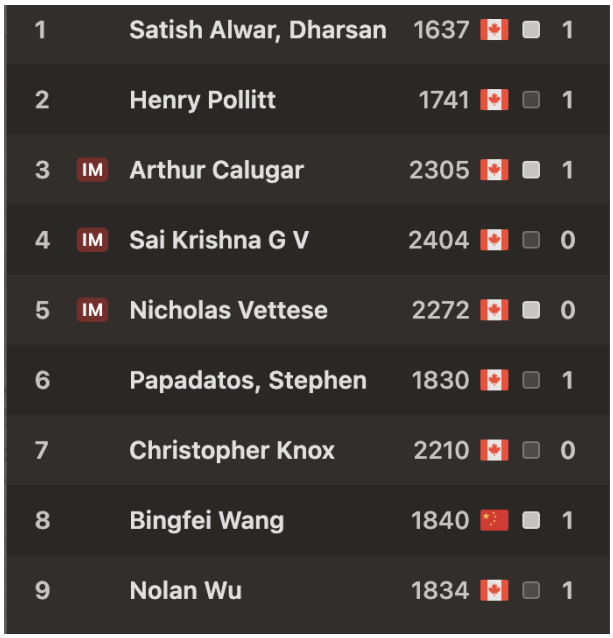


**Figure 3.5.** Same score but different opponent's rating distribution.

In this dataset, the opponent player with the highest rating is 2404, which contrasts with the previous dataset. A new computation of the average rating yields the following result:

$$R_O = \frac{1637 + 1741 + 2305 + 2404 + 2272 + 1830 + 2210 + 1840 + 1834}{9} = 2008\,.$$

Consequently, the estimated rating will be as follows:

$$\hat{R}_A = 2128\,.$$

The player's rating is as follows:

$$R_A = 2107\,.$$

Two players participated in the same tournament and achieved the same score, resulting in their placement in the same position. However, they were paired with opponents whose rating ranges differed slightly. Consequently, the estimated strength is disparate.

In the event of a significant discrepancy in ratings between opponents, it is recommended that a more accurate estimate be obtained by utilising a method that accounts for the ratings of all opponents, such as Newton's method.

## 3.7 Conclusion

In this chapter, we examined various issues that can arise when estimating a chess player's rating, as well as potential challenges. Furthermore, we presented the concept of updating the rating or considering games independently, which can be calculated using the defined numerical method, and demonstrated how this can influence the estimate.

**In the initial problem**, we examined a small dataset in which the player was expected to achieve a superior outcome than was actually observed.

**In the second problem**, we considered the converse scenario, namely, the extent to which unexpected performance can be exhibited by a player with a relatively low rating compared to the opponents.

**In the third problem**, we discussed the potential limitations of averaging ratings of opponents in certain tournaments. It was noted that the ratings of players in open tournaments can vary considerably, which may impact the effectiveness of this approach.

# Chapter 4
# Model improvements

As previously discussed, our model is subject to certain limitations in specific instances. It is possible to enhance the model by addressing the issues identified in the previous chapter. It would be advantageous to examine a different type of problems.

This chapter presents a series of experiments and illustrative examples that demonstrate how the model can be enhanced to incorporate details related to the true strength of the player. This chapter will demonstrate potential avenues for enhancing the model, although it is not the intention to identify an optimal solution. Instead, it will illustrate possible approaches to improving the model.

The initial step will be to analyse a novel problem and the factors that may contribute to discrepancies between estimated and actual player ratings. The subsequent sections will present potential enhancements to the model. One such example is the rating advantage enjoyed by the white player. This discrepancy can be attributed to an unequal distribution of players between the white and black sides.

It is important to note that in computer chess, the colour of the player plays a more significant role. This is due to the fact that computers have much greater computational power. Furthermore, the advantage that computers gain from this is much greater than that of a human player. Consequently, this chapter will present only human chess problems.

## 4.1 Long-term non-participation in tournaments problem

The 13th world chess champion in history, Garry Kasparov, participated in the Zagreb Rapid & Blitz 2021 tournament[1]. He retired from professional chess in 2005 after holding the title of world's best player for a record-breaking 20 years.

