Pg. 10, Proof of Theorem 2,
is the proof that ELO ratings corresponding to the betting interpretation
converge to the true strengths.

Notes on project in progress. The main purpose is to find interesting math problems. Various problems are noted; they vary in apparent difficulty and interesting-ness!

Mathematical probability foundations of dynamic sports ratings

1 Introduction

There is a natural probability model (which we call, rather unimaginatively, the *basic probability model*: section 1.1) for sport results: each team has a *strength*, and the probability A beats B is a specified function W of their difference in strengths. One might regard this model as basic to two established fields. Over the last decade, influenced no doubt by *Moneyball* [7] and the growing popularity of fantasy sports leagues, not to mention real-money gambling, there has been intense activity in statistical analysis of sports data and predictive models for professional team sports results. Naturally, such work emphasizes realistic models incorporating the detailed rules of a particular sport, details of the performance of individual players at different positions, and methodology for practical analysis of available data. Perhaps because of this detail-orientation, this field has attracted little attention in the broader applied probability research community, despite its wide use of probability models.

The second field concerns consensus ranking. 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, which imply a ranking order.

Instead of pursuing these well-studied topics, this article discusses how one might use the model in other ways, which apparently have not been studied systematically. Our work is intended as "foundational mathematics" not intended for explicit data analysis or for use as a predictive model, which would require less simplistic models. Here are some topics we address.

Although we have used the word *team* the model clearly can be considered for individual players in sports such as chess, tennis, or boxing, and we write *team* or *player* according to which seems more common in a particular context.

- There is a standard way to design a single-elimination tournament in terms of the seeding (ranking) of players. Within the basic probability model, does this particular design maximize the chance of the highest-ranked player winning the tournament? Does it optimize anything? (section 6).

- Suppose each of n teams plays each other team once during a season. Each team i wins some proportion $q^*(i)$ of its matches. Under a given probability distribution over the entire season match results there is some expected proportion $q(i)$ of matches won by team i. It is easy to characterize the possible values of the sequence $(q(i), 1 \leq i \leq n)$ for a completely arbitrary probability distribution. But under our basic probability model (with arbitrary W and strengths) what are the possible values? (section 5).

- At the end of a league season or tournament we can ask *what is the chance the winner was actually the best team?*. This is fun to do in a popular talk, and of course requires a probability model (section 7).

But our main focus is in another direction. In the usual setting of professional team sports, there are a limited number of teams, systematic scheduling of matches, and extensive public data available. But consider instead games more typically played by many amateur individuals, with non-systematic scheduling and where we only note win-loss (or win-loss-draw) data. Chess and tennis are the classical examples, while online games provide contemporary examples. In such games there are natural algorithms (which we will call *Elo-type* and are based on some *update function* Υ: see section 1.2) which produce *ratings* intended to estimate relative strengths of players. Such algorithms have nothing to do with probability, *a priori*. But as noted in section 2.1, there is an obvious *heuristic* connection between the probability model and the rating algorithm. In trying to understand this connection more sharply, the first question that occurred to the author was the following.

> When a collection of players play repeatedly with results according to the basic probability model, how do the ratings produced

> by the rating algorithm compare to the actual strengths of the players?

Curiously, this question has apparently not be studied very systematically, and we make a start in section 3.

As implicitly noted above, we will need to model the *schedule* by which matches are arranged. We write *league play* for the model of n teams, where each pair of teams plays exactly once (or exactly twice) during a season; this is also called a *round-robin tournament* (unfortunately the word *tournament* is used with many meanings in the real world and in mathematics – see [16] and section 5.2). League play, and the usual *single-elimination tournament*, represent two extremes of what one might call "template" schedules. These are conceptually different from the "many individual players" setting of matches being scheduled on the basis of ratings, either explicitly or more commonly implicitly via player choice. How to model such haphazard scheduling is of course a major conceptual issue (section 8.2).

Another difference between the "league of teams" and the "many individual players" settings is that in the former the teams' strengths are roughly comparable, whereas in the latter the top players may be much stronger than the average player. In the latter case, the mathematically natural model for strengths of the top players is described in section 8.1, where we reconsider the basic probability model and rating algorithms in that setting.

Finally, the basic probability model assumes constant strengths but a major purpose of rankings is to track *changes* in strengths. Models for such changes and effectiveness of ratings in tracking changes are discussed in section 4.

1.1 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 following conditions (which we regard as the minimal natural 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:

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 $W(x_A - x_B \pm \Delta)$;

strengths do not change with time.

Note that we can reinterpret the model as follows. Suppose the winner is determined in the usual way 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 the win probability function W can be interpreted as the distribution function of D, because

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

Note also that by the change of variables $y = e^x$ our "additive" model is equivalent to the "multiplicative" model in which strengths y are nonnegative and

$$\mathbb{P}(\text{A beats B}) = W^*(y_A/y_B); \quad W^*(r) = W(\log r).$$

In particular the additive model with the logistic win probability function

$$\mathbb{P}(\text{A beats B}) = e^{x_A - x_B}/(1 + e^{x_A - x_B}); \quad W(u) = e^u/(1 + e^u) \tag{2}$$

is equivalent to the multiplicative model

$$\mathbb{P}(\text{A beats B}) = y_A/(y_A + y_B)$$

which has an appealing simplicity. However this model makes specific predictions – if $\mathbb{P}(\text{A beats B}) = 2/3$ and $\mathbb{P}(\text{B beats C}) = 2/3$ then $\mathbb{P}(\text{A beats C}) = 4/5$ – and there is no *a priori* reason to believe this relation is empirically true for a given sports league.

1.2 Elo-type rating systems

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 [14] is quite informative about the history and practical implementation. What we 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`)

$$\begin{array}{c} \text{if } i \text{ beats } j \text{ then } y_i \to y_i + \Upsilon(y_i - y_j) \text{ and } y_j \to y_j - \Upsilon(y_i - y_j) \\ \text{if } i \text{ loses to } j \text{ then } y_i \to y_i - \Upsilon(y_j - y_i) \text{ and } y_j \to y_j + \Upsilon(y_j - y_i)\ . \end{array} \tag{3}$$

Note that the sum of all ratings remains constant; it is mathematically natural to center so that this sum equals zero. We require the function $\Upsilon(u),\ -\infty < u < \infty$ to satisfy the qualitative conditions

$$\Upsilon : \mathbb{R} \to (0, \infty) \text{ is continuous, strictly decreasing, and } \lim_{u\to\infty} \Upsilon(u) = 0. \tag{4}$$

We will also impose a quantitative condition

$$\kappa_\Upsilon := \sup_u |\Upsilon'(u)| < 1. \tag{5}$$

To motivate the latter condition, we want the functions $x \to x + \Upsilon(x - y)$ and $x \to x - \Upsilon(y - x)$, the rating updates when a player with (variable) strength x plays a player of fixed strength y, to be an *increasing* function of the starting strength x.

Note that if Υ satisfies (4) then so does $c\Upsilon$ for any scaling factor $c > 0$. So given any Υ satisfying (4) with $\kappa_\Upsilon < \infty$ we can scale to make a function where (5) is satisfied.

Discussion of rating systems. Elaborations of such systems can be, and are, used as real-world ratings, which can be interpreted as estimates of relative skills without any relation to probability models. The reader is likely quite aware of the real-world significance and interest of ratings (for the record we give a brief account in section 9.1). Our *mathematical* goal is to seek to elucidate the connection between Elo-type rating systems and the basic probability model for match results. These are logically different. Even assuming the probability model, for unknown W and strengths x_A, there is no clear connection between the natural (MLE) statistical estimates $\hat{x}_A$ of the strengths and the ratings produced by an Elo-type rating system with some chosen Υ. For instance, the statistical estimates place the same weight on recent-past match results as on distant-past ones, whereas an Elo-type system implicitly places greater weight on recent results. At a practical level they differ in an obvious way. The Elo-type rating scheme (3) is an instance of a *dynamic rating* where a player's rating changes only with the results of the matches played by that player. In contrast, using statistical estimates $\hat{x}_A$ in the probability model as ratings, one should in principle re-compute these ratings for all players after every single match.

Nevertheless there is an intuitively plausible connection, which we will describe and seek to formalize in section 2.

2 The probability model and dynamic rating

2.1 Initial heuristics

We 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 given by the update rule (3) with update function Υ. When player i plays team j, the expectation of the rating change for i equals

$$\Upsilon(y_i - y_j)W(x_i - x_j) - \Upsilon(y_j - y_i)W(x_j - x_i). \tag{6}$$

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

$$\Upsilon(u)/\Upsilon(-u) = W(-u)/W(u), \quad -\infty < u < \infty. \tag{7}$$

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. Recall that in the probability model we can center the strengths so that $\sum_i x_i = 0$, and similarly we will initialize ratings so that $\sum_i y_i = 0$. These observations suggest that 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.

In the context of real world data we could use this suggested effect in either of two directions.

