Long-Term Stable, Quick-Reacting Table Tennis Ratings via Poisson Jumps

David J. Marcus

September 24, 2023; revised July 27, 2025

Overview

- Ratings Central
- Model and Algorithm
- Other Rating Systems
- Ratings Central Priors
- Ratings Central Problems
- New Temporal-Update Model
- A Rapidly-Improving Player
- Things That Did Not Work That Well
- Status
- Future Work

Ratings Central

Ratings Central

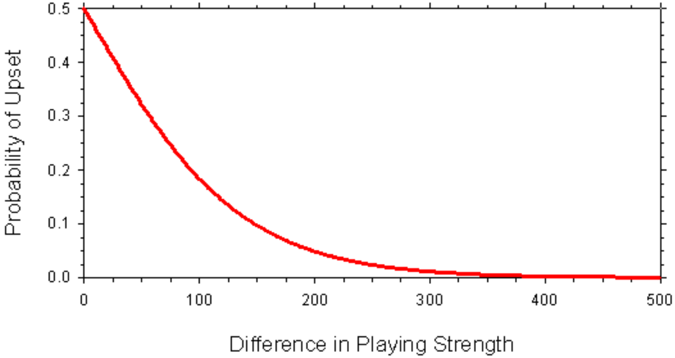
- Ratings Central (https://www.ratingscentral.com) has been providing table tennis ratings for clubs and organizations around the world since 2004.
  - Official ratings/rankings for Austria and Australia.
  - Unofficially, we process some ITTF (International Table Tennis Federation) tournaments.
- As of July 27, 2025, the system contained

  53,324 events,
  124,642 players,
  5,676,985 matches,
  1,316 clubs.
- An *event* is a set of matches that are submitted and processed together as a group, e.g., a tournament or a round in a league. Events can have up to 5,000 matches or 2,000 players. Average is 110 matches and 65 players.
- Algorithm developed 1997–1999 when I was on the USATT (USA Table Tennis) Ratings Committee. USATT did not adopt the algorithm due to politics and nonsense.

Model and Algorithm

The Bayesian Model

- Each player has a *playing strength*, i.e., a number that quantifies how strong the player is.
- Define the *probability-of-upset function* $\pi(x) := 1/(1 + e^{x/67})$.



- The probability that a player with playing strength s will beat (upset) a player with playing strength t is $\pi(t - s)$.

The Bayesian Model (cont.)

- A player's playing strength is not known, so model it as a random variable (the *player's law*) with a normal prior.
- The *temporal update* models a player's playing strength changing with time:
  - Add a zero-mean normal random walk to the player's law with a standard deviation of 70 rating points per year.

Event Update

- The *event update* is the calculation of the posterior law for each player given the match results in the event.
- Let N be the number of players in an event. Let L_j be the start-of-event law for player j.
- Let M be the number of matches in the event. Let $p(m)$ be the number of the player who wins the mth match. Let $q(m)$ be the number of the loser.
- Define

$$U(x_1) := \int_{\mathbf{R}^{N-1}} \prod_{m=1}^{M} \pi(x_{q(m)} - x_{p(m)})\, dL_2(x_2) \cdots dL_N(x_N) L_1(x_1).$$

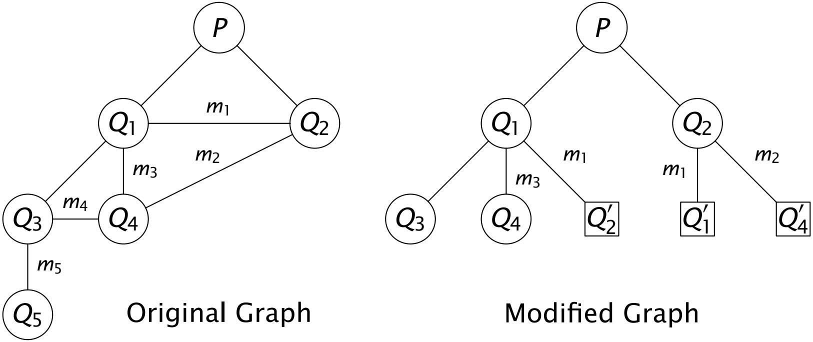
- The posterior law for player 1 is $U / \int_{\mathbf{R}} dU$.

