Intrinsic Chess Ratings

AAAI 2011

Kenneth W. Regan[1] Guy McC. Haworth[2]

University at Buffalo (SUNY) University of Reading (UK)

August 8, 2011

[1] Various projects in progress, co-workers named orally in-context. Sites: http://www.cse.buffalo.edu/∼regan/chess/fidelity/ (my homepage links), http://www.cse.buffalo.edu/∼regan/chess/ratings/ (**all data, not yet linked**).

[2] Various other projects—Google Haworth Reading CENTAUR to find http://centaur.reading.ac.uk/view/creators/90000763.html

Fallible Agent Model – Inputs and Outputs

- Player skill parameters $s, c, \ldots$
- At each decision point (game turn), 'objective'/'hindsight' *utility values* $v_0 \geq v_1 \geq v_2 \ldots$ for the available options (moves) $m_0, m_1, m_2, \ldots$.

- Probabilities $p_0, p_1, p_2, \ldots$ of each option.
- Skill assessment of the option m_j that was actually chosen.
- Also generates projected confidence intervals for various statistics.

Why? What to do with it?

Intrinsic skill based on decisions made rather than outcomes.

- In chess, outcomes of games subject to opponent's play, 'luck'.
- 50-odd *games* per year is a small sample—
- —but 1,500 *moves* from those games is a healthy sample.

1. Have outcome-based ratings been consistent over time? (In chess, has there been 'Elo Inflation'?)
2. Compare players of different historical eras.
3. Measure effect of thinking time on skill.
4. Evaluate statistical claims of cheating with computers, or alternatively, 'sandbagging'.
5. Human player training.

Main Principle

The probability p_i of an option m_i depends primarily on its value v_i *in relation to the values of other options.*

Other principles/working assumptions:

1. Decisions at different *turns* t are independent.
2. For agents of all skill levels, the higher v_i, the higher p_i.
3. Weaker agents prefer weaker moves.
4. Values v_i need not be omniscient, only high enough above the agents being modeled to represent an authority figure (*in chess:* 200–400 Elo higher).

Well, 1. is contradicted by human players having multi-turn *plans*, while 2. and 3. contradict each other. But we argue 1. gives "local-limited dependence" while our *full model* harmonizes 2. and 3. via linear combinations—though it violates 4.

Skill Parameters for Agents P

- *Sensitivity* s: How well can P perceive small differences in value?

$$x_i = \frac{v_0 - v_i}{s}$$

  Lower s is better. (Necessary to convert from utility units to dimensionless?)

- *Consistency* c: How (in)frequent are mistakes?

$$p_i \sim \frac{1}{- - (- - x_i - -)^c - -}$$

  Higher c is better.

Full model—not this paper—not as general?

- *Depth of calculation*. Applies to alternating-move games, transactional decision making, 'looking ahead'.

Basic Model

General form: for certain *relation function* R and *curve family* $g = g_{s,c}$,

$$R(p_i, p_0) = g_{s,c}(v_i, v_0).$$

Simplest in the sense that the dependence on v_j for $j \neq i$ is only thru p_0 and the constraint $\sum_j p_j = 1$.

This paper: First *scale* down differences in value according to the overall imbalance in the position, defining

$$x_i = \delta(v_i, v_0)/s.$$

Then take R to be a ratio of logs and g an inverse-exp curve, namely:

$$\frac{\log(1/p_0)}{\log(1/p_i)} = e^{-x_i^c}.$$

Full Model—better? needs attention?

- Values v_i^d from chess engines at different search depths d.
- Apply basic model at each d to get p_i^d.
- Additional skill parameters w_d called *weights*, with $\sum_d w_d = 1$. (Hopefully specifiable by one or two scalars.)
- Overall probability then given by the linear combination

$$p_i = \sum_d w_d p_i^d.$$

Idea: A move whose goodness appears only at higher depths ("swing-up" move) should be less probable than a move whose good value is evident at all depths. Likewise a "swing-down" move (a good 'trap' in chess) should be more probable than a less-deep error, even though both have the same value at the reference depth.

Working assumption of using basic model: Swing-up and swing-down cases offset over large-enough sets of moves.