(i) First, we could assume the basic probability model with known W and then choose Υ to satisfy (7) – the simplest choice being $\Upsilon(u) = cW(-u)$ for some scaling factor c. After sufficiently many games the ratings y_i (or some time-average $\bar{y}_i$ used to smooth the random fluctuations from individual matches) should approximate the strengths x_i. These ratings now have some meaningful and testable interpretation: the probability that player i will beat player j is approximately $W(\bar{y}_i - \bar{y}_j)$.

(ii) Alternatively, without assuming any probability model we could just pick a plausible update function Υ and observe the process of ratings, in which players i should have some long-run average ratings $\bar{y}_i$. This indicates an approximate equilibrium in which the expectation of rating changes for each player is approximately zero, so it resembles the basic probability model with W satisfying (7). Now the solution (for W given Υ – see Lemma 1 below) of (7) is $W_\Upsilon(u) = \Upsilon(-u)/(\Upsilon(u) + \Upsilon(-u))$. So as above these ratings now have a meaningful and testable interpretation: the probability that player i will beat player j is approximately $W(\bar{y}_i - \bar{y}_j)$, for this particular W.

The verbal arguments above, which we call the *basic rating heuristic*, are clearly the motivation for using Elo-type update ratings.

Project. `Can you find some discussion of this?`

But it is far from clear how to relate this to any specific mathematical result.

Perhaps the main concern about the heuristic, mathematically speaking, is as follows. Consider realizations of the probability model with an unknown W^*, with ratings defined using a given Υ. Write W_Υ for the function (9) determined by Υ. The long-run average ratings $\bar{y}_i^*$ will depend on W^*. It is quite possible that, for a given pair i, j of players, we have

$$W^*(\bar{y}_i^* - \bar{y}_j^*) \approx H_\Upsilon(x_i - x_j) \tag{8}$$

which is what the folklore conclusion says. But (for generic W^*) (8) cannot be true for all pairs of players. Consider three players i, j, k such that (in the probability model using W^*)

$$\mathbb{P}(i \text{ beats } j) = 0.6, \ \mathbb{P}(j \text{ beats } k) = 0.6$$

and suppose the approximation (8) holds for these two pairs. Then the value of $\mathbb{P}(i \text{ beats } k)$ in the model depends on the function W^*, whereas in our folklore conclusion (8) the value is pre-determined by our choice of Υ – it equals $W_\Upsilon(2u_0)$ where u_0 solves $W_\Upsilon(u_0) = 0.6$.

2.2 Relating W and Υ

Let's start by understanding the solutions of (7), which requires a little care because W has the constraint that $W(u) + W(-u) = 1$ while Υ has no corresponding constraint. Taking the required care, we easily find

Lemma 1. *(a) If Υ satisfies the qualitative constraints (4) then*

$$W(u) = \Upsilon(-u)/(\Upsilon(u) + \Upsilon(-u)) \tag{9}$$

is the only solution of the equation (7) which satisfies the qualitative constraints (1).
(b) If W satisfies the qualitative constraints (1) then, for any function ϕ,

$$\Upsilon(u) = W(-u)\phi(|u|) \tag{10}$$

satisfies the equation (7), and these are the only solutions of (7).

Note the implication of (b) is as follows. Given W satisfying (7), to find the general Υ satisfying the requirements (4,5) we first take Υ of form (10) for some continuous ϕ, but we still need to check the remaining conditions of (4,5). In the special case where we take

$$\Upsilon(u) = cW(-u)$$

for some $c > 0$, we only need to check $c < 1/\kappa_W$.

Returning to the heuristic argument at (6), when $x_i = y_i \ \forall i$ the variance of the increment of i's rating when i plays j is

$$\Upsilon^2(u)W(u) + \Upsilon^2(-u)W(-u), \quad u = x_i - x_j.$$

Given W, as an alternative to the choice $\Upsilon(u) = cW(-u)$ of solution of form (10) one could make the choice that makes the variance be constant, that is

$$\Upsilon(u) = c\sqrt{W(-u)/W(u)}.$$

This choice seems "statistically efficient" in some heuristic sense, but combined with the usual choice of logistic win probability function W leads to a "clash of heuristics". In a match where player strengths differ by a large amount u but the weaker player wins, the update is order $\exp(u/2)$, which for large u is much larger than the original difference and so clearly ridiculous. In fact if we want to use the constant-variance solution and ensure our quantitative property (5) for Υ, we must use a win probability function such that $W(-u)$ decreases no faster than u^{-2} as $u \to \infty$. From this constant-variance viewpoint, taking W as the Cauchy distribution function seems quite reasonable, in that we than have $\Upsilon(-u)$ of order $u^{1/2}$.

2.3 Stationary distributions for the update process

The basic probability model just says what happens when two teams play, but does not specify how the matches are scheduled, which as mentioned earlier varies widely amongst real-world sports. We will study the mathematically simplest, albeit not so realistic, "random matching" scheme in which there are n players and for each match a pair of players is chosen uniformly at random. With the basic probability model for match results, and using the rating system (3), we obtain a continuous-state Markov chain $\mathbf{Y}(t) = (Y_i(t), 1 \leq i \leq n), t = 0, 1, 2, \ldots$, where $Y_i(t)$ is the rating of player i after a total of t matches have been played. We call this the *update process*. Note that this process is parametrized by the functions W and Υ, and by the vector $\mathbf{x} = (x_i, 1 \leq i \leq n)$ of player strengths. We center player strengths and rankings: $\sum_i x_i = 0$ and $\sum_i Y_i(0) = 0$.

Theorem 2. *Under our standing assumptions (1, 4, 5) on W and Υ, for each $\mathbf{x}$ the update process has a unique stationary distribution $\mathbf{Y}(\infty)$, and for any initial ratings $\mathbf{y}(0)$ we have $\mathbf{Y}(t) \to_d \mathbf{Y}(\infty)$ as $t \to \infty$.*

We prove this by standard methods in section 3. Note here we are not assuming the specific relation (7) between W and Υ. Note also that given non-random initial rankings $\mathbf{y}(0)$ the distribution of $\mathbf{Y}(t)$ has finite support for each t, so we cannot have convergence in variation distance.

Problem. `In the final comment above, I'm assuming the limit distribution is continuous, which seems obvious, but I haven't tried to prove it.`

For the record the transition probabilities for the update process are as follows. From $\mathbf{y}$, for each pair $\{i, j\}$

$$\begin{array}{c} y_i \to y_i + \Upsilon(y_i - y_j) \text{ and } y_j \to y_j - \Upsilon(y_i - y_j), \text{ probability } W(x_i - x_j)/\binom{n}{2} \\ y_i \to y_i - \Upsilon(y_j - y_i) \text{ and } y_j \to y_j + \Upsilon(y_j - y_i), \text{ probability } W(x_j - x_i)/\binom{n}{2} \ . \end{array} \tag{11}$$

3 A convergence theorem

3.1 Background: Markov chain convergence on $\mathbb{R}^n$

Here we give a general result (Proposition 3) which is easily proved because it uses strong assumptions; then we will discuss related literature. Let $\mathbf{Y}(t), t = 0, 1, 2, \ldots$ be a Markov chain on $\mathbb{R}^n$, and write $\mathbb{P}_{\mathbf{y}}(\cdot), \mathbb{E}_{\mathbf{y}}(\cdot)$ to indicate initial state $\mathbf{y}$. Write $||\mathbf{y}|| = \sum_i |y_i|$. We will require four properties.

(i) The Feller property. The map $\mathbf{y} \to \mathbb{P}_{\mathbf{y}}(Y(1) \in \cdot)$ is continuous w.r.t. weak convergence.

(ii) The locally bounded jumps property. For each $r_0 < \infty$ there exists $r' < \infty$ such that, for all $\mathbf{y}$ with $||\mathbf{y}|| \le r_0$, we have $\mathbb{P}_{\mathbf{y}}(||Y(1)|| \le r') = 1$.

Note that each of the two properties above holds in the setting of Theorem 2, by continuity of the functions H and Υ.

(iii) The Lyapounov property. There exists a continuous function $\Phi : \mathbb{R}^n \to [0, \infty)$ such that $\Phi(\mathbf{y}) \to \infty$ as $||\mathbf{y}|| \to \infty$, and such that

$$\mathbb{E}_{\mathbf{y}}\Phi(Y(1)) - \Phi(\mathbf{y}) \to -\infty \text{ as } ||\mathbf{y}|| \to \infty.$$

(iv) The coupling property. Given any two initial states $\mathbf{y}', \mathbf{y}''$ we can construct a process $(\mathbf{Y}'(t), \mathbf{Y}''(t)), t = 0, 1, 2, \ldots$ such that the two marginal processes are distributed as the two Markov process started from each given state, and $||\mathbf{Y}'(t) - \mathbf{Y}''(t)|| \to 0$ in probability as $t \to \infty$.

