Do professional chess players draw too often?

Tudor Muntianu

March 7, 2022

Description of changes

I changed all the models to be in principally game-theoretic terms. The models were rewritten, much of the code has changed, but the basic ideas and underlying assumptions have remained. The Elo rating system is still used, but I also modified it to allow it to represent suboptimal play and inform the predictive logistic regressions that are described later.

The only piece of feedback I was unable to incorporate was running simulations to generate data (though I could have done this after training the regressions, it would not have been useful). This is principally because of the way the Elo system works, and its description of an expected value rather than as probabilities of winning, drawing, or losing. So without making baseless assumptions (that would have resulted in useless simulations), generating simulated data would not have been useful. There was also plenty of real data to work with as well.

All code can be found here.

1 Introduction

The prevalence of draws in professional chess has been steadily increasing over the past couple decades. As play becomes more precise and accurate due to extensive computer preparation and analysis, and as more money enters the sport, winning edges are diminished and players are increasingly incentivized not to lose. For instance, at the 2018 World Chess Championship, where Magnus Carlsen played Fabiano Caruana (World No. 1 and No.2 respectively), all 12 classical time-control games were drawn. The championship was decided by a rapid chess tiebreaker. Near-perfect play is not the only factor, however. Since tournament fates, prize pools, and sponsorship deals hinge on performance, losing is highly undesirable. So, the popular argument goes, players play for and agree upon too many draws. This opinion is shared by many strong players and professionals, notably Susan Polgar, a very strong grandmaster who argues that agreed draws should be eliminated from the game. While some agreed

draws are legitimate reflections of drawn game state, others occur when players simply play out opening moves and subsequently agree to draw to minimize variance. This sort of behavior is most egregious, but it demonstrates the most extreme form of this behavior, making it likely that playing passively or "soft-playing" for draws also occurs.

The question we want to answer is: are draws in high-level chess a result of optimal play? or are players sacrificing chances to win to reduce variance as a result of external pressures? Playing "solid" chess (i.e., slow, methodical, less aggressive) that lends itself to more draws is not what we are investigating, but rather if players are playing worse as a result of draw incentives. In other words, are players playing below their true strengths as a result of playing for draw? Put even another way, are players are playing to maximize the single-game outcome (i.e., win if possible, draw if losing), or are players sacrificing performance to reduce their variance?

In this paper the frequency of draws is estimated and compared to existing estimates, namely the very robust Elo rating system used in professional play. We begin by adapting David Smerdon's Fighting Chess Index to generate latent variables representing player behavior and aggression, then use those latent variables in a simple Bayesian model to estimate the probability of winning, drawing, and losing. Then, using the Elo rating system, we compute the expected value of the game's outcome, and compare our estimated probabilities to those generated by ratings to determine whether drawing behavior results in poorer results.

2 Optimal chess

Given two players W and B, we want to know $p = (p_{\text{win}}, p_{\text{draw}}, p_{\text{lose}})$, where p_{outcome} is the probability of W (white) winning, drawing, and losing respectively.

The Elo system, while it does not give a perfect estimate of p, does shed some light.

2.1 Elo ratings

The Elo rating system gives an excellent estimate of the strength of a player. With sufficient games played against many opponents, its application to chess produces robust results. All major chess governing bodies (USCF and FIDE) use Elo as their rating system, chess engines are benchmarked using Elo, and it is used ubiquitously in tournaments to determine seeding and matches.

Given the ratings R_A and R_B of two players A and B respectively, the Elo

rating system predicts game outcomes as follows.

$$Q_A = 10^{R_A/400}, Q_B = 10^{R_B/400}$$
$$E_A = \frac{Q_A}{Q_A + Q_B}$$
$$\text{where } E_A = p_{\text{win}} + \frac{1}{2}p_{\text{draw}}$$

It follows from this that $E_A + E_B = 1$.

2.2 Equilibria under Elo assumptions

The Elo rating system makes an implicit assumption that draw frequency does not directly affect the expected score E, as it does not define p_{win} directly.

There are infinite solutions to $E = p_{\text{win}} + \frac{1}{2}p_{\text{draw}}$. Under Elo assumptions players can play solid passive chess and solid aggressive chess, changing the probabilities of winning and drawing but importantly *not changing* the total expected value E.

This makes any set of strategies chosen by both players a Nash equilibrium under Elo assumptions, since the expected score for any two strategies is equal. Intuitively, under Elo's system, players can play well aggressively and they can play well passively; the choice of playstyle will not affect the outcome.

