Absolute Skill with Bayesian Inference

Pedro Pedrosa
Department of Computer Science
Faculty of Sciences of University of Porto

January 15th, 2024

1 Introduction

In the end of our Statistical Inference Course we scratched the surface of **Bayesian Statistics**. I'll now proceed to apply its core principles to obtain valid models to create an **absolute scale** for a certain player's skill, applied to the domain of chess[1], using as basis the work done by Ken Regan, et al. **[1]**.

All auxiliary code was done in **Python** and was sent as attachment.

2 Motivation

Most skill quantification systems rely heavily on describing a player by a number determining their strength $\equiv$ rating. This rating is modelled so that the expected outcome of a game between players X and Y depends strictly on their ratings[2] - which is able to successfully create a **relative strength** scale to classify the players. At first glance this looks to be satisfactory enough, as we're able to, for example:

- Have a good basis for matchmaking players creating, in principle, fair matches;
- Establish a clear player ranking;

However, as stated, this is still all **relative**! If we simply shifted all ratings by a constant c we wouldn't lose any information - a rating by itself has no meaning, and this has certain disadvantages, such as:

- **Geographical Rating Stagnation**: For example, strong players in less developed countries will have their rating stagnated just because they can't play better opponents.
- **Quantifying Player's Strengths**: Evaluating whether a player is better in the "opening", "middle" or "end" phases of the game;

[1] Without loss of all generality, as with some adaptions this can be applied to a some other sports/activities.
[2] Usually these systems are backed up by the **Bradley-Terry model** for paired comparisons **[2]**.

- **Cheater Detection**: Detecting whether a player is playing suspiciously well, and perhaps cheating [3];
- **Timeless Comparison**: Establishing more accurate comparison between players of different eras.

Therefore, if we're able to obtain a rating model rooted upon an absolute scale, we could in principle address these and a lot of more problems!

3 Theoretical Basis

For such an absolute scale, as stated before, I'll explore a way to construct it supported by the **Bayesian Inference** approach in **[1]**. Let's start by defining our problem more concretely: to have an absolute scale, we'll be aided with a computer which I'll define as a "**perfect player**". If we then create a **parametric fallible model**, we can create our absolute strength scale by comparing it to our perfect entity.

Let's now formally define our tools. Our "see-all" program is labeled as a **reference chess engine** E - it's armed with an analysis function $f_E : p \to \{(m_i, v_i)\}$ - that is, we can evaluate a position p and obtain a value v_i for every legal move m_i in p.

Now that we have a reasonable approximation for the perfect player, we can use it to model the human players (fallible) model: we create a **stochastic chess engine** $E(c)$ [4]- while E always plays to perfection, $E(c)$ can play any of the available m_i moves in position p, each with a probability defined by a likelihood $L\big[E(c), (p, m_i)\big]$. $E(c)$ must satisfy some properties **[4]**:

- $E(c = 0) \Longrightarrow L = 1$, after normalization we see that this is equivalent to a random-moving player!
- $E(c \to \infty)$, in a similar argument, leads to the perfect E player;
- $c > 0 \Longrightarrow$ better moves are more likely than worse moves;

L is defined in the support paper **[1]** such that:

$$L(\text{E(c) plays } m_i \text{ in p}) = (v_{max} - v_i + K)^{-c}$$

Where $v_{max} \equiv$ the maximum value of all the possible moves, $v_i \equiv$ the chosen move's value, c is our fallibility parameter and K. To obtain a valid PDF we simply normalise it:

$$P\big[E(c), (p, m_i)\big] = \frac{L[E(c), (p, m_i)]}{\sum_{m_j} L[E(c), (p, m_j)}$$

To make sense of this L let's fix v_{max} and plot the different c-likelihood curves in function of v_i:

[3] Quite the buzz recently - https://www.nytimes.com/2023/12/25/crosswords/chess-hikaru-vladmir-kramnik-cheating.html