Using MLE, we can estimate rating. Let's take a look at the dataset from the tournament in Figure 4.1.

The mentioned player had a rating of 2812 points but only achieved a final score of $\frac{0.5}{9}$. This was despite his opponents having an average rating of 2749.

Estimating the player's rating using the derived formula:

$$\hat{R}_A = R_O - 400 \log_{10} \frac{8.5}{0.5} .$$

Estimated player's rating:

$$\hat{R}_A = 2257 .$$

[1] https://www.chess.com/news/view/grand-chess-tour-zagreb-rapid-blitz-day-4-kasparov

**Zagreb Rapid & Blitz 2021 | Day 4 Standings**

| Rank | Fed | Player | Rtg | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | Rds | Pts |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | Ian Nepomniachtchi | 2792 | ♟ | 1 0 | 1 ½ | 1 1 | 1 1 | 2 0 | 2 ½ | 2 1 | 0 | 1 1 | 1 | 18 | 17 |
| 2 | | Maxime Vachier-Lagrave | 2749 | 1 1 | ♟ | 1 0 | 2 ½ | 2 ½ | 0 1 | 1 1 | 1 ½ | 1 | 1 1 | 1 | 18 | 16.5 |
| 3 | | Anish Giri | 2776 | 1 ½ | 1 1 | ♟ | 2 1 | 1 ½ | 1 ½ | 1 ½ | 2 0 | 0 | 1 ½ | 1 | 18 | 15.5 |
| 4 | | Viswanathan Anand | 2753 | 1 0 | 0 ½ | 0 0 | ♟ | 1 1 | 1 ½ | 1 ½ | 2 1 | 1 | 2 1 | 1 | 18 | 14.5 |
| 5 | | Jan-Krzysztof Duda | 2738 | 1 0 | 0 ½ | 1 ½ | 1 0 | ♟ | 2 1 | 1 ½ | 2 1 | 1 | 1 0 | 1 | 18 | 14.5 |
| 6 | | Shakhriyar Mamedyarov | 2782 | 0 1 | 2 0 | 1 ½ | 1 ½ | 0 0 | ♟ | 1 0 | 1 0 | 1 | 2 1 | 1 | 18 | 13 |
| 7 | | Alexander Grischuk | 2778 | 0 ½ | 1 0 | 1 ½ | 1 ½ | 1 ½ | 1 1 | ♟ | 1 1 | 1 | 1 ½ | ½ | 18 | 13 |
| 8 | | Anton Korobov | 2683 | 0 0 | 1 ½ | 0 1 | 0 0 | 0 0 | 1 1 | 1 0 | ♟ | 2 | 2 1 | 1 | 18 | 11.5 |
| 9 | | Ivan Saric | 2653 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 0 | ♟ | 1 | | 9 | 10 |
| 10 | | Jorden van Foreest | 2688 | 1 0 | 1 0 | 1 ½ | 0 0 | 1 1 | 0 0 | 1 ½ | 0 0 | 1 | ♟ | 1 | 18 | 9 |
| 11 | | Garry Kasparov | 2812 | 0 | 0 | 0 | 0 | 0 | 0 | ½ | 0 | | 0 | ♟ | 9 | 0.5 |

**Figure 4.1.** The tournament leaderboard on which Kasparov played.

The actual expected score is:

$$E = \frac{w_n + \frac{d_n}{2}}{w_n + d_n + l_n} \doteq 0.06\,.$$

The standard error is:

$$err = \left(\frac{\hat{E}(1-\hat{E})}{n}\right)^{1/2},$$

$$err = \left(\frac{\frac{0.5}{9}\cdot\frac{8.5}{9}}{9}\right)^{1/2} = 0.08\,.$$

Based on the standard error, our estimate should fall within the given interval.

$$(0, 0.13)\,.$$

The expected score based on the pre-tournament rating should have been:

$$E_{exp} = \frac{1}{1 + 10^{\frac{R_O - R_A}{400}}} \doteq 0.6\,.$$

The standard error is uninformative and unhelpful. To calculate the 99% confidence interval for the true proportion, we will use the Wilson score interval method, as in the previous chapter. The value of $z$, which is the standard normal interval half-width corresponding to the desired confidence, is equal to 2.576.