But if players play an overly passive style to achieve more draws, they decrease their total expected value E to increase p_{draw}. This is the sort of behavior we are interested in, whether players play too passively or draw too often *at the expense* of playing optimal chess.

This is Polgar's prediction, as described here; i.e., real-world circumstances encourage drawing over playing optimally. As a result of chess being played suboptimally, we would expect that as a result of a player playing for a draw more aggressively, they sacrifice some results. In other words, if you play to minimize variance, you decrease your expected score. To describe this, we introduce a modified Elo system.

3 (Potentially) suboptimal chess

First we describe the framework for describe hyperpassive, suboptimal chess before discussing potential equilibria.

3.1 Modified Elo game

There are two players, W, B, white and black, with ratings R_W, R_B respectively.

Each chooses how "overly passive" to play, that is, how much they want to sacrifice expected value for increased draw percentage.

For both W, B, their strategies are to choose

$$\alpha_W, \alpha_B \in [0, 1]$$

Their payoffs, Π_W, Π_B, are represented by the following modified Elo equations. Similar to the standard Elo equations above, they represent the expected score of each player, but instead the expected score is scaled by α_W, α_B.

$$Q_W = 10^{R_W/400}, Q_B = 10^{R_B/400}$$
$$\Pi_W = \frac{\alpha_W Q_W}{\alpha_W Q_A + \alpha_B Q_B}$$
$$\Pi_B = \frac{\alpha_B Q_B}{\alpha_W Q_A + \alpha_B Q_B}$$

If a player A chooses $\alpha = 1$, then they play optimally, that is $\alpha_A Q_A = Q_A$. If they choose $\alpha = 0$ then they "turtle," give up all chances of winning only to increase the likelihood of drawing. Importantly, if $\alpha_W = \alpha_B$, that is if both players are playing equally passively, then $\Pi_W = E_W$ and $\Pi_B = E_B$; their payoffs remain the same as in the standard Elo system, as they are *both* playing equally suboptimally.

Importantly, scaling Q_W, Q_B by α_W, α_B is equivalent to changing each player's effective rating by $R_W + 400 \log_{10}(\alpha_W), R_B + 400 \log_{10}(\alpha_B)$. This fact is used in the code when estimating α_A, α_B.

This difference in rating is not the only effect that α_A, α_B have, since this modified Elo description deals only with suboptimal play and changing expected values; the issue of changing draw probability will be addressed in the next section.

3.2 Estimating outcome probabilities

To estimate $p = (p_{\text{win}}, p_{\text{draw}}, p_{\text{lose}})$, we initially use a multinomial regression as follows.

$$\boldsymbol{x} = (R_W, R_B)$$
$$p_{\text{win}} = \frac{\exp(-\boldsymbol{\beta}_w \cdot \boldsymbol{x})}{1 + \exp(-\boldsymbol{\beta}_w \cdot \boldsymbol{x}) + \exp(-\boldsymbol{\beta}_d \cdot \boldsymbol{x})}$$
$$p_{\text{draw}} = \frac{\exp(-\boldsymbol{\beta}_d \cdot \boldsymbol{x})}{1 + \exp(-\boldsymbol{\beta}_w \cdot \boldsymbol{x}) + \exp(-\boldsymbol{\beta}_d \cdot \boldsymbol{x})}$$
$$p_{\text{lose}} = 1 - p_{\text{win}} - p_{\text{draw}}$$

To account for α_A, α_B changing draw probabilities, we make the following modifications to $p_{\text{win}}, p_{\text{draw}}, p_{\text{lose}}$, beginning with the rating adjustments as described

in the previous section.

$$R_{W_{\text{adj}}} = R_W + 400\log_{10}(\alpha_W), R_{B_{\text{adj}}} = R_B + 400\log_{10}(\alpha_B)$$
$$\boldsymbol{x}' = (R_{W_{\text{adj}}}, R_{B_{\text{adj}}})$$
$$\sigma(x) = \frac{1}{1+\exp(-x)}$$
$$\gamma(\alpha_W, \alpha_B) = 1 - \frac{\sigma(\alpha_W)\alpha_W + \sigma(\alpha_B)\alpha_B}{\sigma(\alpha_W) + \sigma(\alpha_B)}$$
$$p_{\text{win}_{\text{adj}}} = \frac{\exp(-\boldsymbol{\beta}_w \cdot \boldsymbol{x}')}{1+\exp(-\boldsymbol{\beta}_w \cdot \boldsymbol{x}') + \exp(-\boldsymbol{\beta}_d \cdot \boldsymbol{x}')}$$
$$p_{\text{draw}_{\text{adj}}} = \frac{\exp(-\boldsymbol{\beta}_d \cdot \boldsymbol{x})}{1+\exp(-\boldsymbol{\beta}_w \cdot \boldsymbol{x}') + \exp(-\boldsymbol{\beta}_d \cdot \boldsymbol{x}')} + \gamma(\alpha_W, \alpha_B)\left(p_{\text{win}} - p_{\text{win}_{\text{adj}}}\right)$$
$$p_{\text{lose}_{\text{adj}}} = 1 - p_{\text{win}_{\text{adj}}} - p_{\text{draw}_{\text{adj}}}$$