[4] $c \equiv$ competence, first introduced by **[3]**

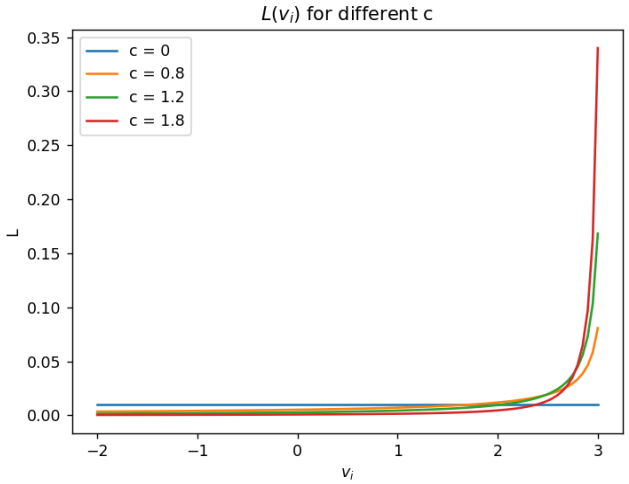

Figure 1: $L(v_i)$ for different c, $v_{max} = 3$

This seems valid! As we can see, $c = 0 \implies$ we have a random player, $c \to \infty \implies$ we tend to a dirac-delta located on v_{max}, and we have a transitory middle ground $\propto x^{-c}$ as $\uparrow c$.

Nevertheless, in this work I'll explore a **different likelihood function** to explore a different transition:

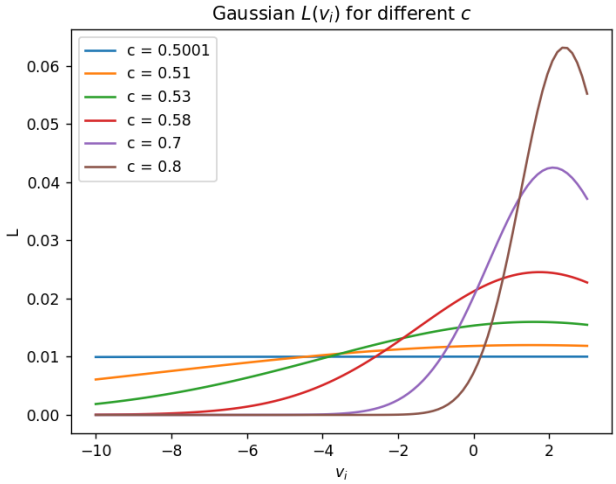

Figure 2: Gaussian $L(v_i)$ for different c, $v_{max} = 3$

The reasoning used was modelling the above - a **Gaussian-likelihood** - so that we still keep the aforementioned properties. We can accomplish this by:

- Restricting $c \in [0.5, 1]$ [5]

[5]Hence defining a perfect player with $c = 1$ and a random player with $c = 0.5$.

- We then obtain the above transition if $\mu(c) = v_{max} \cdot c$ and $\sigma(c) = \frac{1}{c-0.5} - 2$

Which leaves us with **Gaussian-likelihoods** that "shrink and slide"[6] as we increase c-values[7], something that should be interesting to investigate.

4 Inference of the c-distribution

Defining data as the event $e = (m, p)$, by simple application iteration of **Bayes Theorem [5]** we can perform the model's inference:

$$P\big[E(c)|e_i\big] = \frac{P\big[E(c)|e_{i-1}\big] \cdot P\big[e_i|E(c)\big]}{\sum_c P\big[E(c)|e_{i-1}\big] \cdot P\big[e_i|E(c)\big]}$$

Essentially calculating the **posterior distribution** and using it as the **prior** when we're presented with more data, to fit a better model.

I'll use the **"know-nothing" initial uniform prior distribution** - armed with all this, we're able to compute our desired c-distributions.

5 Results

Now that the methodology has been addressed, let's try to compute these c-distributions to define our absolute rating scale for certain players, (dataset as attachment) and use this to [8]:

- **A)** Establish an **absolute** comparison between players of different skills;
- **B)** Detect suspiciously over-performing players - **cheaters**!
- **C)** Compare the world champions of **completely different eras**;

5.1 A - Skill Rating Comparison

Now that we have our new scale, the first thing that pops to mind is to establish a rating comparison between players of different skill levels, by comparing their c-distributions. First off comparing it with the **traditional Elo Rating**, let's evaluate a benchmark for elite level players (rated > 2700) and compare it with a benchmark beginner-intermediate level of players around 1000 rating:

[6] Credit to my colleague Francisco Ferreira for an interesting discussion on this topic.