Implementation Details

1. Values v_i by champion chess program RYBKA 3 run in 50-PV mode to reported depth $d = 13$.
   - Estimated 2650–2700 strength.
   - Usually 6–8 hours per x64 CPU thread per game.
   - RYBKA 3 led field by 200 points in 2008; only program with "Multi-PV cap" feature saving much time; now under cloud for GPL violations.
2. Scaling derived from huge amounts of Single-PV mode data also serving as scientific control:

$$\delta(v_0, v_i) = \int_{v_i}^{v_0} d\mu(x) \quad \text{with} \quad \mu(x) = \frac{1}{1+x}.$$

3. Eliminate turns 1–8 of any game, turns where RYBKA 3 judges more than a 3-pawn advantage, and turns involved in *repetitions*.

A Weird Fact, and its Correction

- In upwards of 10% of moves, Rybka 3 gives equal-top values to two or more moves, i.e. $v_0 = v_1 = \dots$
- All of our models would give equal probabilities to such moves.
- However, in cases of exactly two equal-top moves, the move first-listed by Rybka is preferred **almost 60%** of the time!
- Likewise the second move of an equal-top triad is preferred to the third, although triads etc. are rare enough to ignore.
- Believed cause: The first move gains a higher value at some low depth and stays, since engines keep order of tied moves stable.
- Alas, Rybka 3 hides the lowest plies, so it's hard to tell.
- Hope is that the "full model" will reflect this naturally by w_d for low d picking up the higher value.
- Basic model adjusted by deducting 0.03 from $v_1, v_2, \dots$ in tied cases, which nearly equalizes the probabilities across training sets.

Training the Model

Objective: Find a robust relation between Elo ratings and values $s, c, \ldots$

2800	World's best players
2700	World-class players
2600	Strong grandmaster (GM)
2500	Typical GM
2400	Typical International Master (IM)
2300	Typical FIDE Master (FM)
2200	National Master
2000	Expert (U.S)
1800	Class A (U.S.)
1600	Class B, etc...

Training Test Sets

- All available games under *standard round-robin tournament conditions* where both players were rated within 10 points of an Elo milestone $R = 2700, 2600, \ldots, 2200$, played in three different time periods.
- Time periods 2006–2009, 1991–1994, 1976–1979. (The Elo system was adopted by FIDE in 1971. No 2700 set for 1976–1979.)
- Use statistical fitting to derive s_R, c_R for each R in each period.
- **Main finding:** Little or no difference across time periods for each R, hence no “rating inflation.”
- Since the FIDE rating system was extended below 2200 in the past decade, test sets were compiled for $2100, 2000, \ldots, 1600$, widening the interval from ± 10 to ± 15 or ± 20 to get enough games.
- **Gamescore errors** in these test sets needed manual correction. Frequency about 1% of games, but cause about 20% additional error in master-level sets.

Two-Parameter Fitting: c-fit Convention

- While s varied widely from 0.08 to 0.16 and higher, c was observed to stay in a narrower band, about 0.52 down to 0.43, fairly linearly as R from 2700 to 1600.
- Moreover, the s, c values trade off against each other in long ridges of near-optimality for various fitting methods.
- Hence decided to impose a linear fit to determine c'_R first. (By happenstance almost exactly 0.07 per 100 Elo.)
- Given c'_R, can do one-parameter fit to get s'_R. Quality of fit is reasonably close to that of original s_R, c_R fit.
- *Future: use a tool like Neil Sloane's* gosset to investigate s, c space further.

Fitting Methods—which to use?

1. *Maximum Likelihood:* Maximize the product of probabilities p_j over the moves m_j that were actually played. Gives markedly inferior results.
2. *Bayesian Update:* Used by Haworth in his models, not yet here—should it approach ML values?
3. Solve the two equations that set $\sum p_0$ equal to the actual number of times move m_0 was played, and set the projected error $\sum_i p_i \delta(v_0, v_i)$ summed over all game turns equal to the actual error. (Called 'FF' in tables.)
4. Fit expected vs. actual percentages of playing the second-best, third-best, fourth-best moves $m_1, m_2, m_3, \ldots$ of the analysis engine, as well as m_0. Problem: heteroskedasticity.
5. Fit *probability percentiles* instead, hopefully solving the problem of 4. (Called 'PF', or 'CF' in tables below when c is fitted first.)
6. Some other fitting methods?