Based on the error for 99% confidence interval, our estimate should be within the given interval:

$$(0, 0.49)\,.$$

The error remains significant, and unfortunately, the provided information was not helpful in our case.

Extended periods of non-participation in tournaments may lead to a decline in a player's performance. However, the formula used in this case does not account for this factor.

It is worth noting that the player retired from professional chess in 2005. As the tournament was held in 2021, it is possible that not participating regularly could have had a negative impact on the performance, particularly given the age difference between the player and his opponents.

In this case, it would be reasonable to assume that the player's relative strength has decreased since his last tournament. We can incorporate this into our formula by considering the length of time the player has been absent from tournaments, as well as the number of tournaments played during that time. This can provide a metric for estimating how much weaker the player has become.

## 4.2 Long time no participating in tournaments

In the opening section of this chapter, we discussed the potential impact of a prolonged absence from tournament play on a player's actual strength. Nevertheless, the Elo formula does not appear to account for this phenomenon. In this section, we will attempt to refine our estimation methodology, acknowledging the relevance of expected score in this context. One potential solution to be considered is the number of years that have elapsed since the player's retirement.

In the problem under discussion, the player retired from professional chess in 2005, and the tournament in question was played in 2021. After estimating the player's rating, a reduction of almost 600 points in the player's rating strength is found. A similar technique to that employed in previous section may be employed to include a constant to indicate that the player has not participated in tournaments for an extended period of time. This would suggest that their actual strength is not proportional to their rating, in accordance with the aforementioned assumption. The constant $P$ may be employed to represent a rating penalty in real player strength for not participating in tournaments regularly for an extended period of time.

$$E_A = \frac{1}{1 + 10^{\frac{R_O - (R_A - P)}{400}}}.$$

Prior to examining the parameter $P$, it is necessary to examine a greater number of players and to ascertain how their rating decreased after the peak rating. Subsequently, an attempt can be made to estimate the parameter $P$ based on the length of time that has elapsed since the player retired.

In order to estimate the parameter $P$, a small dataset of experimental data will be employed, comprising players who have been less active since attaining their peak rating. This approach is not the optimal method for estimating this parameter, as it is not a straightforward task. Its purpose is to demonstrate the conceptual possibilities for estimating this parameter.

Table 4.1 indicates a notable decline in the performance of players who were previously rated at a higher level. This decline can be observed over an extended period of time. It is also possible that the decline in rating may be attributed to the natural process of aging. The players in the presented table demonstrated a reduction in their level of activity following the achievement of their peak rating.
It can be observed that there is a correlation between the number of years that have elapsed since a peak rating was achieved and the rating itself.

| Player Name | Peak rating and Date | Rating on 2024 (April) | Points decline |
|---|---|---|---|
| Vasyl Ivanchuk | 2787 October 2007 | 2611 | 176 |
| Anatoly Karpov | 2780 July 1994 | 2617 | 163 |
| Alexander Morozevich | 2788 July 2007 | 2659 | 129 |
| Veselin Topalov | 2816 July 2015 | 2727 | 89 |

**Table 4.1.** Players rating decline over a period of time.

The following Python code calculates the average decline in rating per year for the players from Table 4.1. This allows us to make an assumption about the mean decline after the peak rating and use it for estimating parameter $P$. We can then proceed to examine another problem and how it helped to demonstrate the true strength of the players after retirement.

```
import pandas as pd

# Create a DataFrame
df = pd.DataFrame(data)

# Convert the peak rating date to datetime objects
df['Peak Rating Date'] = pd.to_datetime(df['Peak Rating Date'])

# Calculate the number of years elapsed since the peak rating
current_year = 2024
df['Years Elapsed'] = current_year - df['Peak Rating Date'].dt.year

# Calculate the average decline in rating per year
df['Average Decline per Year'] = df['Points Decline']/df['Years Elapsed']
```

Table 4.2 presents the calculated average decline in rating per year for each player.

| Player Name | Average decline in rating per year |
|---|---|
| Vasyl Ivanchuk | 10.3 |
| Anatoly Karpov | 5.4 |
| Alexander Morozevich | 7.6 |
| Veselin Topalov | 9.9 |

**Table 4.2.** Average decline in rating per year.

In light of the observed average decline, the formula for $P$ is as follows:

$$P = n \cdot D\,,$$

where:
$n$ is the number of years that have elapsed since the player retired or achieved his highest rating.
$D$ is the average decline in rating per year.