Proposition 3. *A chain with properties (i - iv) has a unique stationary distribution* $\mathbf{Y}(\infty)$*, and from any initial distribution we have* $\mathbf{Y}(t) \to_d \mathbf{Y}(\infty)$ *as* $t \to \infty$*.*

Proof. By the Lyapounov property

$$\mathbb{E}_{\mathbf{y}}\Phi(Y(1)) \le \Phi(\mathbf{y}) - b(\Phi(\mathbf{y})) \tag{12}$$

for some function $b(\phi) \uparrow \infty$ as $\phi \to \infty$. By the locally bounded jumps property

$$\sup\{\mathbb{E}_{\mathbf{y}}\Phi(Y(1)) \ : \ \Phi(\mathbf{y}) \le \phi\} < \infty \text{ for each } \phi \ge 0$$

from which it follows that the inequality (12) holds for all $\mathbf{y}$ for some increasing function $b : [0, \infty) \to (-\infty, \infty)$ such that $b(0) > -\infty$ and $b(\phi) \uparrow \infty$

as $\phi \to \infty$. But (12) implies that the process

$$S(t) := \Phi(Y(t)) + \sum_{s=0}^{t-1} b(\Phi(Y(s)))$$

is a supermartingale, and hence $\mathbb{E}_{\mathbf{y}} S(t) \leq \Phi(\mathbf{y})$ for all t. Consider ϕ with $b(\phi) > 0$. We have

$$S(t) \geq b(\phi) \sum_{s=0}^{t-1} 1_{\{\Phi(\mathbf{Y}(s)) \geq \phi\}} \ + b(0)t$$

and so

$$\mathbb{E}_{\mathbf{y}} \left[\tfrac{1}{t} \sum_{s=0}^{t-1} 1_{\{\Phi(\mathbf{Y}(s)) \geq \phi\}} \right] \leq \tfrac{t^{-1}\Phi(\mathbf{y}) - b(0)}{b(\phi)}.$$

Taking U_t uniform on $\{0, 1, \ldots, t-1\}$, this says

$$\mathbb{P}_{\mathbf{y}}(\Phi(\mathbf{Y}(U_t)) \geq \phi) \leq \tfrac{t^{-1}\Phi(\mathbf{y}) - b(0)}{b(\phi)}.$$

This implies that the sequence $\Phi(\mathbf{Y}(U_t)),\ t = 1, 2, \ldots$ is tight, and so the sequence $\mathbf{Y}(U_t),\ t = 1, 2, \ldots$ is tight. Take any subsequential weak limit $\mathbf{Y}(\infty)$; by the Feller property, $\mathbf{Y}(\infty)$ is a stationary distribution. Then the coupling property (applied with initial $\mathbf{y}''$ chosen from the distribution $\mathbf{Y}(\infty)$) establishes the convergence assertion of the Proposition, and uniqueness holds by considering initial $\mathbf{y}'$ chosen from another stationary distribution. □

Discussion. (a) The standard account of discrete-time continuous-space Markov chain theory is [8], but that monograph studies the stronger notion of convergence in variation distance, which cannot hold here.

(b) We do not know any broad "applied probability" treatment of convergence in distribution for Feller chains. A particular version of the general "coupling from the past" method is surveyed in [2].

(c) Results like Proposition 3 have been known for a long time, but it is not easy to discern in the literature the straightforward proof given above. Assuming both the Lyapounov property and the coupling property is overkill, in that (heuristically) we only need one. That is, given the Lyapounov property we only need some extra weak "irreducibility" property, or given the coupling property we only need a weaker tightness property.

(d) Proposition 3 was stated in the setting of state space $\mathbb{R}^d$, but extends naturally to our setting where the state space is $\{(x_1, \ldots, x_n) : \ \sum_i x_i = 0\}$.

3.2 Proof of Theorem 2

As noted in section 3.1, we need only verify the Lyapounov property and the coupling property. Write $\Phi(\mathbf{y}) = \sum_i y_i^2$. The evolution rule (11) for the update process shows that, from an arbitrary configuration $\mathbf{y}$, conditional on the first match involving the pair $\{i, j\}$, the conditional expectation of $\Phi(\mathbf{Y}(1)) - \Phi(\mathbf{y})$ equals

$$W(x_i - x_j) \left[\ (y_i + \Upsilon(y_i - y_j))^2 - y_i^2 + (y_j - \Upsilon(y_i - y_j))^2 - y_j^2 \right]$$

$$+W(x_j - x_i) \left[\ (y_i - \Upsilon(y_j - y_i))^2 - y_i^2 + (y_j + \Upsilon(y_j - y_i))^2 - y_j^2 \right]$$

Set $W(u) = \frac{1}{2} + V(u)$, so $W(-u) = \frac{1}{2} - V(u)$, and set $d_{ij} = x_i - x_j$, and expand separately the "$\frac{1}{2}$" and the "$\pm V(u)$ terms. The former becomes

$$(y_i - y_j)\Upsilon(y_i - y_j) + (y_j - y_i)\Upsilon(y_j - y_i) + \Upsilon^2(y_i - y_j) + \Upsilon^2(y_j - y_i)$$

$$= A(y_i, y_j) \text{ say}$$

and the latter becomes

$$2V(d_{ij}) \left[(y_i - y_j)\Upsilon(y_i - y_j) - (y_j - y_i)\Upsilon(y_j - y_i) + \Upsilon^2(y_i - y_j) - \Upsilon^2(y_j - y_i)\right]$$

$$= 2V(d_{ij})B(y_i, y_j) \text{ say.}$$

So

$$\mathbb{E}_{\mathbf{y}}\Phi(Y(1)) - \Phi(\mathbf{y}) = \frac{1}{\binom{n}{2}} \sum\sum_{\{i,j\}} (A(y_i, y_j) + 2V(d_{ij})B(y_i, y_j)).$$

We want to upper bound the right side as $||\mathbf{y}|| \to \infty$. By a compactness argument we may assume $\mathbf{y}/||\mathbf{y}|| \to \mathbf{r} = (r_1, \ldots, r_n)$ where the sets

$$I_+ := \{i : r_i > 0\} \text{ and } I_- := \{i : r_i < 0\}$$

are non-empty and we order as $r_1 \leq r_2 \leq \ldots \leq r_n$. First consider the case where $\Upsilon(-\infty) := \lim_{u \to -\infty} \Upsilon(u)$ is finite. Here

$$\mathbb{E}_{\mathbf{y}}\Phi(Y(1)) - \Phi(\mathbf{y}) \sim \frac{1}{\binom{n}{2}} ||\mathbf{y}|| S(\mathbf{r}) \tag{13}$$

where

$$S(\mathbf{r}) = \sum\sum_{i<j} (r_i - r_j)\Upsilon(-\infty) + \sum\sum_{i<j} 2V(d_{ij})(r_i - r_j)\Upsilon(-\infty).$$

But $1 + 2V(u) = 2W(u)$ and so

$$S(\mathbf{r}) = 2\Upsilon(-\infty) \sum\sum_{i<j} W(d_{ij})(r_i - r_j) < 0.$$

This and (13) establish the Lyapounov property in the case $\Upsilon(-\infty) < \infty$. The case $\Upsilon(-\infty) = \infty$ holds by a similar argument using the fact

$$u\Upsilon(u) + \Upsilon^2(u) \leq (u(1 - \kappa_\Upsilon) + o(1))\Upsilon(u) \text{ as } u \to -\infty.$$

To verify the coupling property, we first show that condition (5) implies

$$(x - y)(\Upsilon(x) - \Upsilon(y)) + (\Upsilon(x) - \Upsilon(y))^2 < 0, \quad x \neq y. \tag{14}$$

First consider $x > y$: condition (5) implies

$$(x - y) + (\Upsilon(x) - \Upsilon(y)) > 0$$

and (14) follows because Υ is decreasing. The analogous argument gives the same conclusion for $x < y$. Next note that applying (14) to $-x$ and $-y$ gives

$$-(x - y)(\Upsilon(-x) - \Upsilon(-y)) + (\Upsilon(-x) - \Upsilon(-y))^2 < 0, \quad x \neq y. \tag{15}$$

In the update process, when team i meets team j there are two alternatives
(a) $y_i \to y_i + \Upsilon(d_{ij})$ and $y_j \to y_j - \Upsilon(d_{ij})$: probability $W(x_i - x_j)$
or (b) $y_i \to y_i - \Upsilon(-d_{ij})$ and $y_j \to y_j + \Upsilon(-d_{ij})$: probability $W(x_j - x_i)$
where $d_{ij} = y_i - y_j$. Consider the natural coupling of update processes with initial configurations $\mathbf{y}'$ and $\mathbf{y}''$; that is, for each match the winner is the same team in each process. So in the coupled process, when team i meets team j there are two alternatives, whose effect can be written as
(a) $(y'_i - y''_i) \to (y'_i - y''_i) + (\Upsilon(d'_{ij}) - \Upsilon(d''_{ij}))$
and $(y'_j - y''_j) \to (y'_j - y''_j) - (\Upsilon(d'_{ij}) - \Upsilon(d''_{ij}))$: probability $W(x_i - x_j)$
or (b) $(y'_i - y''_i) \to (y'_i - y''_i) - (\Upsilon(-d'_{ij}) - \Upsilon(-d''_{ij}))$
and $(y'_j - y''_j) \to (y'_j - y''_j) + (\Upsilon(-d'_{ij}) - \Upsilon(-d''_{ij}))$: probability $W(x_j - x_i)$.
So the conditional expected update of $s^2 := \Phi(\mathbf{y}' - \mathbf{y}'')$, given team i plays team j, is

$$s^2 \to s^2 + 2(\Upsilon(d'_{ij}) - \Upsilon(d''_{ij}))^2 W(x_i - x_j) + 2(\Upsilon(-d'_{ij}) - \Upsilon(-d''_{ij}))^2 W(x_j - x_i)$$

$$+2(d'_{ij} - d''_{ij})[(\Upsilon(d'_{ij}) - \Upsilon(d''_{ij}))W(x_i - x_j) \; - \; (\Upsilon(-d'_{ij}) - \Upsilon(-d''_{ij}))W(x_j - x_i)]$$