The Percentile-Fitting Method

- Pick a percentile grid, e.g. $q = .02, .04, .06, \ldots, .96, .98$ (used in paper).
- A game turn t is a "hit" (scoring 1) if the played move m_j satisfies

$$q \geq c(j) = \sum_{i=0}^{j-1} p_i.$$

- Partial hit if $c(j-1) \leq q \leq c(j)$, scoring $\frac{q - c(j-1)}{c(j) - c(j-1)}$.
- Let r_q be the proportion of hit scores for q over all turns t.
- Minimize $\sum_q (r_q - q)^2$.
- (Alternatives: multiply by $H(q)$ and/or minimize some other ℓ_p distance instead of least-squares.)

Can Basic Model Be Tuned Better?

- PF and CF are observed fairly regularly to:
  - over-predict p_0 by about 0.003 (i.e., 0.3%),
  - under-predict p_1 by about three times as much, and
  - predict probabilities of $m_2, m_3, \dots$ fairly closely.
- Thus PF and CF turn out to be *biased* estimators of p_0...
  - ...but helpfully hedge against false positives in cheating testing.
- FF is an unbiased estimator of p_0, but so-far seems to give less control on other p_j.
- This holds for all reasonable curve families g tried thus far, such as

$$g_c(x) = \frac{1}{1 + x^c} \qquad \text{or} \qquad g_c(x) = \frac{1}{(1+x)^c}.$$

- Changing relation R to simply p_i/p_0 rather than ratio of logs warps the model markedly the other way. Use some R intermediate between them?

Results For 2006–2009

Elo R	s	c	c'_R	s'_R	$\mathrm{mm}_p/\mathrm{mm}_a$	$\mathrm{ad}_p/\mathrm{ad}_a$	Q_{fit}
2700	.078	.503	.513	.080	56.2/56.3	.056/.056	.009
2600	.092	.523	.506	.089	55.0/54.2	.063/.064	.041
2500	.092	.491	.499	.093	53.7/53.1	.067/.071	.028
2400	.098	.483	.492	.100	52.3/51.8	.072/.074	.016
2300	.108	.475	.485	.111	51.1/50.3	.084/.088	.044
2200	.123	.490	.478	.120	49.4/48.3	.089/.092	.084
2100	.134	.486	.471	.130	48.2/47.7	.099/.102	.034
2000	.139	.454	.464	.143	46.9/46.1	.110/.115	.065
1900	.159	.474	.457	.153	46.5/45.0	.119/.125	.166
1800	.146	.442	.450	.149	46.4/45.4	.117/.122	.084
1700	.153	.439	.443	.155	45.5/44.5	.123/.131	.065
1600	.165	.431	.436	.168	44.0/42.9	.133/.137	.129

The s'_R column shows a reasonable progression.

Results Over Time

With c'_R fixed, corresponding s'_R values:

Elo	c'_R	2006–9	1991–4	1976–9
2700	0.513	0.080	0.084	n.a.
2600	0.506	0.089	0.087	0.087
2500	0.499	0.093	0.092	0.091
2400	0.492	0.100	0.103	0.103
2300	0.485	0.111	0.117	0.116
2200	0.478	0.120	0.122	n.a.?

Beyond Paper: Supporting evidence (of “no inflation”) from the Single-PV control data that Average-Difference (*ad*) levels for round-robin tournaments of each given FIDE *category* are consistent over time.

Operating the Model: 1. Skill Assessment

To compute an **Intrinsic Performance Rating** (IPR) for a player P over a set of games:

- Fit s, c for those games.
- Iterate to find a stable point on the s'_R, c'_R line.
- Output the corresponding R.

Beyond Paper—Some Old Masters:

Player	Years	IPR
Howard Staunton	1841–1858	1990
Adolf Anderssen	1851–1878	2060
Paul Morphy	1857–1859	2310
Wilhelm Steinitz	1860–1876	2220

Figures are somewhat subject to change.