Algorithm

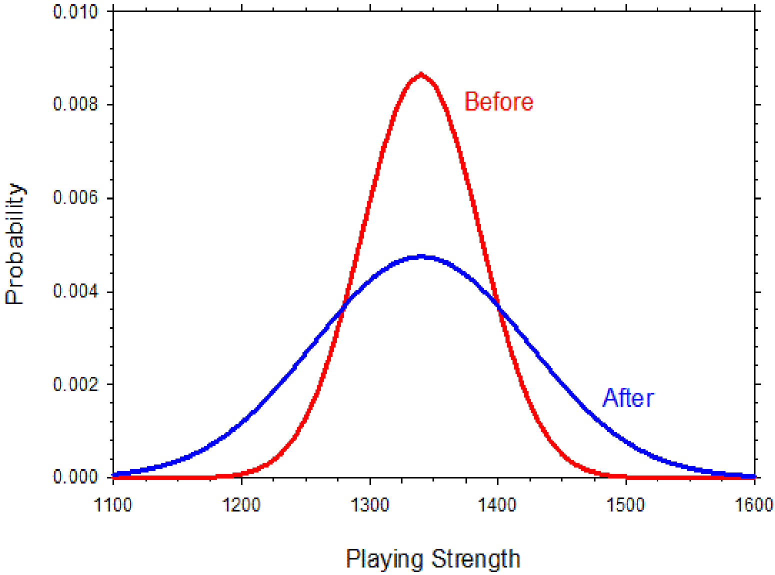
- The system calculates an approximation to the posterior law for each player.
- Calculations are done by replacing continuous laws with discrete laws on $\{0, 10, \ldots, 3500\}$.
- Temporal update is a convolution. Because integrand in formula for the posterior law is a function of playing-strength difference, event update for one match is also a convolution.
  - Convolutions may be rapidly calculated via the FFT.

Tournament Surgery

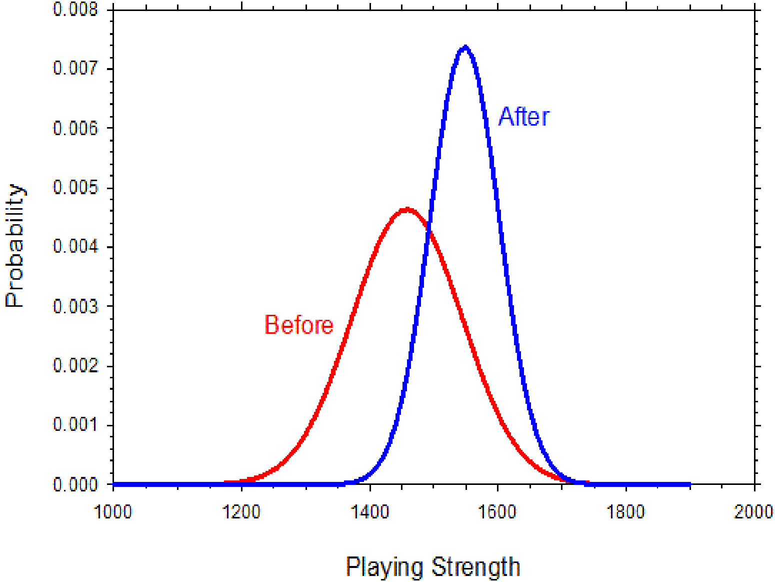
- For an event, consider the graph where each player is a node and each match is an edge connecting two players.
- For each player P, construct a modified graph and use the modified graph to calculate P's posterior law. See the references (on a later slide) for details.