Now (14,15) show this increment is strictly negative unless $d'_{ij} = d''_{ij}$. So for the coupled process,

$$\mathbb{E}_{\mathbf{y}',\mathbf{y}''}\Phi(\mathbf{Y}'(1) - \mathbf{Y}''(1)) \leq \Phi(\mathbf{y}' - \mathbf{y}'')$$

and the inequality is strict provided $\mathbf{y}' \neq \mathbf{y}''$. This says that $\Phi(\mathbf{Y}'(t) - \mathbf{Y}''(t))$ is a nonnegative supermartingale, which therefore converges a.s. to some limit $\Phi_\infty \geq 0$. We want to prove $\Phi_\infty = 0$ a.s. Because we have already established the Lyapounov property (this argument is a bit of a hack, but avoids quantifying the strictness of the inequality above), the argument in the proof of Proposition 3 shows there is some subsequential weak limit $(\mathbf{Y}'(\infty), \mathbf{Y}''(\infty))$ of $(\mathbf{Y}'(U_t), \mathbf{Y}''(U_t))$ which must be a stationary distribution for the coupled process. Such a stationary distribution must have $\mathbf{Y}'(\infty) = \mathbf{Y}''(\infty)$ a.s., so $\Phi(\mathbf{Y}'(U_t) - \mathbf{Y}''(U_t)) \to \Phi(\mathbf{Y}'(\infty) - \mathbf{Y}''(\infty)) = 0$ in distribution. But by supermartingale convergence $\Phi(\mathbf{Y}'(U_t) - \mathbf{Y}''(U_t)) \to \Phi_\infty$ a.s., so indeed $\Phi_\infty = 0$ a.s.

Discussion. The theorem was stated in the context of the ‘‘random match" schedule (choose two teams at random to play the next match) which fits most simply into the Markov chain framework. But in fact the key inequalities in the proof remain true conditionally on which two teams are chosen to play the next match. So the arguments can be applied more generally. For instance if we choose some deterministic season schedule, and then repeat that schedule each season, the process of ratings at the end of successive seasons will be a Markov chain, and the arguments in the proof above will show that the distribution of this chain will converge to some limit distribution. As another instance, if we schedule matches in some arbitrary way not depending on ratings, then the coupling argument shows that the influence of the initial ratings will asymptotically disappear. However if we schedule in a way depending on ratings then the simple coupling argument breaks down.

3.3 Future work

Theorem 2 shows that the update process has a stationary distribution $\mathbf{Y} = (Y_i)$ which depends of the triple $(W, \Upsilon, \mathbf{x})$. As outlined in section 2.1, intuition suggests that the equilibrium rating Y_i should be close to the strength x_i. Trying to quantify this idea is a main goal of this project. In principle the arguments in the proof of Theorem 2 give explicit bounds but those are undoubtedly very crude.

Below are some other ways to think about this issue.

3.3.1 The dynamical system limit

Given a pair (W, Υ) satisfying our standing assumptions, we can study the update process using the scaled update function $c\Upsilon$. Heuristically, as $c \downarrow 0$ this process should, after time-rescaling, converge to the deterministic dynamical system

$$\mathbf{y}'(t) = \Gamma_{\mathbf{x}}(\mathbf{y}(t))$$

governed by the expectations in (11), that is

$$y_i' = \sum_j (\Upsilon(y_i - y_j)W(x_i - x_j) - \Upsilon(y_j - y_i)W(x_j - x_i)).$$

It seems intuitively clear that the arguments in the proof of Theorem 2 can be reused in this deterministic setting to prove

$$\text{for all } \mathbf{x}\text{, there is a unique fixed point } \mathbf{y} = \mathbf{y}(\mathbf{x}) \text{ of the operator } \Gamma_{\mathbf{x}}. \quad (16)$$

Problem. `Prove (16).`

Of course when (7) holds we have the fixed point $\mathbf{y} = \mathbf{x}$ but here we are interested in the general case. In this problem we would also want to have convergence to the fixed point from arbitrary start.

Conceptually, this starts to address the "deterministic error" caused by not knowing W. That is, if we use $\Upsilon_0(u) = cW_0(-u)$ for some "guessed" function W_0, then our update procedure gives us (in the dynamical system limit) some ratings $\mathbf{y}(\mathbf{x})$ which depend on the unknown true W; we would like to get some bound on the error $\mathbf{y}(\mathbf{x}) - \mathbf{x}$ in terms of some measure of the distance between W and W_0.

Problem. `Find such a bound.`

3.3.2 The Ornstein-Uhlenbeck approximation

To avoid the issue above let's consider the case $\Upsilon(u) = cW(-u)$. The second-order behavior associated with the dynamic process limit will be an Ornstein-Uhlenbeck approximation, as outlined below. Fix small c. Consider ratings $\mathbf{y}$ with $\mathbf{y} - \mathbf{x}$ small. When i plays j, the increment Δ_{ij} of i's rating is, to second order,

$$\Upsilon(y_i - y_j) \approx \Upsilon(x_i - x_j) + ((y_i - x_i) - (y_j - x_j))\Upsilon'(x_i - x_j)$$

or

$$-\Upsilon(y_j - y_i) \approx -\Upsilon(x_j - x_i) - ((y_j - x_j) - (y_i - x_i))\Upsilon'(x_j - x_i)$$

with probabilities $W(x_i - x_j)$ and $W(x_j - x_i)$ respectively. Because $\Upsilon(u) = cW(-u)$ and $W'(u) = W'(-u)$ the expectation reduces to

$$\mathbb{E}\Delta_{ij} \approx -c((y_i - x_i) - (y_j - x_j))W'(x_i - x_j)$$

and $\mathrm{var}\Delta_{ij} \approx c^2 W(x_i - x_j)W(x_j - x_i)$. So when i plays a random opponent the rating increment Δ_i is such that

$$\mathbb{E}\Delta_i \approx -c\sum_{j\neq i} \tfrac{1}{n-1}((y_i - x_i) - (y_j - x_j))W'(x_i - x_j)$$

$$\mathrm{var}\Delta_i \approx c^2 \sum_{j\neq i} \tfrac{1}{n-1} W(x_i - x_j)W(x_j - x_i).$$

Note also that when i plays j we have

$$\mathbb{E}[\Delta_i\Delta_j] \approx -\mathrm{var}\Delta_{ij} \approx -c^2 W(x_i - x_j)W(x_j - x_i).$$

In the update process $\mathbf{Y}(t)$ the parameter t was total number of matches played, We now rescale to a continuous-time process $\mathbf{Z}^{(c)}(t)$ defined by

$$Z_i^{(c)}(t) = c^{-1/2}\ \left(Y_i\left(\lfloor \frac{nt}{2c} \rfloor\right) - x_i\right)$$

Now each team plays an average of $1/c$ matches per rescaled time unit.

It is now intuitively clear that $\mathbf{Z}^{(c)}$ approximates the multidimensional Ornstein-Uhlenbeck process $\mathbf{Z}$ defined by

$$dZ_i(t) = -\tfrac{1}{n-1}\sum_{j\neq i}(Z_i(t) - Z_j(t))W'(x_i - x_j)\ dt\ +\ dB_i(t)$$

where $\mathbf{B}(t) = (B_i(t))$ is the Brownian motion associated with the mean-zero Gaussian distribution with covariance matrix

$$\mathrm{var}B_i(1) = \sum_{j\neq i} \tfrac{1}{n-1} W(x_i - x_j)W(x_j - x_i)$$

$$\mathrm{cov}(B_i(1), B_j(1)) = -\tfrac{1}{n-1} W(x_i - x_j)W(x_j - x_i), \quad j \neq i.$$

Project. A not very exciting project is to think about
details of proof of O-U limit
checking the matrix condition for stability
bounding the variance in the stationary distribution
the analog in the general setting (16).

