URS™ Universal Rating System

The Universal Rating System – A Performance Rating Across All Time Controls

The URS™ rating algorithm was designed and developed by our research team, which consists of Mr. Maxime Rischard, Dr. J. Isaac Miller, Dr. Mark Glickman, and Mr. Jeff Sonas. The work has been funded for the last two years through a collaborative research project funded by the Grand Chess Tour, the Kasparov Chess Foundation, and the Chess Club and Scholastic Center of Saint Louis.

There are many differences between the URS™ and the FIDE Elo systems. The most striking difference is that the URS™ calculates only one rating for each player, informed by their results at all rates of play from Classical to Blitz (5 minutes per game). This published rating is the URS™ system's assessment of each player's strength at Classical chess (defined as a rate of play where each player has at least 2 hours for their first 60 moves).

Furthermore, the URS™ is a weighted performance rating, calculated across several years of previous game results for all players. Older games are given less importance than recent games, by applying an exponential decay rate. URS™ Ratings are calibrated so that they use a scale comparable to traditional Elo ratings. It is critical to note, however, that the URS™ does not incorporate Elo ratings anywhere within its actual calculation.

Comparison to Elo-based approach

A URS™ Rating is more like a performance rating than it is like an Elo rating.

In Elo systems, each player retains an Elo rating that is incrementally adjusted based on the results of any new games played. It is essential to know what the Elo ratings of the two players were at the time that a game between them was played, since each player's Elo rating is used to calculate their expected score. This in turn directly impacts how their Elo rating is increased or decreased due to the actual result of each game.

In comparison, the URS™ involves computing ratings simultaneously over a substantial period. The only game information included in the URS™ calculation is who had White and Black, what the outcome was (win/loss/draw), when the game was played, and what the time controls were. It doesn't matter what each player's rating was in the past, because there is no concept of a rating that is incrementally adjusted up or down.

Instead of adjusting an existing rating, the URS™ simply includes the new games within the large pool of existing data that it analyses whenever it is time to calculate a rating list. The ratings of every single player in the pool are then recalculated, in what is essentially a complex performance rating calculation, using their entire pool of games within the database. In Statistics terminology, the URS™ involves a time-weighted regularized maximum likelihood calculation, an approach that has solid statistical foundations.

What is a Simultaneous Performance Rating?

Anyone who understands how performance ratings are calculated may now be wondering how it is possible to calculate a performance rating across a large pool of games while ignoring the opponents' pre-event ratings at the time of each game? The solution is to treat all the games as if they were played in one giant tournament and determine the ratings that are simultaneously most consistent with the game outcomes. This cannot be accomplished using a simple formula like the Elo updating formula. Instead, the computation needs to be iterative. This type of iterative procedure is well-established in applied mathematics. In fact, Arpad Elo suggested a particular instance of an iterative procedure decades ago when he developed his rating system. Under this scenario, we assume initially that everyone has the same rating (we'll call it "R") and we then calculate a Tournament Performance Rating (TPR) for every single player across this hypothetical tournament. Then once you have those TPR's for everyone, you recalculate a TPR for everyone, but this time, instead of using "R" as the rating for each opponent, you actually use each opponent's latest TPR. And then you keep doing this, over and over.

In this way, each time you have an iteration of calculating a performance rating, you obtain a more self-consistent set of performance ratings, which in turn makes the next iteration of TPR even more self-consistent. If you do this for long enough, you will typically reach a stable equilibrium where the TPR's are changing by only negligible amounts from one iteration to the next. When you reach this point, each player's performance rating is consistent with the performance ratings of all their opponents. This can be called a "simultaneous performance rating".

Performance Ratings Explained

This brings us to a major point: what exactly do we mean when we say the URS™ is like a "performance rating"? In fact, there are lots of ways of calculating performance ratings. Some of the ways are simple, and some are far more complex.

The simplest, most popular, and most easily understood method, is to first calculate the average rating of your opponents across the different games played in a tournament. You then convert your overall percentage score in those games into a rating advantage/disadvantage, and add that positive/negative number to the average rating of the opponents faced, in order to calculate the performance rating.

There are different variations of this calculation (e.g. removing the game against the weakest opponent and so forth) but they typically involve a relatively simple formula that allows you to calculate the performance rating directly. This is essentially how Jeff Sonas's Chessmetrics simultaneous performance rating calculation worked, years ago. US Chess has a similar but more complex way of computing performance ratings that are the basis for determining provisional ratings. However, the URS™ took a more thorough approach.

There is another way of viewing performance rating, one that is not as easy to calculate using a direct formula. Under this approach, a performance rating is "the rating that would have led to the performed results". We can then consider several possible ratings for our player, assess what their overall expected score would be (with that rating) across all their games, and then pick the rating that yields an expected result that most closely matches what actually happened.

Rather than inventing a specific formula that can be used to calculate ratings directly, like there is for the Elo system and for performance ratings, the URS™ involves a probability model that analyses a large domain of possible ratings for each player, with some ratings being more likely than others (based upon the overall population distribution of chess strength). Across those possible ratings, our system then determines how likely the actual results would have been to occur, and ultimately determines the most likely overall set of ratings, for all players at once, in order to best explain the actual results.