The calculated $D$ value for the collected dataset is 8.3. We will now examine another tournament in which Garry Kasparov participated following his retirement. In 2017,

Garry Kasparov participated in the Saint Louis Chess Rapid & Blitz tournament[2], where he achieved a final score of 12.5 out of 27. His average opponent's rating was 2768. Consequently, the estimated rating is:

$$\hat{R}_A = 2743\,.$$

Given the parameter $P = 12{\cdot}8.3$ and the competitor's rating of 2812 at the tournament, it can be reasonably assumed that his true strength should have been:

$$R = 2812 - 100 = 2712\,.$$

Despite the limitations of our small experiment, we were able to demonstrate an improvement in the score, despite the absence of a comprehensive database of players. In this instance, the updated Elo formula demonstrates a more accurate representation of the player's strength following retirement, as it incorporates a longer period of time after the peak rating.

## 4.3 Rising stars and their rating growth

It has been observed that some players who have recently commenced playing chess tend to exhibit a significantly faster rate of rating growth than their older counterparts. For the purposes of this section, I will refer to such players as **rising stars**. There have been numerous instances where these players have demonstrated greater strength than anticipated. This chapter will examine the issue of estimating the strength of rising stars.

In Table 4.2, we observed a decline in the rating of an older player. In this section we will examine a table of ratings for rising stars and their corresponding growth. Let us examine Table 4.3 to ascertain how a player who commenced playing chess at an age of less than 10 years was able to enhance his rating over a 10-year period.

| Player Name | First written Rating and Date | Rating in 10 years |
|---|---|---|
| Firouzja Alireza | 1756 September 2012 | 2778 |
| Praggnanandhaa Rameshbabu | 1346 September 2011 | 2617 |
| Erigaisi Arjun | 1584 September 2011 | 2626 |
| Gukesh Dommaraju | 1291 August 2013 | 2751 |

**Table 4.3.** Rising stars rating growth over time.

It is evident that young players experience rapid growth in their abilities. Let us now examine the tournament performance of one of these players and consider how the rating reflects his true strength.

In the 2018 World Rapid Championship[3], Alireza Firouzja, then 15 years of age, had a rating of 2412 points. He commenced the tournament as the 169th seed, competing against 206 other participants. Despite the considerable disparity in their ratings, he

[2] https://theweekinchess.com/chessnews/events/saint-louis-rapid-blitz-2017
[3] https://www.chess.com/events/2018-world-rapid-championship/results

achieved a score of 10/15 points by playing against opponents with a higher rating. The opponents' average rating was 2723[4].

The estimated rating is as follows:

$$\hat{R}_A = 2843\,.$$

The discrepancy is considerable. The young player demonstrated considerable strength in the tournament, exceeding expectations. The discrepancy is approximately 400 points. His performance rating was the second-best of the tournament, trailing only that of the eventual winner, Daniil Dubov. It would be beneficial to ascertain whether the 99% confidence interval would be of assistance in this instance, and to determine whether the error falls within the interval.

A 99% confidence interval was calculated using the Wilson score interval method:

$$(0.35, 0.88)\,.$$

The rating would be in range:

$$(2615, 3069)\,.$$

His pre-tournament rating was:

$$R_A = 2412\,,$$

The expected score based on the player's pre-tournament rating is as follows:

$$E = 0.14\,.$$

The error remains significant, and unfortunately, the provided information was not helpful in our case. One potential solution is to update the player's live rating, given that the expected score gain was considerably lower than the actual score gain. The player's rating has increased significantly since the tournament. This may result in the player being situated within the desired range.

$$R_{post} = R_{pre} + K \cdot n(\hat{E} - E)\,.$$

The factor of 20 will be employed, as this is the factor officially used by FIDE for rapid and blitz[5],

$$R_{post} = 2570\,.$$

The updated rating gives an expected score of:

$$E_{post} = 0.29\,.$$

Nevertheless, the player's rating did not improve to the extent that our estimate was enhanced. This solution is not effective in this case, as it does not fall within the 99% confidence interval.

