Lecture 3: Sports rating models

David Aldous

August 31, 2017

- **Sports** are a popular topic for course projects – usually involving details of some specific sport and statistical analysis of data.
- In this lecture we imagine some non-specific sport; either a team sport – *(U.S.) football, baseball, basketball, hockey; soccer, cricket* – or an individual sport or game – *tennis, chess, boxing*.
- We consider only sports with matches between two teams/individuals. But similar ideas work where there are many contestants – *athletics, horse racing, automobile racing, online video games*.

Let me remind you of three things you already know about sports.

Reminder 1. Two standard "centralized" ways to schedule matches: league or tournament.

2016–17 Premier League

Standings

#	Team	MP	W	D	L	GF	GA	GD	Pts
1	Chelsea	38	30	3	5	85	33	52	**93**
2	Tottenham	38	26	8	4	86	26	60	**86**
3	Man. City	38	23	9	6	80	39	41	**78**
4	Liverpool	38	22	10	6	78	42	36	**76**
5	Arsenal	38	23	6	9	77	44	33	**75**
6	Man United	38	18	15	5	54	29	25	**69**
7	Everton	38	17	10	11	62	44	18	**61**
8	Southampton	38	12	10	16	41	48	-7	**46**
9	Bournemouth	38	12	10	16	55	67	-12	**46**
10	West Brom	38	12	9	17	43	51	-8	**45**
11	West Ham	38	12	9	17	47	64	-17	**45**
12	Leicester City	38	12	8	18	48	63	-15	**44**
13	Stoke City	38	11	11	16	41	56	-15	**44**
14	Crystal Palace	38	12	5	21	50	63	-13	**41**
15	Swansea City	38	12	5	21	45	70	-25	**41**
16	Burnley FC	38	11	7	20	39	55	-16	**40**
17	Watford	38	11	7	20	40	68	-28	**40**
18	Hull City	38	9	7	22	37	80	-43	**34**
19	Middlesbrough	38	5	13	20	27	53	-26	**28**
20	Sunderland	38	6	6	26	29	69	-40	**24**

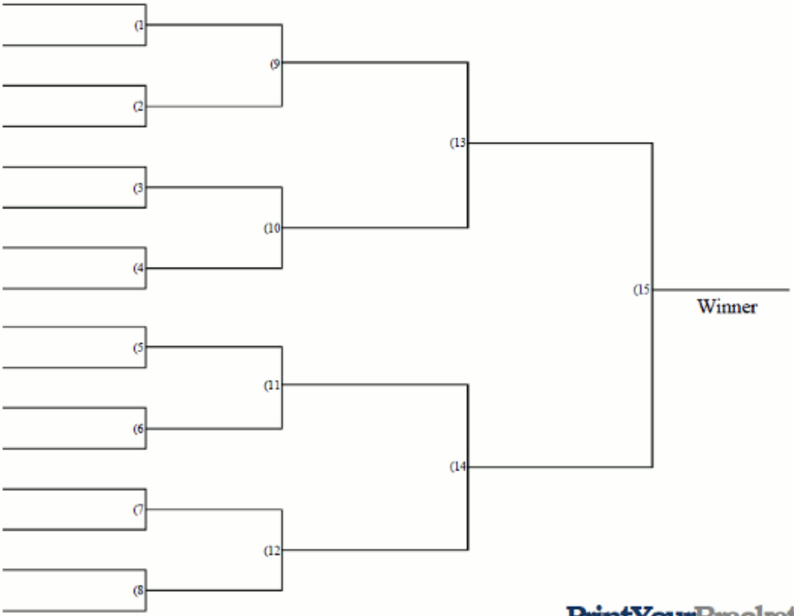
Show less

Premier League Table: Final Standings For 2016-2017 Season

heavy.com/.../english-premier-league-bpl-epl-table-standings-leader-final-uefa-relegat... ▾

May 21, 2017 - For the second time in three years, the **Premier League** season ends with Chelsea on top.