Temporal Update



Event Update



References

- For more information on the algorithm, see the webpage https://www.ratingscentral.com/HowItWorks.php. The webpage also contains links to the following more technical references:

  - Marcus, David J. (2001) “New Table-Tennis Rating System“. Journal of the Royal Statistical Society: Series D (The Statistician), 50: 191–208. doi: 10.1111/1467-9884.00271.
  - Marcus, David J. (2011a) “Ratings Central: Accurate, Automated, Bayesian Table Tennis Ratings for Clubs, Leagues, Tournaments, and Organizations”. Joint Statistical Meetings, July 30–August 4, 2011.
  - Marcus, David J. (2011b) “Ratings Central: Accurate, Automated, Bayesian Table Tennis Ratings for Clubs, Leagues, Tournaments, and Organizations”. NESSIS (New England Symposium on Statistics in Sports), September 24, 2011.

Other Rating Systems

Other Rating Systems

- Elo: Arpad Elo. 1978. Player's rating is a single number. Probability of winning is function of difference in ratings. Not Bayesian; update formula ad hoc. No temporal update. Very popular in chess.
- Glicko: Mark Glickman. 1999. Bayesian. Model very similar to Ratings Central model. Event update only goes one level deep in event graph (i.e., start-of-event ratings used for opponents). Posterior mean and standard deviation calculated directly using approximate formulas (so posterior assumed to be normal).
- Glicko-2: Mark Glickman. 2001. Modification of the Glicko system. Adds a rating volatility to each player's rating. This allows the temporal update to be different for different players. Update algorithm uses an iterative procedure.

Ratings Central Priors

Ratings Central Priors

- Priors come from the event directors.
- We provide guidance on how to set priors.
- Our guidance has become more explicit and more restrictive over the years because we have learned that most directors cannot set sensible priors without our help.
- As part of the project that led to the improved temporal model (with the Poisson jumps), we helped several organizations improve the priors that they were using and helped them fix the priors that they had used in previous submissions.

Strategies for Setting Priors

- Sometimes an organization can use a single prior for all their events. This can work if they run many tournaments so players play a wide range of other players.
- If an organization runs mostly leagues, where the leagues are organized by playing level, then it is usually necessary to use different priors for each division.
  - For example, league structure might be best players in Division A, next best in Division B, with eight or more divisions, plus maybe divisions for juniors or seniors.
- Event directors like to set individual priors for players.
  - This is sometimes useful, e.g., if the director submits few events or a particular player is very different from the main population.
  - We have been discouraging this because it produces a proliferation of model parameters and can produce unexpected results.

Ratings Central Problems

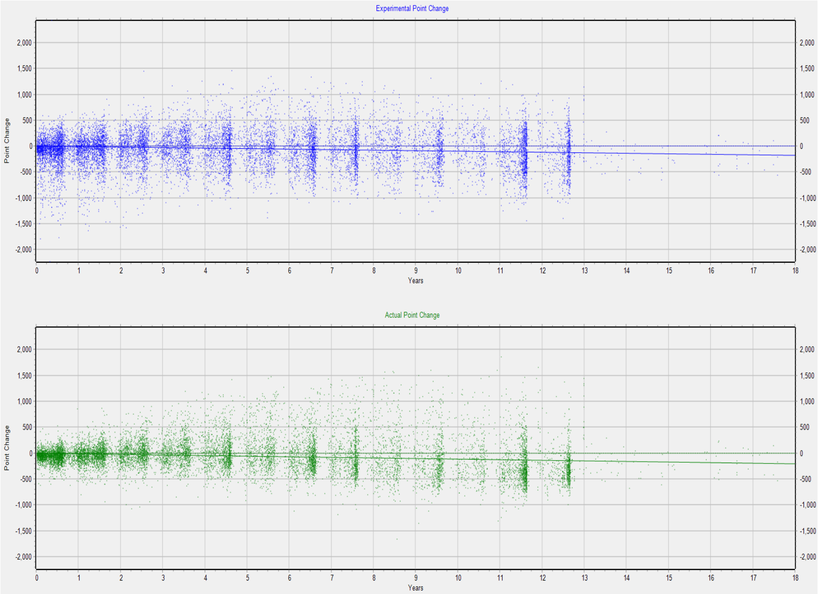
Ratings Central Problems

- The Austrian Association reported deflation for their players.
  - Deflation not evident for adult ITTF tournament players.
- Rapidly-improving players (typically juniors) were sometimes underrated.