The function γ is a sigmoid-weighted average of the two α parameters. It represents the fraction of draw probability that is gained by sacrificing win percentage. In other words, it represents the decrease in variance that comes at the cost of decreasing expected value.

If both players choose the same α, then the weighted average will come out to just α, otherwise the average biases toward the player playing more optimally (that is, the player with the greater α). That is, it is more punishing to the player playing more suboptimally passively; the player playing worse does not receive a proportional increase in draw percentage.

3.3 Defining α_W, α_B

For modelling player behavior, the following feature variables were used, inspired by David Smerdon's Fighting Chess Index. Variables computed for each player in the dataset were

- total games
- average rating
- average absolute difference between player ratings
- average number of draws
- average number of draws in 30 moves or less
- average number of draws in 30 moves or less when player was White
- average number of moves in drawn games

Smerdon uses these features as inputs to PCA to compute an index of player aggression. Here, we use these features to estimate α_W, α_B.

We add two vectors of weights $\boldsymbol{w}_W, \boldsymbol{w}_B$ to the model to represent linear combinations of the above features, $\boldsymbol{f}_W, \boldsymbol{f}_B$ for white and black respectively, with

intercept terms δ_W, δ_B. Importantly, all features above were z-scored.

$$\alpha_W = \delta_W + \boldsymbol{w}_W \cdot \boldsymbol{f}_W$$
$$\alpha_B = \delta_B + \boldsymbol{w}_B \cdot \boldsymbol{f}_B$$

3.4 Learning parameters

There are no hyperparameters in the model. All parameters $\boldsymbol{\beta}_d, \boldsymbol{\beta}_w, \delta_W, \delta_B, \boldsymbol{w}_W, \boldsymbol{w}_B$ are learned through maximum likelihood estimation according to the computed multinomial distribution $p = (p_{\text{win}_{\text{adj}}}, p_{\text{draw}_{\text{adj}}}, p_{\text{lose}_{\text{adj}}})$.

For more information see the GitHub repo.

4 Results

4.1 Data

The dataset used was Caissabase, a free database of 5.61 million professional chess games. Simul games, Chess960 (a chess variant) games, and consultation (team) games were removed. Games played by players with fewer than 10 games were also removed. Games played by chess engines were also removed, as were games with invalid player names.

Player names were cleaned and standardized using Levenshtein distance fuzzy string matching, see this Python package for more details.

This left 4,695,527 games in the dataset.