4 Tracking changing strengths

The setting envisaged in section 3 – an increasing number of matches between players of unchanging strengths – is not so realistic, and if one believed the basic probability model then the "statistical" method of calculating MLEs of strengths from the complete set of match results would surely be more accurate than any Elo-style updating scheme. In this section we consider the more realistic setting where strengths may change. One could use continue the "statistical" approach by making parametric models of how strengths change with time. For instance [4] use a model where strengths $x_i(t)$ have a "smoothing prior" under which the changes $x_i(t+1) - x_i(t)$ are Normal$(0, \sigma^2)$. But it seems conceptually difficult to find plausible models, and many parameters would be needed. In contrast, the (heuristic) point of Elo-type updates is that they are expected to be good at tracking changing strengths.

To give a mathematical analysis we must not only model how strengths change but also specify more precisely what ratings are intended to do. There is an (intuitively) obvious trade-off in the choice of scaling constant c used for updating: a larger c will track changes more quickly but cause larger random variability. How we make this trade-off will depend on whether we are more interested in tracking changes of strengths or in accurate assessment of unchanging strengths.

4.1 An O-U model for changing strengths in league play

`Description/argument below very sketchy`

Consider n teams in ongoing league play. Take the unit of time as $n-1$ matches per team. Model strengths $X_i(t)$ as discretely-sampled Ornstein-Uhlenbeck process

$$dX_i(t) = -\mu X_i(t)\ dt + \sigma\sqrt{2\mu}\ dB_i(t)$$

for parameters $\mu, \sigma > 0$ (we need to center, but for large n this has small effect – we will ignore for now). So X_i has stationary distribution Normal$(0, \sigma^2)$ and correlations $\mathrm{cor}(X_i(0), X_i(t)) = e^{-\mu t}$. Assume σ is order 1, so that teams have comparable but different strengths.

We study ratings $Y_i(t)$ in the setting of section 3.3.2, with c small. Let us assume there is an approximate joint Normal mean-zero stationary distribution for (X_i, Y_i) and calculate its covariance structure.

`There is a do-able calculation which I haven't had the patience to actually do, but below is what I think is the right order-of-magnitude` result.

As in section 3.3.2, the increment of $Y_i(t)$ for a match by i has mean of order $-c(Y_i - X_i)$ and variance of order c^2, and the resulting variance of the stationary distribution of $Y_i - X_i$, assuming Y_i does track X_i closely, will be order c. Our time unit is now $n-1$ matches for i, so the drift rate for $Y_i - X_i$ is order cn; we want this to be a multiple of μ, the drift rate of X_i, in order that Y_i can track X_i closely.

So this suggests we should take c as a multiple of μ/n, and then the typical error $Y_i - X_i$ will be order $c^{1/2} = (\mu/n)^{1/2}$.

4.2 Two component rating models

Intuition suggests that Elo-type systems – which rely on updating only a single numerical rating – are in fact rather inefficient as rating schemes, because of the unavoidable trade-off mentioned above. As a natural alternative one can consider schemes which use two numbers to summarize past performance. In such a scheme one needs both an update rule and also a rule for predicting win probabilities.

One such scheme is used in the TrueSkill ranking system [3] on Xbox Live. 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 one first computes the conditional distribution of X_i given $X_i > X_j$, where X_i has Normal(μ_i, σ_i^2) distribution, and then updates 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 $X_i < X_j$. So when i is about to play j we are implicitly predicting the probability that i wins as $\mathbb{P}(X_i > X_j)$.

Project. This scheme seems a bit strange from a theory viewpoint, as follows, so a project is to study it more carefully (and check my description of the scheme is reasonable). Consider the case of unchanging strengths. In this scheme the σ's will typically decrease after each match (see Kenneth's write-up). So in the long run we get to a situation with very small σ's and where the update for the μ's is, for $\mu_i > \mu_j$

if i wins then $\mu_i \to \mu_i + \varepsilon$, $\mu_j \to \mu_j - \varepsilon'$ for some small $\varepsilon, \varepsilon'$

if j wins then each of μ_i and μ_j is updated to (almost) the same intermediate value.

This seems like some non-very-plausible Elo scheme.

An alternative scheme is as follows. Use the Elo rating as one component – write $Y_i(s)$ for i's Elo rating after i has played s matches. Now introduce a parameter $\lambda < 1$ and take the other component $Z_i(s)$ as the discounted past average of the $Y_i(\cdot)$, that is we update as

$$Z_i(s) = \lambda Z_i(s-1) + (1-\lambda) Y_i(s).$$

The intuition for this scheme is as follows, via a comparison with single-number Elo for some given $c\Upsilon$. Use for Y_i Elo with $c'\Upsilon$ for $c' > c$. Then choose λ such that the size of increments of Z_i are small than the comparison single-number Elo update sizes. We then use Z_i as the "rating" to predict win probabilities; this scheme is intend to reduce the effect of recent randomness of match results.

Dan's simulations show this in fact doesn't work well on baseball data. Likely hero reason is that there's not enough non-linearity is the basic Elo process.

Project. There are many projects here. Repeat theory analysis for unchanging strengths in style of section 3, or for changing strengths in style of section 4.1. Both for the TrueSkill model and our model above. Do simulation studies of these models.

5 Numbers of games won by different teams

Figure xxx shows number of matches won by Major League Baseball teams over the years xxx.

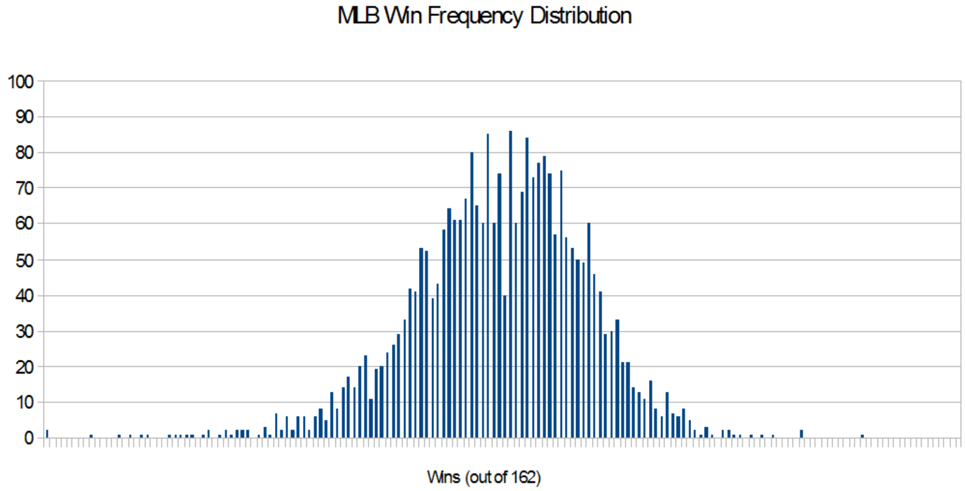

Figure xxx. thanks Dan Lanoue

A very basic mathematical question is the following. In league play, what are the possible combinations of numbers of wins for the teams (after all, two different teams cannot both win all their matches)?

5.1 Arbitrary probability model

Project. `It is likely that some or all of this is already known. Need literature search for results related to Proposition 4 and the conjecture below (20).`

Consider a league of n teams; each pair of teams plays once, with a definite win-lose result. Write "season results" for the collection of all game results. Write I for a uniform random team, and write U_n for the uniform distribution on $\{0, 1, 2, \ldots, n-1\}$.

Proposition 4. *Consider an arbitrary probability distribution over season results, and write*

$$q(i) = \mathbb{E}(\text{number of wins for team } i). \tag{17}$$

Then

$$q(I) \preceq U_n \tag{18}$$

in the sense of convex order. Conversely, given a function q from the set of teams to $[0, n-1]$ such that (18) holds, there exists a probability distribution over season results such that (17) holds,

Here we use the notion of *convex order* on distributions (see e.g. [10] section 2.A): $\mathrm{dist}(Y) \preceq \mathrm{dist}(U)$, which we write less formally as $Y \preceq U$, means

$$\mathbb{E}\phi(Y) \leq \mathbb{E}\phi(U) \ \text{ for all convex } \phi$$

or equivalently

$$Y =_d \mathbb{E}(U|Z) \text{ for some } Z. \tag{19}$$

Consider now the case of deterministic season results – what we actually see at the end of a season. In this case

$$q^*(i) = \text{number of wins for team } i \tag{20}$$

is integer-valued, and (18) says $q^*(I) \preceq U_n$. Conversely, given integer-valued Y such that $Y \preceq U_n$, we **conjecture** that there are deterministic season results such that $q^*(I) \overset{d}{=} Y$. But Proposition 4 gives only the weaker result that there are *random* season results such that $q(I) =_d Y$. The answer to this conjecture is likely already known – see section 5.2.