[7] As it should, this L also converges to a dirac-delta located in v_{max}.

[8] There are endless possibilities here, but I chose these for illustration purposes.

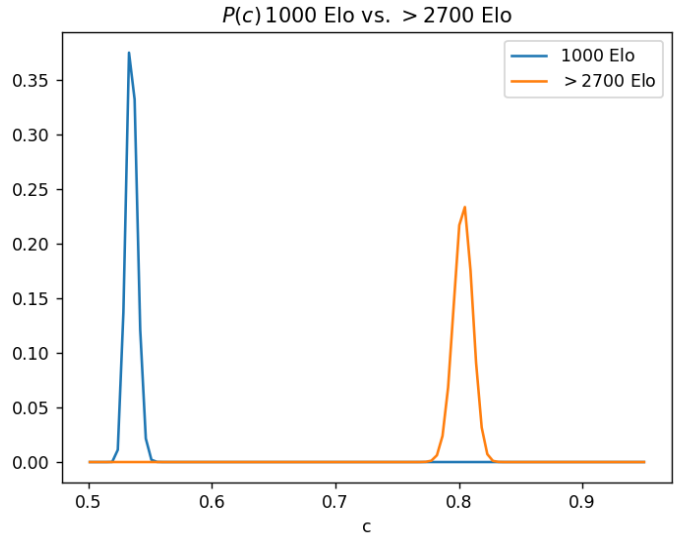

Figure 3: Beginners vs. Elite Players' c-distributions

1000 ELO	> 2700 ELO
$\mu = (0.5350 \pm 0.0010)$	$\mu = (0.8033 \pm 0.0526)$
$\sigma^2 = 0.0002$	$\sigma^2 = 0.0720$

Figure 4: μ and σ for both groups

We see two clearly different distributions; the latter has slightly converged better, but either way this provides us with a reasonable metric for what are **consistently** humanly attainable performances, in this scale.

5.2 B - Cheating Detection

For this part, generating c-distribution for a certain player gives us the necessary information about him to detect over-performances: if we have a collection of games where the player is suspected of using external assistance, we can simply create a c-distribution for said games and perform a test such as **Kolmogorov-Smirnov** with a significance $\alpha = 2.857 \cdot 10^{-7}$ (corresponding to at least 5σ [9]).

For the sake of example, I'll be the test subject here [10]:

[9] To confirm cheating we have to be quite certain!

[10] I cheated by using stronger computer assistance... against computer opponents.

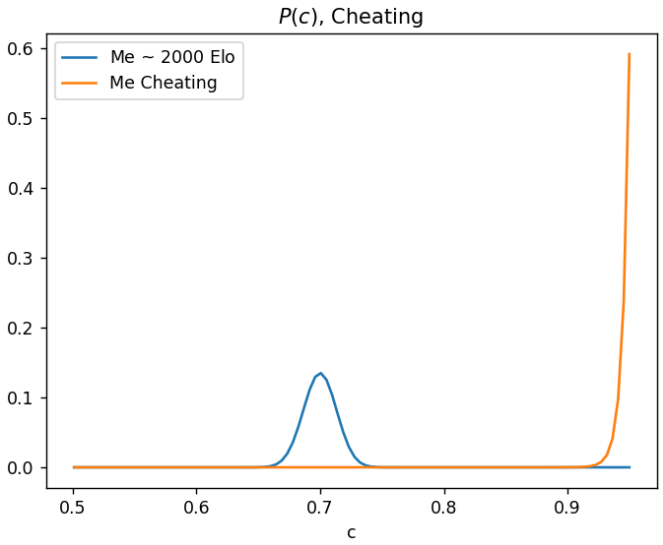

Figure 5: Me vs. Cheater Me c-distributions

Me	Cheater Me
$\mu = (0.6997 \pm 0.003)$	$\mu = (0.94678 \pm 0.0010)$
$\sigma^2 = 0.0002$	$\sigma^2 = 0.00003$

Figure 6: μ and σ for both distributions

```
p_value(KS) = 2.2087606931995054e-59
```

Figure 7: p-value obtained with **KS** test

In this case my cheating was obvious, and clearly detected here. Nowadays the problem isn't detecting this, rather detecting **smart cheating**.