16 Team Single Elimination



PrintYourBrackets.com

These schemes are clearly “fair” and produce a “winner”, though have two limitations

- Limited number of teams.
- Start anew each year/tournament.
- Require central organization – impractical for games (*chess*, *tennis*) with many individual contestants.

Reminder 2. In most sports the winner is decided by *point difference*. One could model point difference but we won't. For simplicity we will assume matches always end in win/lose, no ties.

Reminder 3. A main reason why sports are interesting is that the outcome is uncertain. It makes sense to consider the **probability** of team A winning over team B. In practice one can do this by looking at gambling odds [next slide]. Another lecture will discuss data and theory concerning how probabilities derived from gambling odds or prediction markets change over time.

Winner of 2017-18 season Superbowl – gambling odds at start of season, converted to implied probabilities (PredictWise)

8/30/2017 2017-18 NFL Super Bowl – PredictWise

Outcome	PredictWise	Derived Betfair Price	Betfair Back	Betfair Lay
New England Patriots	22 %	$ 0.200	5.00	5.10
Pittsburgh Steelers	8 %	$ 0.080	12.00	12.50
Green Bay Packers	8 %	$ 0.080	12.50	13.00
Seattle Seahawks	8 %	$ 0.077	12.50	13.00
Dallas Cowboys	6 %	$ 0.056	17.00	18.00
Atlanta Falcons	5 %	$ 0.054	18.00	19.00
Oakland Raiders	5 %	$ 0.050	20.00	21.00
New York Giants	3 %	$ 0.036	27.00	28.00
Carolina Panthers	3 %	$ 0.031	32.00	34.00
Kansas City Chiefs	3 %	$ 0.029	32.00	36.00
Tennessee Titans	3 %	$ 0.029	34.00	36.00
Arizona Cardinals	3 %	$ 0.028	34.00	36.00
Denver Broncos	2 %	$ 0.025	40.00	44.00
Minnesota Vikings	2 %	$ 0.025	36.00	42.00
Tampa Bay Buccaneers	2 %	$ 0.023	40.00	44.00
Houston Texans	2 %	$ 0.021	44.00	48.00
Philadelphia Eagles	2 %	$ 0.021	46.00	50.00
Baltimore Ravens	2 %	$ 0.017	60.00	65.00
Los Angeles Chargers	1 %	$ 0.014	60.00	70.00

Obviously the probability A beats B depends on the **strengths** of the teams – a better team is likely to beat a worse team. So the problems

- estimate the strengths of A and B
- estimate the probability that A will beat B

must be closely related. This lecture talks about two ideas for making such estimates which have been well studied. However the connection between them has not been so well studied, and is suitable for simulation-style projects.

Terminology. I write **strength** for some hypothetical objective numerical measure of how good a team is – which we can't observe – and **rating** for some number we can calculate by some formula based on past match results. Ratings are intended as estimates of strengths.

Idea 1: The basic probability model.

Each team A has some "strength" x_A, a real number. When teams A and B play

$$\mathbb{P}(\text{A beats B}) = W(x_A - x_B)$$

for a specified "win probability function" W satisfying the conditions

$$\begin{gathered} W : \mathbb{R} \to (0,1) \text{ is continuous, strictly increasing} \\ W(-x) + W(x) = 1; \quad \lim_{x \to \infty} W(x) = 1. \end{gathered} \tag{1}$$

Implicit in this setup, as mentioned before

- each game has a definite winner (no ties);
- no home field advantage, though this is easily incorporated by making the win probability be of the form $W(x_A - x_B \pm \Delta)$;
- not considering more elaborate modeling of point difference

and also

- strengths do not change with time.

Some comments on the math model.

$$\mathbb{P}(\text{A beats B}) = W(x_A - x_B)$$

$$W : \mathbb{R} \to (0,1) \text{ is continuous, strictly increasing}$$

$$W(-x) + W(x) = 1; \quad \lim_{x \to \infty} W(x) = 1.$$

There is a reinterpretation of this model, as follows. Consider the alternate model in which the winner is determined by point difference, and suppose the random point difference D between two teams of equal strength has some (necessarily symmetric) continuous distribution not depending on their common strength, and then suppose that a difference in strength has the effect of increasing team A's points by $x_A - x_B$. Then in this alternate model

$$\mathbb{P}(\text{A beats B}) = \mathbb{P}(D + x_A - x_B \geq 0) = \mathbb{P}(-D \leq x_A - x_B) = \mathbb{P}(D \leq x_A - x_B).$$

So this is the same as our original model in which we take W as the distribution function of D.