Proof. First consider the deterministic case. Fix a convex function ϕ. Order teams so that $0 \leq q^*(1) \leq q^*(2) \leq \ldots \leq q^*(n)$. Suppose $q^*(1) \geq 1$. Change the results of the games won by team 1, one game at a time, to make them a loss for team 1. At each such step $q^*(1)$ decreases by 1 and some $q^*(i)$ increases by 1, and by convexity the value of $\mathbb{E}\phi(q^*(I))$ can only increase. Continue until reaching a configuration with $q^*(1) = 0$. Now suppose $q^*(2) \geq 2$. Again change the results of the games won by team 2 against teams $i > 2$, one game at a time, to make them a loss for team 2. Again this can only increase $\mathbb{E}\phi(q^*(I))$. Continue until reaching a configuration with $q^*(1) = 0$ and $q^*(2) = 1$. Eventually we reach the configuration with $q^*(i) = i - 1, 1 \leq i \leq n$. So the original configuration satisfies $\mathbb{E}\phi(q^*(I)) \leq \mathbb{E}\phi(U_n)$, establishing (18) in the deterministic case.

In the random case write Q_i for number of wins by team i. So $q(i) = \mathbb{E}Q_i$, and by Jensen's inequality $\phi(q(i)) \leq \mathbb{E}\phi(Q_i)$. So

$$\mathbb{E}\phi(q(I)) = \tfrac{1}{n}\sum_i \phi(q(i)) \leq \tfrac{1}{n}\sum_i \mathbb{E}\phi(Q_i) = \mathbb{E}\left[\tfrac{1}{n}\sum_i \phi(Q_i)\right]$$

and the quantity in brackets is bounded by $\mathbb{E}\phi(U_n)$, by the deterministic case. This establishes (18) in general.

For the converse, consider permutations $\pi : \{1,2,\ldots,n\} \to \{1,2,\ldots,n\}$. Such a permutation defines a special kind of "season results", in which team i loses to team j if and only if $\pi(i) < \pi(j)$, and therefore team i wins exactly $\pi(i) - 1$ games. By considering this special type of "totally ordered" season result, the converse will follow from the lemma below. □

Lemma 5. *Let $q : \{1,2,\ldots,n\} \to [1,n]$ be a function such that $q(1+U_n) \preceq 1+U_n$. Then there exists a probability distribution over permutations π such that*

$$\mathbb{E}\pi(i) = q(i), 1 \le i \le n.$$

Proof. Write I_n and J_n for RVs with the uniform distribution on $\{1,2,\ldots,n\}$. The relation $q(I_n) \preceq J_n$ is equivalent, using (19), to saying that we can construct a joint distribution for (I_n, J_n) such that

$$\mathbb{E}(J_n | I_n = i) = q(i),\ 1 \le i \le n.$$

Now the matrix with entries

$$p_{ij} := n\mathbb{P}(I_n = i, J_n = j)$$

is doubly stochastic, so by Birkhoff's theorem it is a mixture of permutation matrices. In other words, there is a probability distribution over permutations π such that

$$p_{ij} = \mathbb{P}(\pi(i) = j).$$

But this just says

$$\mathbb{P}(\pi(i) = j) = \mathbb{P}(J_n = j | I_n = i)$$

and so

$$\mathbb{E}\pi(i) = \mathbb{E}(J_n | I_n = i) = q(i).$$

□

Remark. Obviously a general "season results" is not a mixture over the special "totally ordered" results. However, implicit in our result and its proof is the fact that, if a given function $q(\cdot)$ arises as at (20) from some arbitrary "season results", then it also arises as at (17) from some mixture of "totally ordered" results. One can in fact prove this directly via the following argument, somewhat analogous to the "probabilistic" (martingale) proof of Birkhoff's theorem. `Cycle-flipping argument omitted`.

5.2 Tournament graphs

In graph theory a *tournament* graph is defined to be the complete graph on n vertices with each edge directed in one direction. This topic has attracted a large literature (*MathSciNet* shows over 1,400 papers with *tournament* in their title, most concerning this or related concepts), and an introductory account is found in the 1968 monograph [9]. Our notion of "season results in league play" is mathematically equivalent to "tournament graph" (a directed edge ij means i beat j) but most results about tournament graphs do not have natural interpretations in terms of sports results. However, our "number of wins" quantity $(q^*(i), 1 \leq i \leq n)$ is simply the out-degrees of a tournament graph, so our conjecture below (20) may have been answered in the graph theory literature.

5.3 Win proportions in the basic probability model

We now consider the same question in the context of the basic probability model. That is, there is a league of n teams, each pair of teams plays once, with results determined by the model, for some W and $x_1, \ldots, x_n$, and

$$q(i) = \mathbb{E}(\text{number of wins for team } i)..$$

As a special case of Proposition 4 we have

$$q(I) \preceq U_n. \tag{21}$$

where as before I denotes a uniform random team, and U_n denotes the uniform distribution on $\{0, 1, 2, \ldots, n-1\}$. But it is not clear whether or not the converse is true.

Open Problem 6. *Given $(q(i), 1 \leq i \leq n)$ satisfying (21), can we always find W and $x_1, \ldots, x_n$ such that*

$$q(i) = \sum_{j \neq i} W(x_i - x_j), \ 1 \leq i \leq n. \tag{22}$$

(There is a small detail that the extreme case $q(i) \equiv i - 1$ does not have this representation. A precise version of the problem is: is the set of $(q(i))$ of form (22) dense in the set of form (21) ?)

Below is discussion related to this problem. Given W we have $n-1$ equations for $n-1$ unknowns (because of the sum constraints), so at first sight it is plausible that the set of equations should have a solution for typical

W. The simplest W to study is the linear (uniform distribution function) case

$$W(u) = \tfrac{1}{2}(1+u), \ -1 \le u \le 1.$$

Equations (22) become

$$\begin{aligned} q(i) &= \sum_{j \neq i} \tfrac{1}{2}(1 + x_i - x_j) \\ &= \tfrac{1}{2}((n-1) + n x_i) \text{ because } \sum_j x_j = 0 \end{aligned}$$

and the solution is

$$x_i = \tfrac{2q(i)-(n-1)}{n}.$$

However, the model with this W only makes sense when $\max_i x_i - \min_i x_i \le 1$. So the functions q that can arise from this W are exactly the functions with

$$\max_i q(i) - \min_i q(i) \le n/2$$

as well as the deterministic constraint $\sum_i q(i) = n(n-1)/2$.

Project. `So one next step in studying the problem above would be to try to solve equations (22) for some other` W`, for example the logistic.`

5.4 A mean-field version

Here is an "$n \to \infty$ limit" version of the problem above. Instead of a list of "unknown" strengths $x_1, \ldots, x_n$ of n teams we imagine a very large number of teams whose strengths follow some (unknown) probability density function $g(x)$, with corresponding distribution function G. And we imagine each team plays each other team, with sufficiently many games that we can ignore "finite-sample random effects". In this setting the function

$$f(x) := \text{ proportion of games won by a team with strength } x$$

is given by the convolution formula

$$f(x) = \int W(x-y) g(y) \ dy.$$

Writing X for the strength of a uniform random team, so that X has density function $g(x)$, then the distribution of $f(X)$ corresponds to the distribution of proportions of games won by the different teams.

The conceptual point is that G and W are not directly observable; what is easily observable is the proportions of games won by the different teams – our $f(X)$ in this mean-field world. Note that in the real-world finite setting the randomness would make the distribution of "proportions of games won by the different teams" more spread out.

Because $x \to f(x)$ is increasing, $\mathbb{P}(f(X) \leq f(x)) = \mathbb{P}(X \leq x) = G(x)$, so the distribution function of our "observable" is

$$\mathbb{P}(f(X) \leq u) = G(f^{-1}(u)).$$

So the limit analog of Open Problem 6 is

Open Problem 7. *What are the possible distributions for $f(X)$, as we vary W and G?*

As before we really mean the *closure* of the set of possible distributions. It is intuitively clear that we can take limits in Proposition 4 below `not worth writing details now`) to show that

> any distribution $f(X)$ arising from some W and G must satisfy $f(X) \preceq U(0,1)$.

But it is not at all clear whether the converse is true.

5.5 Decomposition of variance

`This is only loosely related discussion.`

Consider first arbitrary winning probabilities (p_{ij}). There are several ways one might measure the effective variability of team strengths as indicated by winning probabilities. One way is to consider the variance of the numbers p_{ij}. Another way is to think of $\rho_{ij} := (p_{ij} - \frac{1}{2})^2$ as a measure of the difference in strengths of i and j, and then average over pairs (i,j). These are in fact equivalent. That is, writing (I,J) for an uniform random pick of two distinct teams, because $\mathbb{E}p_{IJ} = \frac{1}{2}$ we have

$$\mathrm{var}\ p_{IJ} = \mathbb{E}\rho_{IJ} := v, \text{ say.}$$

Conceptually one can view this as the general "law of total variance"

$$\tfrac{1}{4} = \mathrm{var}\ 1_A = \mathbb{E}\mathrm{var}(1_A | I, J) + \mathrm{var}\ \mathbb{P}(A|I,J)$$

applied to the event $A := \{I \text{ beats } J\}$. From this viewpoint v measures variability at the level of individual matches: the "uncertainty" var $1_A = \frac{1}{4}$

decomposes as a sum $v + (\frac{1}{4} - v)$ where v is the contribution from the choice of teams and $\frac{1}{4} - v$ is the contribution from the match result.

