CSE702 Fall 2025 Week 3 Monday: Elo and "Objective" Metrics

The meeting began with a question about why the full predictive analytic model (to come next week) is loglog-linear rather than log-linear. This is put in more detail (than the introductory slides) on slide 7 of my talk at RIT last November. All forms of my model currently have two main free parameters, called "s" for "sensitivity" and "c" for "consistency" (a third parameter called "h" or "e_v" intended to reflect depth of thinking is tied to the other two---plus there are hyperparameters with fixed values). They can all be fitted on a sample of games G by solving the two equations in two unknowns that equate the projected T1-match and ASD on G with the actual values (thus making them into **unbiased estimators**). This is in preference to maximum-likelihood estimation (MLE) and various other minimizable loss functions that can be configured in my C++ code. The question then becomes: how do numerous other estimators behave under the resulting fit? Those estimators include the frequencies F2 and F3 of playing the engine's second or third listed move, or "T3-match" which is playing one of the top three moves, or the frequencies of small, medium, or large errors. The upshot shown by my "London Calling" article is

- when the log-linear model is fitted, F2 and F3 are horribly skewed away from their projections.
- when the loglog-linear model is fitted, F2 and F3 and the other mentioned estimators not only match their projections well, their errors are close to Gaussian and within their variances as projected by the model.

The Gaussian aspect was explored in the seminar last year, but no one chose to make a fully rigorous study out of it. The degree to which loglog-linear works makes me speculate that some kind of "natural law" is involved, tying in a larger reason why the probabilities p_i of playing the i-th best move are most economically represented as powers not multiples of p_1.

The first large segment of the meeting reviewed the investigation from last Tuesday. I updated the "Titled Tuesday" data to include last week's play and firmed up the choices of data to display. This is summarized in 20 pages of charts which I went through. It includes a link to the annotated spreadsheet of data used. The **main conclusions** are:

- Empirical support for the theoretical stance that the T1 and ASD metrics should be linear in the real-world skill metric (Elo rating).
- The FIDE Standard rating system has recovered sufficiently well from the pandemic and the (March 1, 2024 one-time correction to the) issues identified by FIDE statistician Jeff Sonas.

But these are only clearly favored by the January 2025 data set alone, and are subject to review as February 2025 data starts coming in. There is also evidence that 1400 cannot be a hard floor of the FIDE rating system, since it corresponds to under 0.20 in ASD under the newest fits, but a substantial number of FIDE-rated recently-active players have higher ASD values.

In the third segment, I used the same spreadsheet and transposed the axes to investigate further questions.

Raw Metrics and Their Analytical Horizons

I call simple-counting metrics "objective" because they do not involve comparisons to projections made by a model. For example, in baseball or cricket:

- The number of runs a player P scores (or bats in) is objective.
- The number of runs above what a "replacement-level" player R would provide is less objective. Even if R is defined by the statistics of the least regular player in a given year (instead of a model projection of that playing level), that definition pertains to that year.

Neither of these measures takes into account the relative difficulty of the situations faced by P. If P plays home matches on a bone-dry easy wicket, or in a small warm baseball park with wind blowing out, then P's statistics are likely to be inflated. If P plays in a league or division overfull with top-class bowlers/pitchers, then P will be deflated. Similar considerations apply in chess to the difficulty of positions a player faced. Did the opponents make things easy or tough for P? Normalizing for such factors requires a predictive model, either in the foreground for the particular games or in the background for setting general thresholds. This becomes more "subjective."

Before we get to predictive modeling, we will develop the objective metrics some more. We have satisfied ourselves (maybe) that T1 and ASD

- theoretically should be linear in the real-world skill metric (Elo rating), and
- are linear in Elo ratings as they stand now.

The next question is one that has attracted headline attention in the past:

What is the rating of perfect play---

- that corresponds to "perfection" in these metrics?
- overall?