This basic probability model has undoubtedly been re-invented many times; in the academic literature it seems to have developed "sideways" from the following type of statistical problem. Suppose we wish to rank a set of movies $A, B, C, \ldots$ by asking people to rank (in order of preference) the movies they have seen. Our data is of the form
(person 1): C, A, E
(person 2): D, B, A, C
(person 3): E, D
....
One way to produce a consensus ranking is to consider each pair (A, B) of movies in turn. Amongst the people who ranked both movies, some number $i(A, B)$ preferred A and some number $i(B, A)$ preferred B. Now reinterpret the data in sports terms: team A beat B $i(A, B)$ times and lost to team B $i(B, A)$ times. Within the basic probability model (with some specified W) one can calculate MLEs of strengths $x_A, x_B, \ldots$ which imply a ranking order.

This method, with W the logistic function (discussed later), is called the *Bradley-Terry* model, from the 1952 paper *Rank analysis of incomplete block designs: I. The method of paired comparisons* by R.A. Bradley and M.E. Terry.

An account of the basic Statistics theory (MLEs, confidence intervals, hypothesis tests, goodness-of-fit tests) is treated in Chapter 4 of H.A. David's 1988 monograph *The Method of Paired Comparisons*.

So one can think of **Bradley-Terry as a sports model** as follows: take data from some past period, calculate MLEs of strengths, use to predict future win probabilities.

Considering Bradley-Terry as a sports model:

positives:

- allows unstructured schedule;
- use of logistic makes algorithmic computation straightforward.

negatives:

- use of logistic completely arbitrary: asserting

$$\text{if } \mathbb{P}(i \text{ beats } j) = 2/3, \ \mathbb{P}(j \text{ beats } k) = 2/3 \text{ then } \mathbb{P}(i \text{ beats } k) = 4/5$$

as a universal fact seems ridiculous; (**project;** any data?)

- by assuming unchanging strengths, it gives equal weight to distant past as to recent results;
- need to recompute MLEs for **all** teams after each match.

The Bradley-Terry model could be used for interesting course **projects** – one is described next.

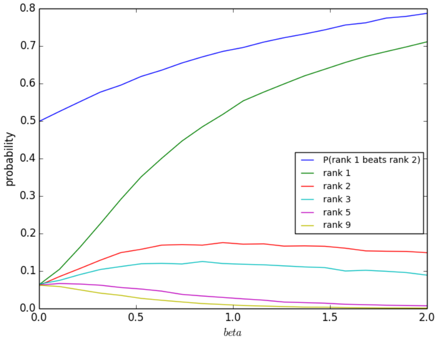
Another possible **project:** take the Premier league data and ask *what is the probability that Chelsea was actually the best team in 2016-17?*.

Robustness of second seed winning probability. A mathematically natural model for relative strengths of top players is $(\beta\xi_i, i \geq 1)$ where $0 < \beta < \infty$ is a scale parameter and

$$\xi_1 > \xi_2 > \xi_3 > \dots$$

form the inhomogeneous Poisson point process on $\mathbb{R}$ of intensity e^{-x} arising in extreme value theory. So we can simulate a tournament, in the conventional deterministic-over-seeds pattern to see which seed wins. The next graphic gives data from simulations of such a tournament with 16 players, assuming the seeding order coincides with the strength order. The interpretation of the parameter β is not so intuitive, but that parameter determines the probability that the top seed would beat the 2nd seed in a single match, and this probability is shown (as a function of β) in the top curve in Figure 1. The other curves show the probabilities that the 16-player tournament is won by the players seeded as 1, 2, 3, 5 or 9.

Figure: Probabilities of different-ranked players winning the tournament, compared with probability that rank-1 player beats rank-2 player (top curve).