It is perhaps more interesting to consider variability at the level of teams. Consider the expected proportion of wins by i when playing all other teams once

$$q(i) := \tfrac{1}{n-1} \sum_{j \neq i} p_{ij}$$

(this is the quantity from (17), now normalized). So

$$v' := \mathrm{var}\ q(I) = \mathbb{E}(q(I) - \tfrac{1}{2})^2$$

is a natural measure of variability at the level of teams. So we now have a decomposition

$$\tfrac{1}{4} = v' + (v - v') + (\tfrac{1}{4} - v)$$

where the first term is the contribution to variance from the choice of home team, the second term is the contribution from the choice of opponent, and the third term is the contribution from the match result.

Project. `There should be some way to pursue these ideas in the setting of the basic probability model.`

6 Tournament design and the basic probability model

Consider a single elimination tournament with $n = 2^m$ players. One could assign players to slots at random; a common alternative is to *seed* (i.e. rank) players and place them at positions in a design template depending on their rank. The Wikipedia page [15] shows the standard design, in which the two highest-ranked teams are placed in separate halves, the four highest-ranked teams are placed in separate quarters, etc.

Is this design optimal in some sense? In particular, the following question comes to mind.

> (*) Assuming the basic probability model with given W, is the probability that the top seeded player wins the tournament maximized by the standard design (rather than some other design) for all values of players strengths?

The answer is: for 4 players this is true for all W, but for 8 players it is not true for the logistic function W.

First consider 4 players of strengths $x_1 \geq x_2 \geq x_3 \geq x_4$. In the standard design players 1 and 4 play in the first round. Clearly the probability that player 1 wins the tournament is

$$G(x_1, x_4; x_2, x_3) := (14)[(23)(12) + (32)(13)]$$

where we write (ij) for $W(x_i - x_j)$. There are two alternative other designs, and the assertion that the standard design maximizes the probability in question is the assertion

$$G(x_1, x_4; x_2, x_3) \geq \max(G(x_1, x_3; x_2, x_4),\ G(x_1, x_2; x_3, x_4)). \qquad (23)$$

We can verify this as follows. Using the fact $(ij) + (ji) = 1$ we have

$$G(x_1, x_4; x_2, x_3) = (14)[(13) + (23)[(12) - (13)]]$$

$$G(x_1, x_3; x_2, x_4) := (13)[(14) + (24)[(12) - (14)]]$$

and therefore

$$G(x_1, x_4; x_2, x_3) - G(x_1, x_3; x_2, x_4) = (14)(23)[(12) - (13)] - (13)(24)[(12) - (14)].$$

One can now check that the right side equals

$$(12)(23)[(14) - (13)]\ +\ (13)[(24) - (23)][(14) - (12)]$$

and every term is non-negative. This establishes the first part of (23) and the second part is similar.

I haven't actually checked the second part but it must work ...

Turning to the case of 8 players, Figure xxx show the positions of the 4 top-ranked players in the standard design and in an alternate design.

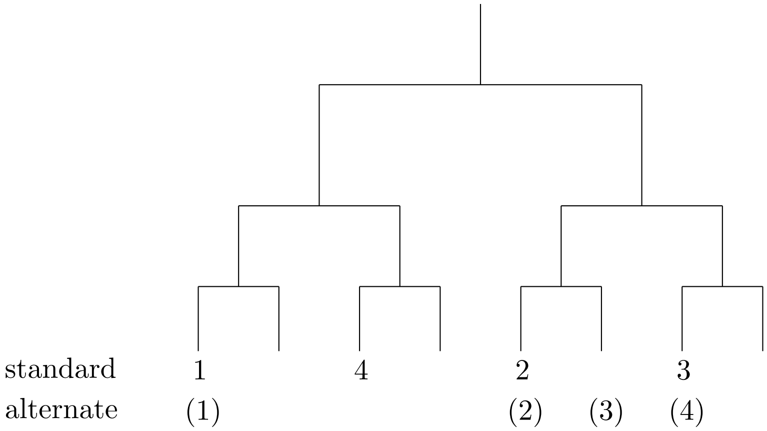

Figure xxx.

Consider the case where the strengths of the 4 top-ranked players are of the form

$$x_1 = x_2 + a; \quad x_3 = x_2; \quad x_4 = x_3 - a$$

and where the other players have negligible chance of beating any of the top 4. In the standard design, up to the negligible terms, the probability that player 1 wins the tournament is

$$W(2a)W(a).$$

In the alternate design the probability is

$$W^2(a) + W(2a)W(-a).$$

As noted earlier, for the logistic W we can choose a so that $W(a) = 2/3$, $W(2a) = 4/5$ and the probabilities become $24/45$ and $32/45$. So property (*) does not hold.

6.1 Possible future work

This is all ‘‘discussion" -- many questions arise from the material above. It's not clear if there is any closely related literature.

1. Our counter-example seems "unfair to player 2" – can one define a class of "fair" tournament designs within which the standard design is optimal, in the sense above?

2. The counter-example above implies counter-examples to other naive optimality conjectures. For instance one might conjecture that the standard design maximizes the chance that the final is between the best two players, but (for 16 players) the counter-example implies one can arrange alternate designs in each half of the bracket to increase the chances of each reaching the final.

3. One could define different "optimality" criteria. For instance the "excitement" of a match where the win probabilities are $(p, 1-p)$ could be defined as $p(1-p)$, the variance of the indicator RV. So in the standard design for a 4-player tournament, the expected excitement of the final is

$$H(x_1, x_4; x_2, x_3) := (14)(23)(12)(21)+(14)(32)(13)(31)+(41)(23)(42)(24)+(41)(32)(43)(34)$$

in the notation above.

Open Problem 8. *Is the standard design optimal in this sense for 4 players – that is, do we have the analog of (23)*

$$H(x_1, x_4; x_2, x_3) \geq \max(H(x_1, x_3; x_2, x_4),\ H(x_1, x_2; x_3, x_4)). \qquad (24)$$

If so then it would be interesting to study whether this optimality result extends in some way to the general $n = 2^m$ tournament.

4. Yuval Peres in conversation suggested another possible optimality criterion

$$\sum_i (n-i)\mathbb{E}(\text{number of matches played by player } i).$$

But again, for 4 players the standard design is not always optimal.

In fact we make the vague conjecture that, for general $n = 2^m$ players, there is no reasonable optimization criterion under which the standard design is not always optimal.

6.2 Broad sense tournaments

Project. There are many schemes for ‘‘broad sense" tournaments, with structure between single elimination and round robin. An interesting recent blog [12] discusses some of these in the context of online games. Can we say anything interesting about the properties of such schemes under the basic probability model?

7 Did the best team win?

At the end of a league season or tournament we can ask *what is the chance the winner was actually the best team?* As mentioned earlier this is fun to do in a popular talk, and of course requires a probability model. For instance [11] describes the following method for the Premier League; each team i has an observed mean μ_i and s.d. σ_i of points earned per match; sample independent Normal$(m\mu_i, m\sigma_i^2)$ (m is number of matches per team) RVs ξ_i and use these – viewed as an independent random repeat of the season – as ratings giving a random ranking of the teams.

To exploit our basic model, the "honest" way would be Bayesian: for instance fix W, take a prior on ratings $\mathbf{x}$, use the complete season results data to give a posterior on ratings $\mathbf{x}$, which provides some probability that the highest-posterior-rated team was the actual winner (for simplicity we suppose here and below that a unique team i^* attains the maximum $\max_i q^*(i)$ actual number of wins).

Project `to actually do this with some (say) baseball or soccer data.`

Here are two related math conjectures, in the *league play* context, and taking a fixed W in our model.

Conjecture 9. *Suppose the season results are arbitrary but with unique most-winning team i^*. Consider the MLE estimates of strengths $\hat{x}_i$. Then it is* **not** *necessarily true that $\hat{x}_i$ is maximized by i^*.*

Conjecture 10. *Consider the basic probability model with strengths $\mathbf{x}$. From a realization of season results calculate the MLEs of strengths – these are now RVs $\hat{X}_i$ depending on the realization, and so there is some random $\hat{I} = \arg\max_i \hat{X}_i$. Then it is always true that $\mathbb{P}(\hat{I} = i)$ is maximized by $\arg\max x_i$.*

`I haven't thought about these but likely they are easy.`

8 More stuff to think about

8.1 The top players

In the "many individual players" setting, the mathematically natural model for strengths of the top players is the inhomogeeous Poisson point process on $\mathbb{R}$ of intensity e^{-x}. That is, we imagine an infinite number of players, whose strengths in decreasing order

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

are such that

$$N(x) := \text{ number of players of strength } \geq x$$

has Poisson(e^{-x}) distribution. Note that here we must abandon our "centered" convention for strengths.

Project. I haven't gotten to any interesting do-able problem in this setting. An immediate conceptually interesting issue is how to model match scheduling amongst top players. oven such a model we can study the cases of unchanging ratings, or of the natural diffusion model for rating fluctuations.