It is possible to enhance the Elo formula by incorporating a constant that indicates that a player is young and anticipates a rapid increase in their rating. This implies that the player is anticipated to demonstrate enhanced performance on average, thereby

[4] https://ratings.fide.com/calculations.phtml?id_number=12573981&period=2019-01-01&rating=1
[5] https://ratings.fide.com/calculator_rtd.phtml

exhibiting a strength that is greater than that indicated by the rating. The constant will be designated as $Y$, and the revised Elo formula will be expressed as follows:

$$E = \frac{1}{1 + 10^{\frac{R_o - (R_A + Y)}{400}}},$$

where:
$Y$ is the rating advantage attributed to the young player.

One possible approach to estimating $Y$ is to examine the rate of change in the players' ratings over a specified period, as indicated in Table 4.3. However, it would be beneficial to examine the growth of the rating from the period of time when those players achieved a rating of 2400. It can be observed that after the players achieve a rating of 2400 in the classical format, the factor $K$ remains constant at 10. This implies that the rating growth will be more consistent and the parameter estimate will be more accurate.

Table 4.4 presents the updated data on the achievement of a rating of over 2400 and the rating in 3 years.

| Player Name | Rating and Date of rating over 2400 | Rating in 3 years |
|---|---|---|
| Firouzja Alireza | 2455 January 2016 | 2618 |
| Praggnanandhaa Rameshbabu | 2429 July 2016 | 2540 |
| Erigaisi Arjun | 2458 February 2018 | 2559 |
| Gukesh Dommaraju | 2401 April 2018 | 2578 |

**Table 4.4.** The growth of the rating following the achievement of a rating of 2400.

From Table 4.4 we can see that young players' rating growth is significant over time. From this data the mean value of each year growth will be an estimate of our parameter $Y$. The $Y$ from this data is 66.

## 4.4 Higher probability of white winning

It may be useful to include a small constant, denoted $C$, in our formula to indicate that whites tend to have better expected scores on average than blacks. The constant can be estimated using the MLE method or by analyzing a large database of games to determine the percentage of white and black wins.

Let:
$R_W$ be the rating of the player playing white,
$R_B$ be the rating of the player playing black,
$C$ be the rating advantage given to white.

Then the expected score for a white player is:

$$E_W = \frac{1}{1 + 10^{\frac{R_B - (R_W + C)}{400}}}.$$

The expected score for the black player is:

$$E_B = \frac{1}{1 + 10^{\frac{(R_W+C)-R_B}{400}}} .$$

Let:
$p_w$ be the probability of white winning,
$p_b$ be the probability of black winning.

An estimate of $C$ can be obtained by solving the equation:

$$E_W = \frac{1}{1 + 10^{\frac{R_B-(R_W+C)}{400}}} = \frac{p_w}{p_w + p_b} .$$

The formula for $C$ can be obtained from the equation:

$$C = R_B - R_W - 400 \log_{10} \frac{p_b}{p_w} .$$

To calculate the equation, we shall utilise the $p_w$ and $p_b$ values provided in Table 2.2, as it represents the most frequently used openings in the database. For equal ratings, with $p_w = 0.3$ and $p_b = 0.2$, the following is obtained:

$$C = 70 .$$

On average, white players have a 70-point rating advantage according to our estimate.

One potential solution is to assign a value of $C$ that is proportional to the number of games in which the player or his opponents played white. This can be expressed as follows:
Let:
$n_p$ be the number of games player played white.
$n$ be the total number of games.

The value of $C$ for the player is:

$$C_p = \frac{n_p \cdot C}{n} .$$

The value of $C$ for the opponents is as follows:

$$C_o = \frac{(n - n_p) \cdot C}{n} .$$

The new formula for expected score looks like this:

$$E_p = \frac{1}{1 + 10^{\frac{(R_o+C_o)-(R_p+C_p)}{400}}} .$$

Let us now attempt to devise a formula for estimating ratings with the introduction of new parameters, and examine the differences between this and the original formula for rating estimate.