The results here are broadly in accord with intuition. For instance it is obvious that the probability that the top seed is the winner is monotone in β. What is perhaps surprising and noteworthy is that the probability that the 2nd seed player is the winner is quite insensitive to parameter values, away from the extremes, at around 17%.

Is this 17% prediction in fact accurate?
How robust is it to alternate models?
As a start, data from tennis tournaments[1] in Table 1 shows a moderately good fit.

Table: Seed of winner, men's and women's singles, Grand Slam tennis tournaments, 1968 - 2016.

seed of winner	1	2	3	4	5+	total
frequency	148	94	42	29	77	390
percentage	38%	24%	11%	7%	20%	100%
model, $\beta = 0.65$	41%	17 %	11 %	7 %	24%	100%

[1] Wimbledon, and the U.S., French and Australian Opens form the prestigious "Grand Slam" tournaments.

Reminder 4. Another aspect of what makes sports interesting to a spectator is that strengths of teams change over time – if your team did poorly last year, then you can hope it does better this year.

In the context of the Bradley-Terry model, one can extend the model to allow changes in strengths. Seem to be about 2-3 academic papers per year which introduce some such extended model and analyze some specific sports data. Possible source of course **projects** – apply to different sport or to more recent data.

Anchor data: [show World Football Elo ratings]

(1) The International football teams of Germany and France currently (August 2017) have Elo ratimgs of 2080 and 1954, which (as described later) can be interpreted as an implictly estimated probability 67% of Germany winning a hypothetical upcoming match.

(2) Assertion **Ratings tend to converge on a team's true strength relative to its competitors after about 30 matches.**

Central questions.
(1) Why is there any connection between the ratings and probabilities?
(2) Is there any theory or data behind this **thirty matches suffice** assertion?

Re (2), by analogy a search on **seven shuffles suffice** gets you to discussions which can be tracked back to an actual theorem Bayer-Diaconis (1992).
(Note **project**: math theory and data for card-shuffling).

History: Elo ratings originally used for chess, then for other individual games like tennis, now widely used in online games. I write "player" rather than team.


Safari File Edit View History Bookmarks Develop Window Help
Thu 7:37 PM David Aldous
aldousdj in Jungle Gin > Social > Dr. Livingston, I Presume?
monthly.tex
Jungle Gin | Pogo.com Free Online Card Games
www.pogo.com/games/gin#game_gin2-plgin2sf027:table_23
Club Pogo Apple Yahoo! Google Maps YouTube Wikipedia News Popular
How to Play
My Friends List
Legend
Rated Game
Oklahoma
Table Info
Beginner
Intermediate
Advanced
Expert
Master
STAND UP
How to Play
tokens
1,043,250
KNOCK
Who's Here (2)
Rating
yeloroses1 1446
aldousdj 1519
aldousdj gets 28 points
yeloroses1 aldousdj
44 90
yeloroses1:
aldousdj:
Your rating has been updated. Your new rating is 1505. Your old rating was 1525.
Your rating has been updated. Your new rating is 1519. Your old rating was 1505.
100 POINT GAME
500 POINT MATCH
yeloroses1 44
aldousdj 90
SCORE summary
My Fav
See all f

Idea 2: Elo-type rating systems

(not ELO). The particular type of rating systems we study are known loosely as Elo-type systems and were first used systematically in chess. The Wikipedia page *Elo rating system* is quite informative about the history and practical implementation. What I describe here is an abstracted "mathematically basic" form of such systems.

Each player i is given some initial rating, a real number y_i. When player i plays player j, the ratings of both players are updated using a function Υ (`Upsilon`)

$$\begin{aligned} \text{if } i \text{ beats } j \text{ then } y_i \to y_i + \Upsilon(y_i - y_j) \text{ and } y_j \to y_j - \Upsilon(y_i - y_j) \\ \text{if } i \text{ loses to } j \text{ then } y_i \to y_i - \Upsilon(y_j - y_i) \text{ and } y_j \to y_j + \Upsilon(y_j - y_i)\ . \end{aligned} \tag{2}$$

Note that the sum of all ratings remains constant; it is mathematically natural to center so that this sum equals zero.