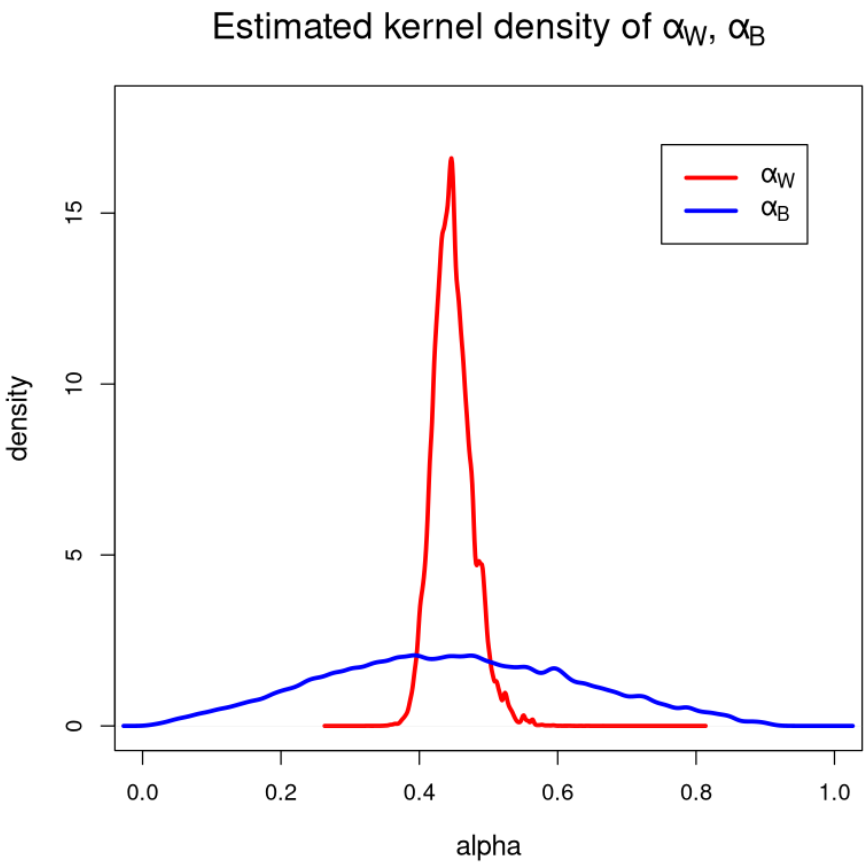
4.2 Model selection

To test the statistical significance of all the additional parameters, a parsimonious model was formulated that excludes α_W, α_B and all additional parameters, instead only taking R_W, R_W.

A likelihood ratio was computed followed by a chi-squared test, yielding a p-value of < 0.000001. This is not unexpected, since those parameters describe draw behavior, whereas ratings and the Elo system more generally, as discussed below, do little to describe draw probabilities.

4.3 Distributions of α_W, α_B

Below are kernel density estimates of the densities of α_W, α_B.



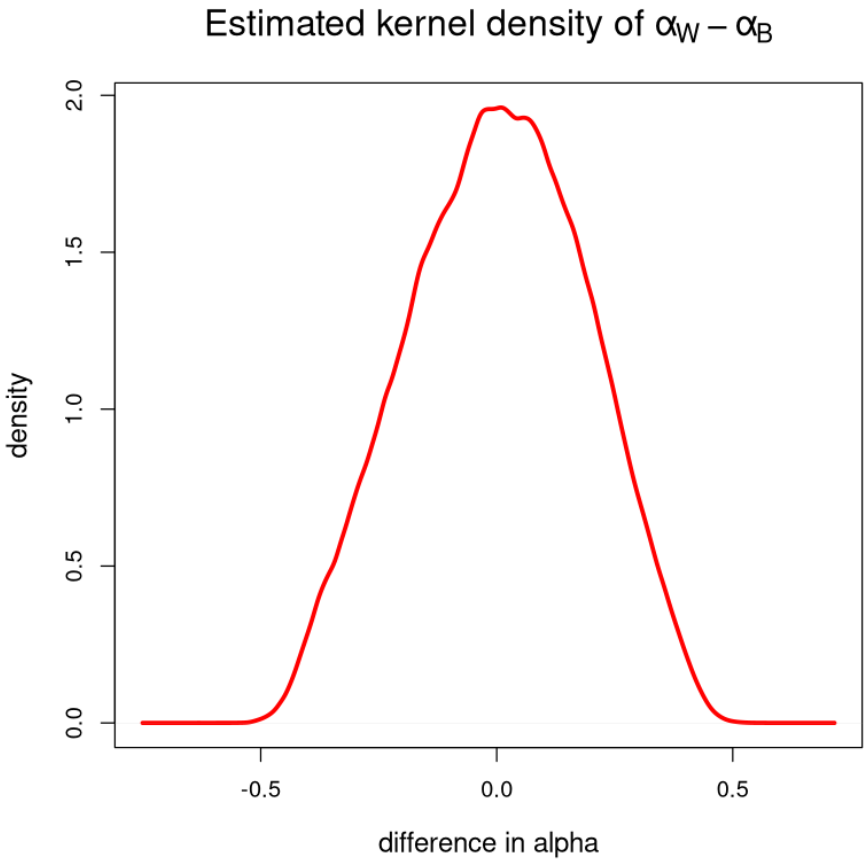
Interestingly, the density at $\alpha_W = \alpha_B = 1$ is almost 0. This might suggest that no players are playing perfectly optimally, but more likely than not, this indicates one of the flaws of the Elo rating system.