The derived formula is as follows:

$$\hat{R}_p = R_o + (C_o - C_p) - 400 \log_{10} \left( \frac{l + \frac{1}{2}d}{w + \frac{1}{2}d} \right) .$$

We can also write the formula the following way:

$$\hat{R}_p = R_o + A_w - 400 \log_{10} \left( \frac{l + \frac{1}{2}d}{w + \frac{1}{2}d} \right) .$$

The rating advantage or disadvantage, designated as $A_w$, is a value assigned to the rating of the player whose rating is to be estimated. This value is calculated according to the following formula:

$$A_w = C_o - C_p .$$

The sole distinction between this formula and the previous one is the inclusion of a difference between constants for a white advantage. This formula is logical and coherent. The greater the advantage enjoyed by opponents when playing white, the greater the rating difference between the player and the expected rating of the opponent.

Let us now attempt to estimate the rating of the player in the first problem once more. We know that the player who managed to win this player played white, thereby enjoying an advantage. Furthermore, the player who managed to achieve a draw was of a lower rating but also had the advantage of playing white. Finally, the player won another player while he was playing black. In conclusion, all of these factors can be used to improve our estimate, thus placing it within the confidence interval. Nevertheless, it is uncertain how to incorporate the strength of the player into the estimation formula. The average rating is calculated based on a set of games, rather than considering each game independently.

The revised estimate of the rating for the problem in Section 2.1 is as follows:

$$\hat{R}_A = 2620 .$$

This has led to an improvement in our estimate. However, in our case, it would be advisable to utilise Newton's method with a minor enhancement, whereby a constant $C$ is incorporated into the rating of the individual player who played white in the specific game. A better solution would be to calculate the ratings separately for each game. This is because the previously estimated formula would be less accurate in this case, as a player could play half of his games as white and the other half as black. This would result in the formula being ineffective, as it would only be accurate when the player has played with their particular colour for a greater proportion of the games.

Accordingly, the revised formula for Newton's method will be as follows:

$$\hat{E} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + 10^{\frac{R_i + \delta_i \cdot C - R_A}{400}}} ,$$

where $\delta_i$ is defined as follows:

$$\delta_i = \begin{cases} 1, & \text{if player } i \text{ played white,} \\ -1, & \text{otherwise.} \end{cases}$$

The same code as in Section 3.4.2 is employed, with the additional inclusion of a constant $C$ for players who played white, with the value of $C$ set to 70. The result is as follows:

$$R_A = 2607 .$$

In this instance, the additional information does not alter our estimate.

From this section, it can be seen that in a series of games in a real tournament, incorporating the rating advantage of a white player as in the regular tournament is not a sensible approach. In the regular tournament, games played with a certain colour account for approximately half of all games, and thus do not significantly influence the estimate change. However, this approach may be more appropriate in instances where the player has participated in the majority of games with one colour.

## 4.5 Conclusion

The initial approach permitted a more precise estimate of the strength of retired players. The method can be employed in the re-rating of players who have resumed participation in chess after an extended period of inactivity. This preliminary investigation employed a relatively modest dataset, which yielded encouraging results.

In the second problem, a study was conducted on the observation of young players and their accelerated rating growth. This factor has the potential to influence the outcome of a series of games. The performance of young players tends to be less predictable than that of players who have been engaged in the sport for a longer period of time.

Finally, the attempt to include a white player advantage demonstrated only minor influence. This is evidenced by the observation that in the majority of tournaments, players utilise a single colour for approximately half of their games. Nevertheless, it can be employed in certain instances where the player has only one colour available for the majority of the games or when the ratings of opponents with black and white differ significantly. This may result in a discrepancy in performance of current method, compensated by our proposed correction.

# Chapter 5
# Discussion

An investigation was conducted to identify potential issues with MLE, with a focus on chess problems. The objective was to ascertain the underlying causes of these issues. Nevertheless, it is necessary to consider a few additional points and to discuss potential future developments.

## 5.1 Different time formats

The thesis did not address the issue of different time controls and their implications for the rating and games played under these controls. It is important to note that the $K$ factor for blitz and rapid chess is 20, whereas the $K$ factor for classical chess is 10 for players who are rated above 2400.

Additionally, players in the blitz and rapid formats are able to participate in a greater number of games due to a shorter time control. However, the accuracy of their moves is often inferior to that observed in a classical game, as players have less time to consider their options.