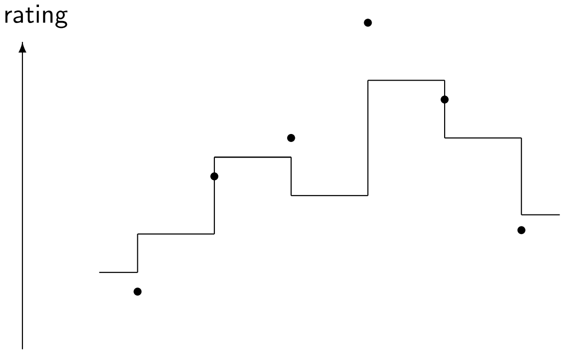
[Note: in practice the scheme is adapted to each specific sport, for instance for International Football . . .]

The Elo ratings are based on the following formulas:

$$R_n = R_o + K \times (W - W_e)$$

- $\mathbf{R_n}$ is the new rating, $\mathbf{R_o}$ is the old (pre-match) rating.
- **K** is the weight constant for the tournament played:
- **60** for World Cup finals;
- **50** for continental championship finals and major intercontinental tournaments;
- **40** for World Cup and continental qualifiers and major tournaments;
- **30** for all other tournaments;
- **20** for friendly matches.
- **K** is then adjusted for the goal difference in the game. It is increased by **half** if a game is won by two goals, by **3/4** if a game is won by three goals, and by **3/4 + (N-3)/8**if the game is won by four or more goals, where **N** is the goal difference.
- **W** is the result of the game (**1** for a win, **0.5** for a draw, and **0** for a loss).
- $\mathbf{W_e}$ is the expected result (win expectancy), either from the chart or the following formula:
- $W_e = 1 / (10^{(-dr/400)} + 1)$
- **dr** equals the difference in ratings plus **100** points for a team playing at home.

Schematic of one player's ratings after successive matches. The • indicate each opponent's rating.



Math comments on the Elo-type rating algorithm.

We require the function $\Upsilon(u), \ -\infty < u < \infty$ to satisfy the qualitative conditions

$$\Upsilon : \mathbb{R} \to (0, \infty) \text{ is continuous, strictly decreasing, and } \lim_{u \to \infty} \Upsilon(u) = 0. \tag{3}$$

We will also impose a quantitative condition

$$\kappa_\Upsilon := \sup_u |\Upsilon'(u)| < 1. \tag{4}$$

To motivate the latter condition, the rating updates when a player with (variable) strength x plays a player of fixed strength y are

$$x \to x + \Upsilon(x - y) \text{ and } x \to x - \Upsilon(y - x)$$

and we want these functions to be *increasing* functions of the starting strength x.

Note that if Υ satisfies (3) then so does $c\Upsilon$ for any scaling factor $c > 0$. So given any Υ satisfying (3) with $\kappa_\Upsilon < \infty$ we can scale to make a function where (4) is satisfied.

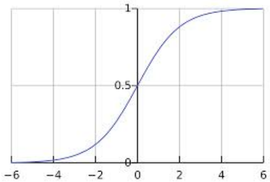
The logistic distribution function

$$F(x) := \frac{e^x}{1 + e^x}, -\infty < x < \infty$$

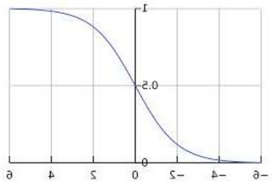
is a common choice for the "win probability" function $W(x)$ in the basic probability model; and its complement

$$1 - F(x) = F(-x) = \frac{1}{1 + e^x}, -\infty < x < \infty$$

is a common choice for the "update function shape" $\Upsilon(x)$ in Elo-type rating systems. That is, one commonly uses $\Upsilon(x) = cF(-x)$.



possible $W(x)$



possible $\Upsilon(x)$

Whether this is more than a convenient choice is a central issue in this topic.

Elo is an algorithm for producing ratings (and therefore rankings) which (unlike Bradley-Terry) does not assume any probability model. It implicitly attempts to track changes in strength and puts greater weight on more recent match results.