8.2 Scheduling matches based on ratings

Project. Think of models for scheduling matches based on ratings.

8.3 A less restrictive basic model

A less restrictive model than our basic probability model is the model wherein players can be ordered so that $p_{ij} := \mathbb{P}(i \text{ beats } j)$ satisfies $p_{ij} > 1/2$ for $i < j$ and also

$$p_{i\ell} \geq p_{jk} \text{ for } i \leq j < k \leq \ell. \tag{25}$$

Intuitively this is more natural than our basic probability model but fits less well with Elo-type updates.

Vague Project: Is there any interesting property that always holds in our basic probability model but does not always hold in the model above?

To digress slightly, the following result (which the author has used for many years as a graduate take-home exam problem, and so will not give the proof here) essentially implies that even weaker conditions suffice for one to determine the best player in linear time.

Proposition 11. *Suppose that amongst n players the winning probabilities p_{ij} are unknown to us, and are arbitrary except that for some unknown "best" player i^* we have $p_{i^*j} \geq 0.5 + \delta \ \forall j \neq i$, for known $\delta > 0$. Then we can design a (broad sense) tournament which in $C(\delta)n$ matches will enable us to determine the best player with probability $\leq \delta$ of error, where $C(\delta)$ is a constant depending only on δ.*

8.4 From league results to tournament probabilities?

Project. Given the numbers $q^*(i)$ of matches won in league play, is there some plausible way to approximate probabilities of winning a single-elimination tournament amongst the n teams?

References

[1] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika*, 39:324–345, 1952.

[2] Persi Diaconis and David Freedman. Iterated random functions. *SIAM Rev.*, 41(1):45–76, 1999.

[3] Thore Graepel and Tom Minka. Trueskill ranking system, 2006. http://research.microsoft.com/en-us/projects/trueskill/.

[4] Leonhard Knorr-Held. Dynamic rating of sports teams. *The Statistician*, 49:261–276, 2000.

[5] Amy N. Langville and Carl D. Meyer. *Who's #1? The science of rating and ranking*. Princeton University Press, Princeton, NJ, 2012.

[6] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, Providence, RI, 2009. With a chapter by James G. Propp and David B. Wilson.

[7] Michael Lewis. *Moneyball: The Art of Winning an Unfair Game*. W. W. Norton, 2004.

[8] Sean Meyn and Richard L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, second edition, 2009. With a prologue by Peter W. Glynn.

[9] John W. Moon. *Topics on tournaments*. Holt, Rinehart and Winston, New York-Montreal, Que.-London, 1968.

[10] Moshe Shaked and J. George Shanthikumar. *Stochastic orders and their applications*. Probability and Mathematical Statistics. Academic Press, Inc., Boston, MA, 1994.

[11] David Spiegelhalter. May the best team win, 2007. http://understandinguncertainty.org/node/61.

[12] Alex Wice. New formats that would improve tournaments, 2014. http://blog.prismata.net/2014/06/10/new-formats-that-would-improve-tournaments/.

[13] Wikipedia. Bowl championship series — wikipedia, the free encyclopedia, 2014. [Online; accessed 25-October-2014].

[14] Wikipedia. Elo rating system — wikipedia, the free encyclopedia, 2014. [Online; accessed 31-October-2014].

[15] Wikipedia. Seed (sports) — wikipedia, the free encyclopedia, 2014. [Online; accessed 24-November-2014].

[16] Wikipedia. Tournament — wikipedia, the free encyclopedia, 2014. [Online; accessed 4-December-2014].

9 Misc discussion fragments

`To be deleted or merged into paper.`

9.1 Real-world uses of ratings and rankings

Here we will remind readers of matters they know perfectly well, our excuse being a desire to emphasize that these matters are not intrinsically related to probability models. Following [5] we use *rating* to mean a real number produced in some specified way (typically algorithmic, but perhaps involving human judgment) intended to indicate strength of a team, and *ranking* to mean ranking a given set of teams as 1 (best), 2 (second best), etc. So a rating scheme determines a ranking scheme but not conversely.

1. Professional team sports use a variety of schemes to determine a "best team" at the end of a season. One extreme is the *English Premier League* style where each pair of teams play twice in a given season, and the other extreme would be one single-elimination tournament; in practice many variations and combinations are used. The former implicitly provides a rating (percentage of matches won, suitably adjusted for ties) and the latter provide a "best team". In such settings further statistical analysis of results is of limited interest. But in other sports – such as U.S. College Football and many individual sports such as chess or tennis, and amateur sport in general – there are many "teams" compared to games per team, and less systematic scheduling of games. In that context ratings provide comparisons between teams who have never played each other.

2. In some sports, invitations to participate in major tournaments, and *seeding* within the tournament, are based on some explicit or implicit notion of ranking. This is more common in individual sports (tennis or golf) than in team sports where eligibility is usually based explicitly on league play results, but occurs notoriously and controversially in the Bowl Championship Series [13] in U.S. College Football.

3. If you are an aficionado of a particular sport then you might just be curious about, for instance, who are the best female tennis players under age 21 in Germany, and rankings provide one way to answer such questions, or to provoke debate. If you engage in a sport yourself then you are even more likely to be curious about your own rating (the author once had a 1774 rating at Hearts on `pogo.com`) and to wish to play against equally skillful opponents.

9.2 Alternate introductory material

`Need more systematic literature review.`

Overview. The goal of this article is to publicize one aspect, in which challenging mathematical open problems arise as soon as one starts thinking about foundational issues. "Foundational" is of course a common euphemism for "totally divorced from reality", a description which fits this article in two main ways. We work within the very simplistic model below. And the available familiar mathematical techniques typically address time-asymptotics (classical Markov chain theory, applied to models of n teams repeatedly playing each other) or size-asymptotics (mixing times for such chains as $n \to \infty$ [6]). Such results can hardly be considered directly relevant to the real world of finite numbers of teams playing finite number of matches.

The basic model. We perceive this basic model as a building block for more elaborate models addressing different aspects of sports results, somewhat analogous to geometric Brownian motion as the *basic model* for stock prices. As the most obvious use, one can take the data for match results amongst a set of teams over some past time interval and fit it to the model, that is make estimates (for instance MLEs) $\hat{x}_A$ of the strengths x_A and use these estimates as ratings. With W of logistic form (2) this is called the *Bradley-Terry model* [1] and has attracted a large literature in the "consensus ranking" interpretation.

Use as a predictive model. Given estimates $\hat{x}_A$ of strengths from past results (as above), one could predict that the probability that A beats B in the next match is $W(\hat{x}_A - \hat{x}_B)$. For this purpose it would be preferable to estimate the function W (along with the strengths) from data, rather than assume *a priori* a form for W. Of course more serious attempts at predictive models would include more data; individual game scores, performance statistics for individual players, etc. For this reason we do not emphasize the predictive aspect, though a natural "academic" question is to what extent these more elaborate models do better than models based only on win/loss data.

Ratings. Part of the motivation for Elo-type ratings is that we do not want to assume any probability model is exactly correct.

Simulation project – how well does Elo track changing strengths?

n teams (n even).

At each time, teams i have strengths x_i; strengths standardized to have ave = 0 and SD = 1.

At each time step we randomly put the teams into $n/2$ pairs who each play a game.

After each time step the team strengths change in the following random way.

There is a parameter $a > 0$ (say $a = 0.03$). We first update each x_i to $(1-a)x_i + \text{Normal}(0, 1-(1-a)^2)$ using independent Normals; then we linearly transform the updated values to have ave = 0 and SD = 1.

Note this strength updating is independent of the match results.

Match results follow the basic probability model: i beats j with probability $W(x_i - x_j)$. We take another parameter b and use

$W(u) = L(bu)$ where L is standard logistic distribution fn $L(u) = e^u/(1+e^u)$.

For the Elo rankings we have two more parameters, c and k, and we use the update function

$$\Upsilon(u) = kL(-cu)$$

That is, when i beats j their ratings update as

$$y_i \to y_i + \Upsilon(y_i - y_j), \quad y_j \to y_j - \Upsilon(y_i - y_j).$$

Run the simulation for many steps, starting with $y_i \equiv 0$. Then we want to see how accurately the probability of i beating j can be estimated using the current ratings y_i, y_j. There are two ways to estimate this. One way is to use the true W: that is, use the estimate $p(i,j) = W(y_i - y_j)$. The other way is to use the estimate implied by our update function; that is, $p(i,j) = L(-c(y_i - y_j))$. In either case we calculate the MSE, that is the average over all pairs (i,j) of $(p(i,j) - W(x_i - x_j))^2$, and average this over several time steps.

Conceptually, the parameters a, b are given by Nature, but we get to choose the parameters c, k. The simulation project is to see how the optimal (minimizing MSE) choice of (c,k) depends on (a,b). For instance, we expect that the optimal c will be close to b, but it is not clear how the optimal k depends on the given parameters.