It is clearly useful (and very effective) for ranking players, but does poorly when applied to draw prediction. This is not relevant for its main purpose, but poses a challenge here. Because the Elo expected value

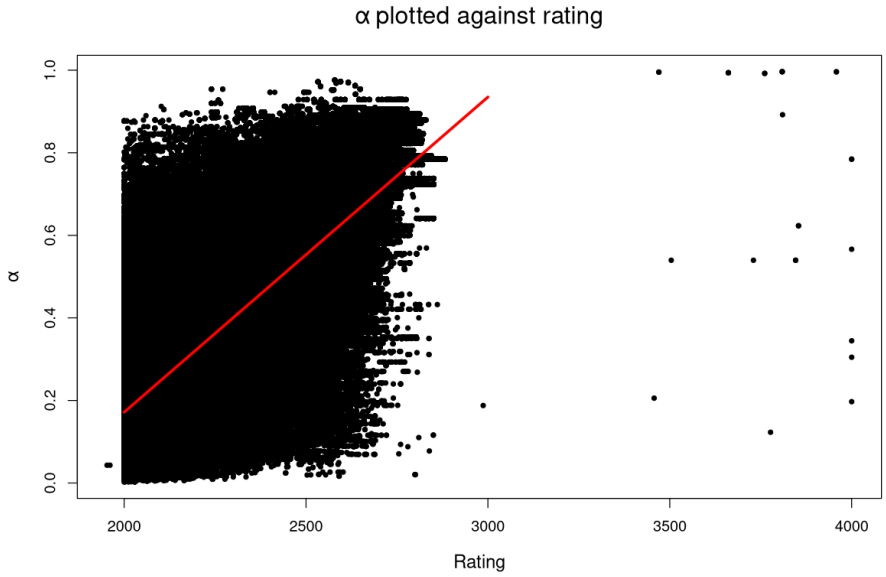
$$E_W = p_{\text{win}} + \frac{1}{2} p_{\text{draw}}$$

neither defines p_{win} nor p_{draw}, instead assigning a "payoff" of $\frac{1}{2}$ to drawing, it becomes difficult to predict draw behavior using it. This is especially visible at high levels of play where draws are quite common; 35% of games in the cleaned dataset were draws. This flaw was also present in the modified Elo expected value, which was more to derive relative playing strength in the form of adjusted ratings, rather than directly compute draw probabilities.

So rather than looking at the individual distributions of α_W, α_B, it is more informative to examine the distribution of their difference.



This is almost zero-mean, with the average difference being 0.0015 and the median difference being 0.0055. It should be expected that white plays less passively, since white wins 37% of the time as compared to black's 28% (in this dataset). Interestingly the tails of the α_B distribution are quite long in comparison to the short tails of the α_W distribution. This could again be due to black's native disadvantage encouraging more deviation and more strategies chosen to either really attempt to win (higher α_B) or attempt to draw at at all costs (lower α_W).



There is also a correlation between rating and α, with an R-value of 0.5047, indicating that at higher ratings there is less suboptimal play. This is not totally unsurprising, since at the highest echelons of play this suboptimal playing behavior is much more easily observed, socially discouraged, and more heavily punished by other players (due to the higher levels of play). At slightly lower levels of play, from national masters to lower-rated international masters, though, we observe lower values of α.

5 Conclusion

These results relate back to Roeder's article/the real-world situation in a couple of ways.

First, they shows that variance-lowering behavior at the cost of expected value does occur in chess. That is, playing suboptimally for draws at higher levels of play occurs often. It is difficult to identify, however, if this is as a result of random noise or variance in playing strength on behalf of players, or if each choice of α_W, α_B for each game is intentional. It is also clear that this behavior, on average, decreases as the level of play increases, as predicted.

That being said, the majority of situations involve players playing similarly, that is choosing $\alpha_W \approx \alpha_B$, so although some suboptimal hyperpassive play occurs, it does not occur sufficiently to pose a threat to the integrity of the game.

Roeder mentions a few measures designed to decrease drawing behavior, but these again are for the highest echelons (super GM levels) of play. These measures increase errors and decrease the likelihood of draws, whether by shortening

time controls, forcing tiebreakers to be played before matches (to force playing for wins), and so on. These would likely be successful at the highest levels of play, where games are sufficiently scrutinized and play is so precise that suboptimal play essentially cannot exist.

However, at slightly lower levels of play (still professional, but just among weaker players), the problem is not absolute precision. Rather, (assuming the results support this), the problem is this variance-reducing behavior. So from a design perspective, the solution would be to penalize this behavior. Making draws worth $\frac{1}{3}$ of a point could discourage taking draws in tournaments, for instance. Removing agreed draws (where players simply agree to a draw rather than playing it out) as a whole could also make this behavior more difficult.

References

[1] Caissabase. Unknown author.

[2] Elo, Arpad (1986). *The Rating of Chessplayers, Past and Present.*

[3] Polgar, Susan (2011). The 700 lbs gorilla issue: To draw or not to draw. *Chess Daily News.*

[4] Roeder, Oliver (2021). Some Humble Suggestions to Save Chess From Itself. *538.*

[5] SeatGeek (2022). TheFuzz, Python package.

[6] Smerdon, David (2021). *The Fighting Chess Index*.