Time weighting adjustments

In a traditional TPR, it makes sense to treat each game with equal importance, since all games are played at nearly the same time. When we extend the concept of a TPR to cover a much longer timeframe, as we have done with the URS™, then of course some of the games were played recently and others were played years ago. It is therefore logical to give the older games less weight than the newer games.

In the case of the URS™ we assign reduced importance to older games through the use of exponential-decay game weightings. The actual decay rate is one of the URS™ system's parameters and affects how sharply or gradually the importance of older games is reduced. We currently calculate ratings across a six-year history of game results.

Rate of play adjustments

Finally, accounting for different time controls was incorporated into the URS™ in a more universal way than just classifying all time controls as either Blitz, Rapid, or Standard. For each event, we determined the maximum number of minutes each player could spend for their first 60 moves. We call this value "M60", where the "M" can be viewed as being an abbreviation for "Minutes".

For some time controls it is easy to calculate M60. For example, for "Game in 5 minutes" or "Game in 90 minutes" the corresponding M60 values are 90 minutes and 5 minutes, respectively. This represents the maximum number of minutes each player could take for their first 60 moves (indeed for their whole game).

However, many time controls have delays or increments, where players receive additional thinking time either during, or after completing, each of their moves. In these cases, we assume the maximum time is taken, whether for increment or delay. Since we are looking for the maximum time that can be taken for 60 moves, it is convenient that increments and delays are typically expressed as N seconds per move, since this also means that it takes N minutes for the first 60 moves. Consequently, if we see something like "Game/5 min + 2 sec/move", we can just add 5+2=7, which means that our M60 value for this time control would be 7 minutes.

And finally, where there are time increments linked to a specific number of moves, then these are also counted (provided the increments start before move 60). Consequently, a time control like "40 Moves/90 min + Game/30 min + 30 sec/move" would have an M60 value of 150 minutes. This is calculated as 90 minutes for the first 40 moves, another 30 minutes (maximum) for the next 20 moves, and 30 minutes' worth of increments through to Move 60, so our value of M60 is 90+30+30=150 minutes. However, if the time control were "40 Moves/100 min + 20 Moves/50 min + Game/15 min + 30 sec/move", then the part about "Game/15 min" would not matter for the calculation of M60, since it doesn't apply until Move 60 has already been completed.

Thus you get the same value of M60=5 minutes for "Game/5 min" and for "Game/3 min + 2 sec/move", and our rating system therefore treats these time controls equivalently. Similarly, the two common time controls "Game/90 min + 30 sec/move" and "Game/120 min" are treated equivalently as M60=120 minutes.

The URS™ uses the M60 values within continuous functions that model the variability of chess results across all rates of play from 5 minutes (blitz) up to 120+ minutes (classical). They are also used to calculate the degree to which individual players' quality and consistency of play degrades as the rate of play moves along the spectrum from classical to faster rates. Player-specific degrees of degradation in quality and consistency of play are expressed as their Rapid Gap (applying to M60=30 minutes) and their Blitz Gap (applying to M60=5 minutes) for each player. A larger Rapid/Blitz Gap means that the player's quality and consistency of player degrades faster as they play at fast rates of play.

While optimizing the rating system, we determined the appropriate continuous functions to use for modelling the variability of results at any value of M60. Thus we treat "Game/41 min" slightly differently from "Game/42 min" and "Game/43 min", but there are no special considerations for these particular rates of play, just as there is nothing special about the treatment of "Game/9 min" versus "Game/10 min" versus "Game/11 min" in our system. They are all treated smoothly across the full spectrum of time controls. By contrast, "Game/9 min" versus "Game/10 min" versus "Game/11 min" are handled completely differently in the Elo system, as some of these games go into the Rapid rating system while others go into the Blitz rating system.

From the perspective of the URS™, the only special points on the spectrum of time controls are at M60=5 minutes and at M60=120 minutes. The URS™ currently does not rate any lightning/bullet results played at faster than "Game/5 min". And all games played at time controls of game in 120 minutes or more (i.e. all classical games) are treated as M60=120. Any slower time controls of play for games that can take longer than 2 hours for each player are thus treated equivalently to "Game/120 min" in the rating calculation.

On the surface, it might seem like a bad idea to mix together a large number of rapid and blitz results with a relatively small number of classical results, when the ultimate goal is to calculate a rating that accurately measures classical chess skill. Indeed, the greater unpredictability of faster chess does mean there is less information to be learned from one rapid or blitz result than from one classical result.

Nevertheless, the URS™ recognizes that there is useful information about a player's over-the-board strength in all game results regardless of the time limit, and can therefore more effectively estimate a player's classical chess strength by also considering their results in games played at faster time controls. As the speed of play increases, the URS™ assigns less and less importance to the game results relative to games played at slower time controls. In this way, we gain useful information about players' classical chess skill without being overwhelmed by the volatility or volume of rapid and blitz games.

We will be providing more details about the rating system in due course, but hopefully this provides an interesting and informative introduction for now.

We welcome all constructive comments that can help us improve moving forward.