The end result will be a much more mundane question: **What is the limit of resolution of your measurement tools? What horizon do they give?**

First, what does "perfect play" even mean? In cricket one cannot expect to hit a 6 or take a wicket with every ball. The NFL's Passer Rating formula tops out at the weird number 158.3, which can be achieved even with some incomplete passes (no interceptions allowed, though). In chess, as I will go into more in the next meeting, it is IMHO better to start with a notion of "top class." In game theory, a [**deterministic**] **strategy** is a function $s: P \to M$ where P is the set of all legal (reachable) game positions and M is the set of legal moves (in those positions, respectively). A **randomized** strategy is a probability distribution of such functions, or equivalently, a mapping from positions π in P to a probability distribution $D(\pi)$ over the legal moves in π.

Definition: A possibly-randomized strategy s is **top-class** if for every opposing strategy s', the points expectation for s versus s' is at least 0.25.

It is possible---I'd say even plausible---that the current best computer chess programs are top-class, running on commodity hardware at the old TCEC time control of 120 minutes + 30 seconds increment per move made, *provided they are randomized a little*. Without randomization, a strategy s' that knows about s in advance could identify and play toward situations where s errs. (In the definition, it does not matter whether s' is randomized. The current TCEC rules specify Rapid chess, 30 minutes + only 3 seconds increment. CCRL allows 15 minutes for every chunk of 40 moves with a single Intel i7-4770k chip as the hardware standard.) If we therefore say that "Stockfish 17 can score at least 25% against God" then "God" is not just a single player but even stronger: a universal quantifier over all possible players.

If that is true, then the Elo rating formula dictates an immediate consequence: **No player can be rated more than 200 points higher than Stockfish 17, hence no higher than the 3800s**. This gives a reason to think of 3500-to-4000 as an envelope for "the rating of perfect play"---call it the "reasonable range."

Our two featured raw metrics come with obvious notions of perfection: **100%** for T1-match and **0.00** for ASD. Zero error surely means perfection, no? How well do they stand up?

The rest was a demo of the intercepts when the axes are transposed to make $y =$ Elo.

The main observations and their upshots were:

- The 100% intercept is for Elo above **5000**. But since major engines agree with each other on the best move only about 75% of the time at the depths of search used in my analysis, the rating at that intercept---shown to be in the reasonable range---is more plausible.
- The ASD **linear** fits are all significantly below the reasonable range. The nub of my 2016 "When Data Serves Turkey" article is that whereas raw ACPL gives an intercept under 3200, ASD gave intercepts near a more-reasonable 3400 figure. But now, the linear fit of ASD for July--December 2024 was only 3193.
- Even for the January 2025 data with "approved" linear fit, only the quadratic fit gave an intercept in the reasonable range---well, barely at its bottom near 3500.
- However, the intercept for Titled Tuesday Blitz under Chess.com's rating system was over 4600. The quality level is shifted by over 600 Elo according to the curve on slide 23 of the RIT talk (can you see this full version?), and Chess.com ratings are shifted up by at least 200 over FIDE. So this may correspond to near Elo 3800 after all.

The last item still has a discrepancy between Elo 3800 and the Elo 3200-or-so given by the estimates from classical chess events under linear fits. If you play around with other data in the spreadsheet, transposing the axes, you may find other discrepancies. All of them may be swallowed up by the final brute fact: The engine analysis that populates all these graphs runs in a time-and-depth limited mode

that equates to about 3100--3200 strength. I have not measured it formally, but this mode did score over 75% in matches against a depth-limited version of the Toga II chess engine whose programmer told me was about 2900 Elo. From general experience---and by extrapolating direct measures of the predictive accuracy of my full model---the resolution is good up to 3100 Elo but then hits a horizon of noise. So:

> All estimations above 3100 from this data require supplementary indicators and reasoning. To be taken otherwise with chunks not just grains of salt.

This is, however, enough resolving power to distinguish human and computer levels in most cases.