How good are Elo-type algorithms? This is a subtle question – we need to

- use Elo to make predictions
- choose how to measure their accuracy
- compare accuracy with predictions from some other ranking/rating scheme (such as Bradley-Terry or gambling odds).

The simplest way to compare schemes would be to look at matches where the different schemes ranked the teams in opposite ways, and see which team actually won. But this is not statistically efficient [board]. Better to compare schemes which predict **probabilities**.

Although the Elo algorithm does not say anything *explicitly* about probability, we can argue that it *implicitly* does predict winning probabilities.

A math connection between the probability model and the rating algorithm.

Consider n teams with unchanging strengths $x_1, \ldots, x_n$, with match results according to the basic probability model with win probability function W, and ratings (y_i) given by the update rule with update function Υ. When team i plays team j, the expectation of the rating change for i equals

$$\Upsilon(y_i - y_j)W(x_i - x_j) - \Upsilon(y_j - y_i)W(x_j - x_i). \tag{5}$$

So consider the case where the functions Υ and W are related by

$$\Upsilon(u)/\Upsilon(-u) = W(-u)/W(u), \quad -\infty < u < \infty.$$

In this case

> *(*) If it happens that the difference $y_i - y_j$ in ratings of two players playing a match equals the difference $x_i - x_j$ in strengths then the expectation of the change in rating difference equals zero*

whereas if unequal then (because Υ is decreasing) the expectation of $(y_i - y_j) - (x_i - x_j)$ is closer to zero after the match than before.

$$\Upsilon(u)/\Upsilon(-u) = W(-u)/W(u), \quad -\infty < u < \infty. \tag{6}$$

These observations suggest that, under relation (6), there will be a tendency for player i's rating y_i to move towards its strength x_i though there will always be random fluctuations from individual matches. So if we believe the basic probability model for some given W, then in a rating system we should use an Υ that satisfies (6).

Now "everybody (in this field) knows" this connection, but nowhere is it explained clearly and no-one seems to have thought it through (practitioners focus on fine-tuning to a particular sport). The first foundational question we might ask is

What is the solution of (6) for unknown Υ?

This can be viewed as the setup for a mathematician/physicist/statistician/data scientist joke.

`Problem.` For given W solve

$$\Upsilon(u)/\Upsilon(-u) = W(-u)/W(u), \quad -\infty < u < \infty.$$

`Solution`

- physicist (Elo): $\Upsilon(u) = cW(-u)$
- mathematician: $\Upsilon(u) = W(-u)\phi(u)$ for many symmetric $\phi(\cdot)$.
- statistician: $\Upsilon(u) = c\sqrt{W(-u)/W(u)}$ (variance-stabilizing ϕ).
- data scientist: well I have this deep learning algorithm ......

These answers are all "wrong" for different reasons. I don't have a good answer to "what Υ to use?" for given W (**project?**). But the opposite question is easy: given Υ, there is a unique "implied win-probability function W" given by

$$W_\Upsilon = \frac{\Upsilon(-u)}{\Upsilon(-u) + \Upsilon(u)}$$

Conclusion: Using Elo with a particular Υ is conceptually equivalent to believing the basic probability model with

$$W_\Upsilon = \frac{\Upsilon(-u)}{\Upsilon(-u) + \Upsilon(u)}$$

Relating our math set-up to data

In published real-world data, ratings are integers, mostly in range $1000 - 2000$. Basically, 1 standard unit (for logistic) in our model corresponds to 174 rating points by convention. So the implied probabilities are of the form [football]

$$\mathbb{P}(\text{ Germany beats France }) = L((2080 - 1954)/174) = 0.67.$$

By convention a new player is given a 1500 rating. If players never departed, the average rating would stay at 1500. However, players leaving (and no re-centering) tends to make the average to drift upwards. This makes it hard to compare "expert" in different sports.

My own, more mathematical write-up of this topic is in a paper *Elo Ratings and the Sports Model: a Neglected Topic in Applied Probability?*, and more possible **projects** are suggested there.