## 5.2 Different rating systems

A lot of alternative chess rating systems and enhancements to the Elo formula exist, exemplified by the use of distinct online chess platforms such as chess.com or lichess, which employ their own rating systems to facilitate effective player pairing.

It is important to note that the rating assigned to an individual by a chess federation may vary. To illustrate, as of April 2015, Hikaru Nakamura had a FIDE rating of 2799 and a USCF rating of 2881. It should be noted that the Elo ratings of these organisations are not always directly comparable. Additionally, the USCF employs distinct $K$ factors that differ from those utilized by FIDE. The focus of this thesis is on the rating system used by FIDE.

## 5.3 Simplifying assumptions

From a contemporary perspective, Elo's simplifying assumptions are no longer necessary because computing power is inexpensive and widely available. Nevertheless, the computational simplicity of the Elo system has proven to be one of its most valuable assets.

## 5.4 Computer chess

Computer chess is a distinct discipline that is challenging to compare with human chess. The advantage of the white pieces for computers is considerable, as they are capable of calculating deeply and thus identifying potential moves for a draw with great ease.

Additionally, it is challenging to ascertain the rating of a computer program, as it is contingent upon the database of openings, as well as the hardware of the computer. To illustrate, the computer Deep Blue, created by IBM, which defeated Garry Kasparov in 1997[1], was designed with the specific intention of challenging Kasparov.

## 5.5 The model can be used for more than just chess

- The European Go Federation (EGF)[2], which is the governing body for the sport of Go in Europe, has adopted an Elo-based rating system. This system was initially pioneered by the Czech Go Federation.
- In tennis, the Elo-based Universal Tennis Rating (UTR)[3] is a system that rates players on a global scale, regardless of age, gender, or nationality.
- Many video games use modified Elo systems in competitive gameplay.

[1] https://www.chess.com/article/view/deep-blue-kasparov-chess
[2] https://www.eurogofed.org/
[3] https://www.utrsports.net/pages/rankings

# Chapter 6
# Conclusion

The thesis examined the MLE problems that can arise when estimating chess players' ratings. In Chapter 2, the essential terminology used in the thesis is introduced, including the chess rating system and the model's intended use for rating estimation. Subsequently, the issues encountered are examined. One of the most significant challenges is the replacement of opponents' ratings with a single value for the average Elo rating.

In chapter 3, we examined the impact of this phenomenon on rating estimates and analysed the underlying causes. We also presented potential solutions and estimated the magnitude of the associated error. Even when utilising a 99% confidence interval, estimation remained challenging for certain problems.

In Chapter 4, we considered other problems and factors that are not directly related to the averaging of opponents' ratings. It is not always the case that averaging Elo is the problem, and the assumption works well in tournaments where opponents' ratings are similar.

The thesis presented a method for estimating a player's rating, highlighting potential limitations and considerations. The thesis utilised the FIDE rating system.

# References


[1] Gil, Amparo, Javier Segura, and Nico M Temme. *Numerical methods for special functions*. SIAM, 2007.

[2] Glickman, Mark E. A comprehensive guide to chess ratings. *American Chess Journal*. 1995, Vol. 3, No. 1, pp. 59–102.

[3] Glickman, Mark E. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society Series C: Applied Statistics*. Oxford University Press, 1999, Vol. 48, No. 3, pp. 377–394.

[4] Glickman, Mark E, and Albyn C Jones. Rating the chess rating system. *Chance-Berlin then New York*. Springer International, 1999, Vol. 12, pp. 21–28.

[5] Wilson, Edwin B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*. Taylor & Francis, 1927, Vol. 22, No. 158, pp. 209–212.

# Appendix A
# AI Assistants

This appendix provides a brief overview of the artificial intelligence (AI) tools used in the research and writing process of this bachelor thesis. The use of these AI tools aligns with the guidelines and permitted extent outlined in the "Framework Rules for the Use of Artificial Intelligence at CTU for Study and Teaching Purposes in Bachelor and Follow-up Master Studies" document (issued on 29th January, 2024).

**DeepL**[1] has been used for the correct use of academic English.

[1] https://www.deepl.com/write