Do professional chess players draw too often?

Tudor Muntianu

February 18, 2022

Disclaimer

I spent most of my time trying to get the models working. It was a complete nightmare. Getting the factor analysis model to work (replacing Smerdon's PCA) took me about 10 hours. Maybe I should have settled for PCA for this draft. Perhaps I should not have tried to relearn Pyro, a probabilistic programming language written in Python. That being said, if I had to build this model in R I likely would have ended up using RStan, which is a binding to another PPL, so the model wouldn't have really been in R either. The data cleaning and manipulation would also have taken me 10x longer since I don't really know R. Hindsight is 20/20.

The code (and this writeup) is in a rough state as a result – by the final draft it will be pretty and commented and clear (despite being in Python). Similarly, since the model was giving me issues I didn't have compute time to train the model on the whole dataset (which is 4.6M games). Again this will not be the case by the final draft. Everything works though; I have a complete pipeline that just needs some clean up, some hyperparameters to be adjusted/played with, and to be run to completion.

Also, there are a couple of things I would like to add (or at least investigate) by the final draft:

- Dataset of online games (ostensibly less incentive to draw here, more "naturalistic" or aggressive gameplay)
- Using the factor model to compute an index, comparing to Smerdon's results

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 The game

In this paper we make a number of assumptions about each chess game observed. First, we assume that each player plays at the strength of their respective Elo rating, or in other words, plays their best. We make no attempt to model player strength directly or measure the variance in their performance. Instead, we want to know how aggressive or how passive they are playing.

Put more formally, 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 functions as follows.

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$
$$\text{where } E_A = p_{\text{win}} + \frac{1}{2} p_{\text{draw}}$$

It follows from this that $E_A + E_B = 1$, unsurprisingly,

2.2 Player strategies and equilibrium

The Elo rating system makes an implicit assumption that draw frequency does not directly affect the expected score E. There are infinite solutions to $E = p_{\text{win}} + \frac{1}{2} p_{\text{draw}}$. So, to simplify chess for our purposes, we assume that each player plays according to their rating, and the choice they must make is how aggressive or passive to play.

If we assume, for instance, that both players W and B, are playing at maximal aggression, then $E_W = p_{\text{win}}$ and $E_B = p_{\text{lose}}$. That is, $p_{\text{draw}} = 0$. But if both players are not maximally aggressive, the probability of a draw increases. To write out all the strategies, we could define an index a_W, a_B for each player that indicates their aggression, and then rebalance $p = (p_{\text{win}}, p_{\text{draw}}, p_{\text{lose}})$ accordingly, but enumerating the strategies formally turns out not to matter. This is because, under Elo assumptions, E_W and E_B are constant, regardless of draw probability. So while players can change the probability of drawing depending on their strategies, they cannot change their expected score.

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.

Our models will estimate these probabilities p in order to determine whether these draw probabilities are in accordance with Elo's system, that is, reflective of player ability and playstyle, or if they result from "irrational play."

The prediction, as described by Polgar here, is that real-world circumstances encourage drawing at the expense of good chess. As a result of chess being played sub-optimally, 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.

3 The models

The dataset used was Caissabase, a free database of over 5M professional chess games. Simul games, Chess960 games, and consultation (team) games were removed. Games played by players with less than 10 games were also removed.

(I will make better citations and give more background and explanations of these models by the final draft I promise. I am rushing at the moment.)

3.1 Player model

For modelling player behavior, a factor model was used, inspired by David Smerdon's Fighting Chess Index. Variables computed for each player 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

See Section 12.2.4 in Bishop's *Pattern Recognition and Machine Learning* for a detailed description of this factor analysis model.

Smerdon uses PCA to form his Fighting Chess Index, but this method has its flaws. As described in Bishop, PCA can be formulated as a maximum likelihood solution of a latent variable model. But PCA does not treat the observed variables as conditionally independent given the latent variables. So if we use principal components to measure "aggression" or "fighting spirit," we are not treating average rating (a measure of player strength) and irrational draw behavior as independent. This is problem since they are correlated. More serious (stronger) players have more on the line, and so will generally play for draws more often. But as we want to measure "fighting spirit" for individual player independent of their strength, we do not want this dependence.

In Smerdon's defense, PCA's components do lend themselves to the construction of his index (whereas the factor model used here does not), but this does not change the problem of the underlying assumption.

The factor model used here does treat observations as conditionally independent of the latent variables. This is tremendously useful, since we want to retain measures of player strength while independently encoding draw behavior, especially as inputs to the following model.

3.2 Game model

(Again I will provide more detail wrt inference, etc in final, in a rush to the deadline as a result of repeated disasters)

For each game, we modelled $p = (p_{\text{win}}, p_{\text{draw}}, p_{\text{lose}})$ using the following Bayesian generative model, where l_w, l_b are the latent variables computed from the player model for the corresponding players, and r is the result of the game.

$$\begin{aligned}
\sigma &\sim \Gamma(\alpha = 1, \beta = 2) \\
\epsilon &\sim \mathcal{N}(\mathbf{0}, \sigma\mathbf{I}) \\
p &= Al_w + Bl_b + \epsilon \\
p_{\text{pos}} &= \text{Softplus}(p) = \log(1 + \exp(p)) \\
p_{\text{norm}} &= \frac{p_{\text{pos}}}{\sum_{i=1}^{3} p_{\text{pos}_i}} \\
r &\sim \text{Cat}(3, p)
\end{aligned}$$

The result r is drawn from a categorical distribution parameterized by p_{norm}, which in turn is a normalized linear transformation of the latent variables l_w, l_b representing the players. The parameters are optimized via stochastic variational inference using a family of multivariate Gaussians as the variational distribution. I will also try MAP estimation and see the effect on the parameters.

Once matrices A and B are found, we can compute p_{norm} for each game from the latent variables. This can then be transformed into $\hat{E}_W$ and $\hat{E}_B$. Similarly, we can compute E_W, E_B from player Elo ratings, as described previously.

4 Results

As of now I did not have enough time to train the model on the full dataset (or even a significant subset).

However, to interpret/compare the results, here a few ideas (a couple of which are quickly explored in the notebooks, but mostly unimplemented):

- Compare $\sum_{i=1}^{ngames} E_{W_i}$, $\sum_{i=1}^{ngames} \hat{E}_{W_i}$, and $\sum_{i=1}^{ngames} r_i$ where r_j is the result of game j, and $r = 1$ if W wins and $r = 0.5$ if it was a draw. This is not a very informative measure, but should express whether the expected scores over- or under-estimate the actual scores.
- Rank players by average draw probability to form an analogous index, compare with Smerdon's index
- Check to see if $\hat{E}_{W_i}$ predicts upsets(victories against players with significantly higher ratings), and if it does so correctly, investigate why
- Visualize variance in latent variables as well as in the generated features
- Compare/regress draw probabilities across different ratings and/or different differences in rating

- Split dataset into training and testing, see if it correctly predicts draws on the test dataset

These (future) results can relate back to the article/real-world situation in a couple of ways. First, they will show whether or not high draw rates are as a result of sub-optimal play rather than higher accuracy at high levels of chess. This would indicate that other pressures induce players into actually playing worse chess to avoid variance. This would provide a new angle from which to attack the problem of increasing draws.

The article mentions measures designed to increase errors and decrease the likelihood of draws. These would likely be successful at the highest echelons of play, where games are sufficiently scrutinized and play is so precise that suboptimal play essentially does not exist. However, at 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.