CSE702 Seminar Week 8: Difficulties With Difficulty

We have already seen three basic issues with the "Fidelity" model's implementation of difficulty:

1. It is dependent on the skill level of the player a position is difficult "for".
2. It is defined only in terms of the projected loss of expectation in the position. (This does keep to the model's supreme agnosticism about details of chess.)
3. It mis-matches real chess difficulty in positions like the one from the Niemann-Shankland game in the Week 5A notes, where White has 14 moves rated completely equivalent in final value because Black is not yet ready to bring a crisis, but the need to find a secure defense looms now.

Now we will evaluate it based on the most desired criterion:

4. Does it accurately select the positions that players find most difficult---on which their performance is observed to suffer?

The answer is neither no nor yes, nor "mu" but a resounding mess.

Before we plow in to demos and details, here is a general matter of expectations. The Intrinsic Pereformance Rating (IPR) is supposed to be a completely extensional measure of a real human player's skill, even though defined in terms of the virtual player Y_P that best fits the given set of games by the human player P. As such, it should give uniform results R_P regardless of the difficulty of the positions π in those games. The R_P is the measured value of a signal emitted by P. The model's operation "rolls with" difficulty because it first makes projections tailored to given positions π and then compares the player's performance against those projections. In more-difficult positions it makes lower projections. In particular, it projects a higher value of centipawn loss or expectation loss in hard positions π---because that is how their difficulty is defined to begin with.

The raw metrics observed, on the other hand, can-and-should change. If the relative difficulty of a position is 2.00, meaning twice as high a projected scaled centipawn loss (ASD) as in an average position, then players P (of all ratings---if the difficulty measure has been correctly normalized across ratings) in such positions should show 2x their overall ASD.

How about the T1-match metric, i.e., the frequency of finding the move the computer recommends? There are three possibilities:

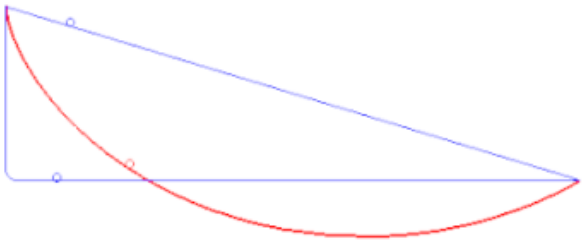
- The T1 match metric could go down, because it is more difficult to find the best moves.
- The T1 metric could be expected to stay the same, because it is an extensional measure of the player's skill like the IPR.
- The T1 metric could go up---why on earth would it do that??

"Brachistochrome Problem" in chess: What is the distribution of move values v_i (viewed as differences from the value v_1 of an optimal move) such that the projected probabilities p_i maximize

$$E[loss] \;=\; v_1 \;-\; \sum_{i=1}^{\ell} p_i v_i ?$$

- If you make the values v_i tail off too slowly, then although the probability p_1 of finding the optimal move (include p_2 etc. if there are tied-optimal moves) goes down, the collective probability of these other reasonable moves will be high enough to keep $E[loss]$ down.
- If you make every move other than the first move a big blunder, then the probabilities p_i of those moves may be driven down far enough so that (especially with the scaling-down of extreme values) the sum of $p_i\, v_i$ over $i \;>\; 1$ stays small.

So the answer is somewhere in the middle. This is analogous to Johann Bernoulli's famous Brachistochrone Problem in 1696. I have neither computed this exactly nor simulated it, but I imagine it is similar to the solution in physics---see Wikipedia's GIF lined above:



If the red curve pretty much describes the shape of dropoff in values v_i, then the first fact is that v_1 is considerably higher than v_2. This means that the projection for p_1 will be higher than in an average chess position. In turn, if we restrict to these positions---which maximize difficulty the way we have chosen to define and measure it---then we should expect the players' T1 match rates to go **up**, not down. And they certainly do...

> Thus the "organic" difficulty metric isolates positions in which the optimal move is unqiue and relatively easy to find. These are, however, exactly the kind of positions in which my model was designed to explain high concordance to computers, going all the way back to the 2006 world championship match (especially game 2 from turn 33 onward).

How about the ASD metric? Well, if the model is sensible, then "brachistochrone" positions maximizing the expected dropoff should show higher ASD values (meaning lower skill) in practice. And they do...but not by as much as expected.

What we really care about is the IPR metric. That really should stay consistent on any sample of positions that is not postconditioned on the player making bad moves in them. The difficulty metric qualifies: it is based only on projected loss, not actual loss.

Demos...

The fairly consistent and majorly dismaying upshot across demos is:

- Over positions the model judges twice as difficult as usual, meaning the projected loss is twice as high as the average over all positions, IPR figures go **way up**. (In EWN mode, these positions ("turns") are those with weight ≥ 2.)
- Over positions the model judges to be easy, say weight ≤ 0.5 in EWN mode, the IPRs go **way down**.
- The actual loss figures do go up in the weight ≥ 2 positions and down in the weight ≤ 0.5 positions, but their ratio stays within a factor of 2, not 4 or more as it should.

This is not a healthy state of affairs. Before we wring our hands over the implementation of difficulty, let's explore the issue of consistency of performance on subsets of positions in more generality.

Performance Compartmentalization

First, this is a completely different question from that of the model's projections on predicates that are not expressly fitted as unbiased estimators. In that matter, we compared the "ersatz z-scores"

$$\frac{actual \ - \ projected}{projected \ stdev}$$

for these predicates with the bell curve. These predicates: playing the 2nd, 3rd, 4th, or 5th-ranked move, making an error of limited or unlimited magnitude, are all ones that are postconditioned on the player having made some kind of mistake.

Here is right away a good example of what in human games we might think of as "difficult": a position where objectively you are substantially behind, in grave danger of losing. Say: positions where the value for the player to move is -2.00 or worse.

More Demos!

And some more examples of turn filters and how to script experiments...

Sets of moves where they're not difficult---would you expect consistency of signal?

- turns 9--16 of any game: late opening---still usually in book.
- turns 17--24. early middlegame
- turns 25--32 crunch middlegame
- turns 33--40 **Zeitnot**
- turns 41--48 late middlegame
- turns 49--56
- turns 57--64 usually endgames
- tunrs 64 and higher often reduced endgames