We have answered our first "central question":

Central questions.
(1) Why is there any correction between the ratings and probabilities?
(2) Is there any theory or data behind this **thirty matches suffice** assertion?

Recall that if strengths did not change then we can just use the basic probability model and use MLEs of the strengths. The conceptual point of Elo is to try to track **changing** strengths. So question (2) becomes

How well does Elo track changing strengths?

How well does Elo track changing strengths?

Too hard as theory – can only study via simulation. What do we need to specify, to make a model we can simulate and assess?

1. Distribution of strengths of teams (at fixed time): has a "spread" parameter σ
2. How strengths change with time (time-stationary): has a "relaxation time" parameter τ
3. Scheduling of matches
4. Update function Υ and win-probability function W; and a scale constant c; update by $c\Upsilon(\cdot)$
5. Assess accuracy?

Item 2 is conceptually hardest . . . ; I use several "qualitatively extreme" models. For item 3 I use random matching of all teams, each time unit.

For distribution of strengths use normal, $\sigma = 0.5$ or 1.0, which matches real data. For changes in strengths over time use

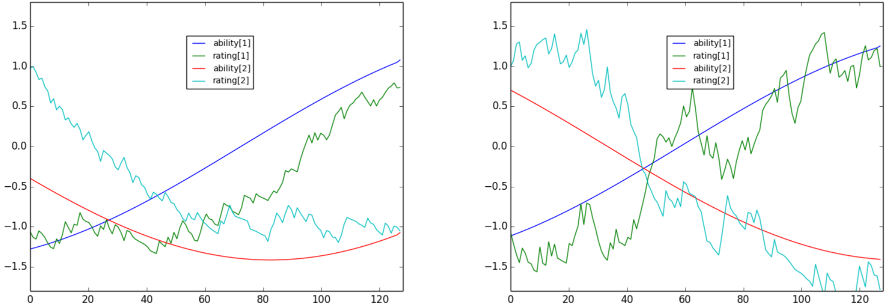
- Cyclic
- Ornstein-Uhlenbeck (ARMA)
- Hold (quite long time), jump to independent random.

each with a "relaxation time" parameter τ. For the win-probability function W and the update function Υ use

- Logistic
- Cauchy
- Linear over $[-1, 1]$

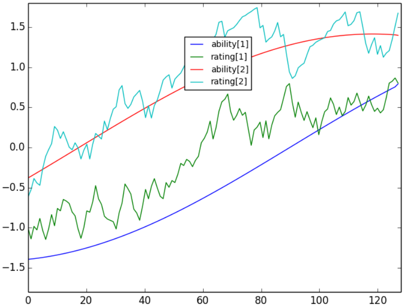
Use different scalings c for updates $c\Upsilon$.

Figure: Realizations of the cycle model: $\sigma = 1$, $\tau = 100$, logistic W and Υ.



$c = 0.17$ (left) and $c = 0.35$ (right).
This shows the intuitively obvious lag-bias versus noise effect.

In the setting above here is the optimal scaling $c = 0.26$



We need to quantify " How well does Elo track changing strengths" in some way. Here's my way.

Consider players A, B at a given time with actual strengths x_A, x_B and Elo ratings y_A, y_B. The actual probability (in our model) that A beats B is $W(x_A - x_B)$ whereas the probability estimated from Elo is $W(y_A - y_B)$. So we calculate

$$\text{root-mean-square of the differences } W(y_A - y_B) - W(x_A - x_B)$$

averaging over all players and all times. I call this **RMSE-p**, the "p" as a reminder we're estimating probabilities, not strengths.

Key question: If willing to believe that (with appropriate choices) this is a reasonable model for real-world sports, what actual numerical values do we expect for RMSE-p?

Short answer; No plausible model gives RMSE-p much below 10%. This is when we are running models forever (i.e. stationary) so hard to reconcile with "30 matches suffice".

Conceptually, there are 3 constituents of error

- **mismatch** between W and Υ.
- **lag** from changes in strengths in our past data
- **noise** from randomness of recent results.