Intrinsic Categories of Historical Tournaments

Beyond Paper: Can do IPR's for entire tournaments too.

Event	Year	IPR
St. Petersburg quad	1896	2390
St. Petersburg prelims	1914	2370
St. Petersburg finals	1914	2570(!)
New York	1927	2620
AVRO	1938	2620
The Hague WC	1948	2640
Curacao CT	1962	2580(!)

Operating the Model: 2. Cheating Testing

1. Let $[m_{E_t}]$ be a sequence of moves over game turns t that are liked by some engine E.
2. Let $[m_{j_t}]$ be the sequence of played moves. How close 'should' these sequences be? for a non-cheating player?
3. Plug in s_R, c_R for the player's rating R, or an upper bound on it.
4. Then $M = \sum_t p(E_t)$ is the expected number of *matches* to E.
5. And $\sigma^2 = \sum_t p(E_t)(1 - p(E_t))$ is the projected variance.
6. Output the *z-score* $z = \frac{\hat{M} - M}{\sigma}$, where $\hat{M}$ is the player's actual number of matches.
7. Is this a reliable way of testing null hypotheses of no cheating?

Beyond Paper: Projected vs. Actual σ

- How close are z-scores from player performances in events to a normal distribution?
- Are they robust enough for use in court cases?
- Is the model capturing a significant enough aspect of skill (at chess)?

Distributional Testing Method: For each Elo milepost R,

1. Generate 10,000 random subsets of 9 games each from the training set for R. [Why 9? Typical for major events.]
2. Also randomly choosing Black or White in each game. Pretend the resulting set S of moves is a performance by one 'player.'
3. Tabulate the z-scores z_S.

Distributional Results: FF Method

z-rng	-4.x	-3.x	-2.x	-1.x	-0.x	+0.x	+1.x	+2.x	+3.x	+4
Targ.	0.3	13	214	1359	3413	3413	1359	214	13	0.3
2700	0	23	248	1350	3223	3266	1508	353	28	1
2600	1	44	236	1313	3127	3220	1708	321	28	2
2500	3	55	430	1684	3059	2960	1415	345	46	3
2400	1	30	309	1484	3128	3121	1510	377	39	1
2300	0	36	322	1606	3133	2996	1465	389	52	1
2200	1	47	364	1451	3231	3136	1420	324	26	0
2100	1	24	299	1418	3131	3281	1477	338	29	2
2000	0	40	344	1410	3102	3212	1500	350	41	1

Somewhat flattened compared to normal distribution, indicating some loss of modeling fidelity. But not too bad.

CF Method—negative skew is deliberate

z-rng	-4.x	-3.x	-2.x	-1.x	-0.x	+0.x	+1.x	+2.x	+3.x	+4
Targ.	0.3	13	214	1359	3413	3413	1359	214	13	0.3
2700	1	29	330	1585	3350	3105	1304	278	18	0
2600	3	63	418	1838	3373	2947	1171	174	13	0
2500	5	97	658	2069	3199	2609	1095	243	24	1
2400	1	45	392	1699	3269	2953	1318	294	29	0
2300	3	63	560	1991	3389	2612	1103	247	32	0
2200	7	97	678	2117	3467	2601	886	138	9	0
2100	3	38	407	1734	3423	2958	1186	233	17	1
2000	2	73	505	1862	3304	2918	1113	205	18	0

Flattening still makes +2.x and +3.x columns higher than desired, but again not too bad.

Conclusions and Future Work

- Model captures enough chess skill to show steady progression of parameters with regard to Elo ratings.
- And to approximate a normal distribution of actual versus expected matching, reasonably closely.
- Scientifically effective on cheating-allegation cases. [Show Elista 2006 demo if time.]

To-do list:

1. Analysis Interchange Format (AIF) standard so others can compile analyzed games.
2. Implement full model, to see if it is really better.
3. Tune model(s) better to eliminate the p_0, p_1, p_2 slight warping.
4. Further application to chess and other games.
5. Extend to (non-game) decision-making settings.