What is the Fastest Way to Correct One's Elo Rating?

Alex Richter

1 Introduction

Suppose Player C has an Elo rating of r, but based on their ability, they should be rated $r + a$. In theory, any opponent that they play will give them a positive expected rating gain, but which rating of opponent will yield the highest expected gain? Assuming the reader already knows what an Elo rating is, they may skip the rest of this introduction and skip to the section labeled "Results: US Chess system". This introduction is to explain what an Elo rating is.

An Elo rating is an attempt to encapsulate a player's level of skill at a game within a single number. Elo ratings are meant to be relative; higher-rated players are generally more skilled. In theory, any 1-on-1 game can feasibly use Elo ratings, but this paper will stick to chess as its game of choice, so as to avoid using the clumsy term "a game" over and over again.

Suppose Player A and Player B, who are both rated, play a game of chess, which ends in a draw. Let's leave aside the question of how their ratings are initialized, which is beyond the scope of the present writing. How are their ratings updated by the outcome of this game? First, each players' expected score is computed, where 0 represents a loss, 0.5 a draw, and 1 a win. Player A and B's expected scores are both a number between 0 and 1, and must add up to 1, since whether someone wins or there's a draw, one point is divided amongst the contestants.

The method by which these expected scores are computed can be anything potentially; there is more than one Elo rating system even in chess alone. Let Players A and B have ratings r_A and r_B respectively. In US Chess, player A's expected score is $\frac{1}{1+10^{-\frac{r_A-r_B}{400}}}$ and player B's expected score is $\frac{1}{1+10^{-\frac{r_B-r_A}{400}}}$ Note that the sum of the two player's expected scores is

$$\frac{1+10^{-\frac{r_A-r_B}{400}}+1+10^{-\frac{r_B-r_A}{400}}}{(1+10^{-\frac{r_A-r_B}{400}})(1+10^{-\frac{r_B-r_A}{400}})}=1$$

Those with experience in statistics may recognize this formula as the cdf of a logistic distribution with location parameter 0 and scale parameter $\frac{400}{\ln 10}$. Then another way of looking at this Elo rating scheme is as a statistical model so that if X is a random variable that follows the distribution just specified, and Player A and B's ratings are r_A and r_B respectively, Player A's expected score is $P(X \leq (r_A - r_B))$ and Player B's expected score is $P(X \leq (r_B - r_A))$. Note again that the two players' expected scores add up to 1 because the logistic distribution is symmetric when its location parameter is 0. International ratings (more commonly known as FIDE ratings) use a similar concept to US Chess's logistic distribution, but international ratings use a normal distribution with mean 0 and standard deviation $200\sqrt{2}$

Without statistical terminology, we would say that player A's expected score is $\int_{-\infty}^{r_A - r_B} \frac{1}{400\sqrt{\pi}} e^{-\frac{x^2}{80,000}} dx$ and for player B, it would be the same except the upper bound is $r_B - r_A$.

Whatever one's approach to computing each players' expected score is, you then take their score (0.5 for both players in this case) minus their expected score and multiply that by a parameter of the rating system called the K factor. If player A's expected score was 0.6 and player B's expected score was 0.4, and both players have a K factor of 20, player A's rating change is $20*(0.5-0.6) = -2$ and player B's rating change is $20*(0.5-0.4) = 2$. A higher K factor leads to a more volatile rating, as the changes in one's rating is directly proportional to one's K factor. Overperforming expectations is more highly rewarded, and underperforming them is more harshly punished. What value, or values, the K factor can have is another feature that separates different Elo systems from each other. For instance, my K factor in US Chess is 15.2, but in FIDE's system it's 20. Two players need not have the same K factor, so rating changes are not necessarily a zero-sum game; rating points can be added to or taken away from the system. If player B had a K of 30 in the example we just used, then player B would have gained 3 rating points, which is more than player A lost.

2 Results: US Chess system

Let's suppose Player C is rated r. Their expected scores will use this value of r, but perhaps Player C's skill level is such that these expected scores are not accurate to what they will actually achieve on average. In particular, let's suppose that an accurate representation of their actual expected scores are those calculated as if their rating were $r + 100$. This is a somewhat longwinded way of saying that Player C is ''underrated" by 100 points. Now, let's find which opponent's rating will give the highest average rating increase. This can be found by maximizing the difference between player C's expected points under their ''true strength" $r + 100$ and their official, nominal rating r, as a function of their opponent's rating. Let the opponent's rating be x. Then this difference, under the US Chess sytem, is:

$$
\begin{aligned}
f(x) &= \frac{1}{1+10^{-\frac{r+100-x}{400}}} - \frac{1}{1+10^{-\frac{r-x}{400}}} \\
&= \frac{(1-10^{-\frac{1}{4}})10^{\frac{x-r}{400}}}{1+(1+10^{-\frac{1}{4}})10^{\frac{x-r}{400}} + 10^{-\frac{1}{4}}10^{\frac{x-r}{200}}}
\end{aligned}
$$