Can do optimal trade-off between latter two by adjusting the scaling c in update $c\Upsilon$, to find the optimal c (given other parameters)

We can estimate the mismatch error from the deterministic limit in which $c \to 0$ for unchanging strengths.

Table: RMSE-p mismatch error.

W Υ	logistic linear	logistic Cauchy	Cauchy logistic	Cauchy linear	linear logistic	linear Cauchy
$\sigma = 0.5$	0.9%	1.1%	1.2%	2.4%	3.2%	6.4%
$\sigma = 1.0$	2.9%	2.9%	2.5%	5.4%	3.2%	6.0%

These errors are perhaps surprisingly small. Now let us take W and Υ as logistic, so no mismatch error. The next table shows the effect of changing the relaxation time τ of the strength change process.

Table: RMSE-p and (optimal c) for O-U model (top) and jump model (bottom)

σ	τ 50	100	200	400
0.5	12.9% (0.11)	11.1% (0.09)	9.5% (0.08)	8.2 % (0.07)
1.0	17.0% (0.28)	14.6 % (0.24)	12.4% (0.16)	10.4% (0.14)

σ	τ 50	100	200	400
0.5	12.8% (0.12)	11.2% (0.09)	9.8% (0.06)	8.4 % (0.05)
1.0	16.9% 0.30	14.5 % 0.24	12.4% 0.19	10.5% 0.14

This shows the intuitively obvious effect that for larger τ we can use smaller c and get better estimates. But curious that numerics in the two models are very close. One can do heuristics (and proofs if one really wanted to) for order-of-magnitude scalings as $c \downarrow 0$ but hardly relevant to real-world cases.

Bottom line from simulations: If you want RMSE-p to be noticeably less than 10% then you need to have played 400 matches and you need that strengths do not change substantially over 200 matches.

Games per year, regular season.

U.S. Football	16
Aussie Rules	22
U.K. Premier League	38
U.S. Basketball	82
U.S. Baseball	162

Some other aspects of rating models.

1. Recent book "The Science of Ranking and Rating" treats methods using undergraduate linear algebra. The lecture in this course in 2014 was based more on that book (link on web page).

2. People who attempt realistic models of particular sports, using e.g. statistics of individual player performance, believe their models are much better than general-sport methods based only on history of wins/losses or point differences. But a recent paper *Statistics-free sports prediction* claims that (using more complex prediction schemes) they can do almost as well using only match scores.

3. I have talked about comparing different schemes which predict **probabilities** – after we see the actual match results, how do we decide which scheme is better? I will discuss this in a different context, the lecture on Geopolitics forecasting.

4. Both schemes are poor at assessing new players. Xbox Live uses its "TrueSkill ranking system" [show page] which estimates both a rating and the uncertainty in the rating, as follows.

XBOX
$59 99
Estimated Retail Price
HALO 4
RP
May contain content inappropriate for children. Visit www.esrb.org for rating information.
Online Interactions Not Rated by the ESRB
Halo® 4 sold separately. *See back for details
Microsoft Studios
343
12 MONTH
XBOX LIVE GOLD MEMBERSHIP
+1
BONUS MONTH
Play Halo 4 online with friends with 13 months of Xbox LIVE Gold.*

Here a rating for player i is a pair (μ_i, σ_i), and the essence of the scheme is as follows. When i beats j
(i) first compute the conditional distribution of X_i given $X_i > X_j$, where X_i has Normal(μ_i, σ_i^2) distribution
(ii) then update i's rating to the mean and s.d. of that conditional distribution.
Similarly if i loses to j then i's rating is updated to the mean and s.d. of the conditional distribution of X_i given $X_i < X_j$.

Discussion. `The authors seem to view this as an approximation to some coherent Bayes scheme, but to me it fails to engage both ‘‘uncertainty about strength" and ‘‘uncertainty about match outcome".`

So another simulation project is to compare this to other schemes. Note this implicitly predicts winning probabilities via $\mathbb{P}(X_i > X_j)$.

People often think that bookmakers adjust their offered odds so that, whatever the outcome, they never lose money. This just isn't true. [show ESPN article]