A difference between an **elite player** and a **world champion** of a certain activity is in the **details** - if an elite player was able to cheat **"once" per game**, getting information for the best move in the most complex and critical position of the game, this could be what could catapult him to the top, and this **smart cheating** surely wouldn't be detected by this approach and will be addressed more carefully in a future, more careful look into the topic (implementing an additional parameter h that would characterise how easy it is for a human to find such a move would be a start).

5.3 C - World Champion Comparison

At last, we'll use this metric to compare the overall performance of world champions from two different eras.

Comparing arguably the current world's strongest player Magnus Carlsen and 1920's world champion José Raul Capablanca, we obtain the following c-distributions:

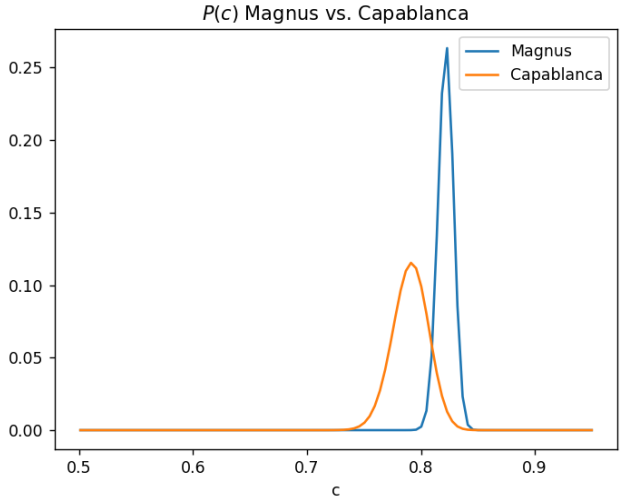

Figure 8: c-distributions for Magnus and Capablanca

Magnus	Capablanca
$\mu = (0.820 \pm 0.001)$	$\mu = (0.7906 \pm 0.0010)$
$\sigma^2 = 0.001$	$\sigma^2 = 0.0246$

Figure 9: μ and σ for both distributions

We see a statistically significant difference [11] between the means of both players, as expected: overall chess study and computers significantly increased the level of play since 100 years ago.

[11] I performed a t-test between two generated samples from both distributions, p-value ≈ 0.

6 Conclusion

The proposed **absolute rating scale** was sucessfully implemented, following the structure of the support paper, albeit using a **different likelihood function**. This scale has a multitude of possible applications and extensions. For example, by having an absolute scale we address inflationary trends, "region-locked" ratings, detect cheaters more systematically, etc.

This metric seems to have great potential for auxiliary purposes, but has also some **drawbacks**: we can't satisfyingly use this in a real tournament, [12] since this scale takes a reasonable amount of data/time to converge and stabilise (doesn't produce instantaneous output such as win/loss).

Like I pointed before, there are really a **LOT** of extension tools that can be created from an absolute scale, that I look forward to explore in further work!

References

[1] - "Skill Rating by Bayesian Inference" - Giuseppe Di Fatta, Guy McC. Haworth and Kenneth W. Regan

[2] - "Rank analysis of incomplete block designs. The method of paired comparisons," - R. A. Bradley and M. E. Terry

[3] - "Reference fallible endgame play" - Guy McC. Haworth

[4] - "Gentlemen, stop your engines!" - Guy McC. Haworth

[5] - "Refresher on Bayesian and Frequentist Concepts" - George Casella

[12] Explaining players they don't gain rating by beating their opponent, rather by mimicking perfect computer-like play seems rather insane...