Differentiating f(x) gives

$$
\begin{aligned}
&k * 10^{\frac{x-r}{400}} * \frac{1+(1+10^{-\frac{1}{4}})10^{\frac{x-r}{400}} + 10^{-\frac{1}{4}}10^{\frac{x-r}{200}} - ((1+10^{-\frac{1}{4}})10^{\frac{x-r}{400}} + 2*10^{-\frac{1}{4}}10^{\frac{x-r}{200}})}{(1+(1+10^{-\frac{1}{4}})10^{\frac{x-r}{400}} + 10^{-\frac{1}{4}}10^{\frac{x-r}{200}})^2} \\
&= k * 10^{\frac{x-r}{400}} * \frac{1-10^{-\frac{1}{4}}10^{\frac{x-r}{200}}}{(1+(1+10^{-\frac{1}{4}})10^{\frac{x-r}{400}} + 10^{-\frac{1}{4}}10^{\frac{x-r}{200}})^2}
\end{aligned}
$$

where $k = \frac{\ln 10}{400}(1-10^{-\frac{1}{4}}) > 0$. This quantity can only be zero when the numerator is 0 i.e. when

$$1 - 10^{-\frac{1}{4}}10^{\frac{x-r}{200}} = 0$$

Which happens when $x = r + 50$. Finally, we take a second derivative to show that this in fact is a maximum. We use our knowledge that the numerator of $f'(x)$ is 0 when $x = r + 50$ to greatly reduce the work involved in doing this; of the three terms in the product rule sum (the $10^{\frac{x-r}{400}}$ term, the numerator, and the denominator), only the one that differentiates the numerator will be non-zero at $x = r + 50$.

$$f''(x)|_{x=r+50} = k * 10^{\frac{x-r}{400}} * \frac{-\frac{\ln 10}{200}10^{-\frac{1}{4}}10^{\frac{x-r}{200}}}{(1+(1+10^{-\frac{1}{4}})10^{\frac{x-r}{400}} + 10^{-\frac{1}{4}}10^{\frac{x-r}{200}})^2} < 0$$

Thus we see that if player C with rating r is ‘’underrated” by 100 points, then their expected rating gain is highest against opponents rated r+50. This calculation can be easily generalized to show that if player C’s ‘’true strength” is $r + a$ where $a > 0$, then they will on average gain the most rating points from opponents rated $r + \frac{a}{2}$. The only change is that $10^{-\frac{1}{4}}$ is replaced with $10^{-\frac{a}{400}}$.

3 Results: FIDE system

In the previous section, we answered the question of which opponents lead to the fastest rating gain under US Chess’s system in case that player C is underrated. This question of which opponents are ‘optimal’ can also be easily answered under FIDE rules, provided one has an understanding of the normal distribution. To remind the reader of how FIDE’s statistical model works, let players A and B have ratings r_A and r_B respectively, and let X be a random variable that follows a normal distribution with mean 0 and standard deviation $200\sqrt{2}$. Then player A's expected score under FIDE's system is $P(X < (r_A - r_B))$ and player B's expected score is $P(X < (r_B - r_A))$. Now let player C, rated r, be underrated by a rating points. Then with an opponent rated x, the difference between player C's expected score with their nominal rating of r and their expected score with their 'true' rating is $f(x) = P((r - x) < X < (r + a - x))$. Since the normal distribution is densest around the mean, and X has a mean of 0, it should make intuitive sense that the value of x which maximizes player C's expected rating gain is $r + \frac{a}{2}$, as that creates an interval centered on 0. This is identical to the result from the previous section, which used a similar statistical model.

4 Conclusion

We have seen that under both US Chess's and FIDE's Elo rating systems, playing slightly higher rated opponents leads, on average, to the fastest rating gain. There exist Elo systems besides US Chess's and FIDE's systems; in particular Mark Glickman's ''Glicko" system is a popular alternative in online chess as well as in games besides chess that use Elo ratings. However, both Glicko and its successor Glicko2 compute expected scores using a logistic distribution, just as US Chess does; Glickman's main innovation is a more involved method of determining K factors. Because both of Glickman's systems use a similar statistical model that US Chess does, I would suspect that it is still optimal to play slightly higher rated opponents under the Glicko or Glicko2 systems.

The motivation to write on this topic came when I was struggling to get my online rating to a level that I thought represented my real strength. Even if I was correct in believing that I was ''underrated", how many games would I be expected to play to regain my ''true" rating? How long could it potentially take? To this end, I was planning to determine some of the properties of the random variable, call it Y, which represents the number of games it takes a player rated r to get to a rating of $r + a$ assuming that $r + a$ is their ''true" rating. I would sample Y via a simulation study where many games would be played against an opponent rated x.

This all seems simple enough, but there is a complicating factor: how to model the draw rate? If a player is expected to score 0.3 points in a game, that could mean a win rate of 30 percent with a draw rate of 0 percent, or a win rate of 0 percent and a draw rate of 60 percent, or any of the infinite possibilities in between. Also, in practice, one's draw rate is different against players of different ratings; I draw most often against players close to my rating, and draw less often against players far higher or far lower rated. One possible approach to this issue, and one that I may well take for a future project, is to assume that draws never occur. After all, in online speed draws, draws are quite rare anyway. My draw rate on lichess.com is about 5 percent, and on chess.com it's about 6.7 percent.