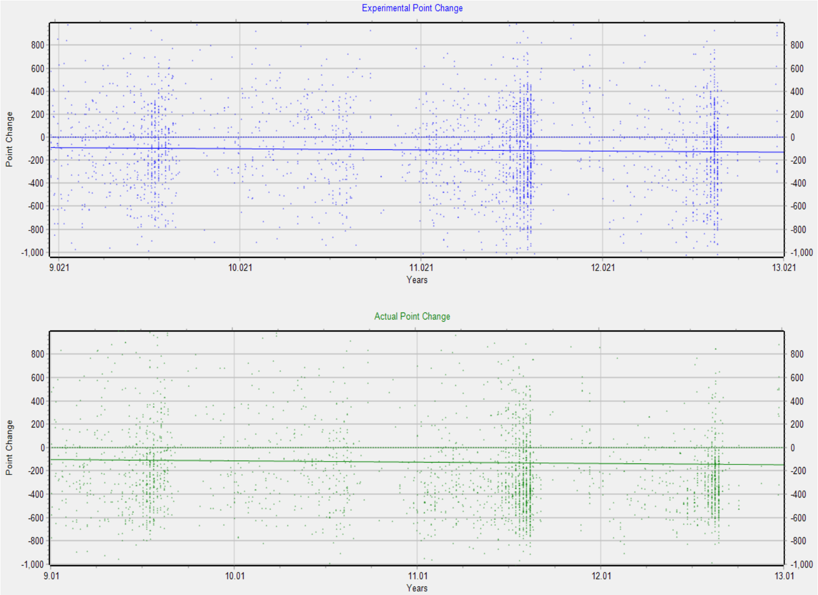
Austrian Priors As of 2018

- The Austrians used many different priors, basing them on division, gender, age, and Austrian state.
  - They have leagues where very young children or beginners play.
- They adjusted the priors each year by looking at the ratings of the players in the division at the end of the previous year.
  - We did not recommend this.
- For the first events in some Austrian states, they used individual priors based on ratings of the players from a different rating system.
  - 1605 different priors in 39 events.
  - We did not recommend this.

Rating Change Graph



Rating Change Graph Zoomed In



Rating Change Graphs: Data

- Approximately 15,000 Austrian players in system at end of 2018.
- Each dot is one Austrian player.
  - Horizontal axis is years between player's first and last event.
  - Vertical axis is player's rating after last event minus rating after first event.
- Bottom graph ("Actual") uses ratings in actual system at the time.
- Top graph ("Experimental") uses ratings from experiment where $N(1500, 450^2)$ prior used for all events submitted by Austrian directors.
  - Austrian players also play in ITTF adult and junior tournaments.
- Each graph includes the regression line.

Rating Change Graphs: Comments

- Horizontal clumping is because league season is from September to May.
- Blobs in Actual graph tend to be asymmetrical, with many of the players in the bottom part of the blob. Blobs in Experimental graph look much more symmetrical with many players gaining points.
- Actual graph has more players who gained a lot of points. This is odd.
- Actual graph has much tighter blobs, especially in early years.
- In Actual graph, deflation looks worse than regression line. Regression line probably pulled up by players who gained a lot of points.
- Experimental graph looks much more reasonable than Actual graph.

Conclusion

- Deflation partly caused by poor priors, but even with better priors, there is still some deflation.
  - Approximately 10 points per year.

New Temporal-Update Model

Cause of Deflation

- Probability-of-upset function basically just sets scale.
  - Shape could be wrong, but unlikely to cause a problem.
- Event update is approximate, but is exact for tournament after surgery. So, unlikely to cause deflation.
- Temporal update may be wrong.
  - If players are improving more than the temporal model thinks likely, the ratings would deflate.
- Sometimes, players improve very rapidly.
  - A player may master a new technique (e.g., footwork, stroke, serve) well enough to use in matches, and is then a level better.
  - 200 rating point improvement almost overnight.

Candidate Temporal-Update Models

- Normal random walk with mean of 7 or 10 points per year and standard deviation of 70 points per year.
- Poisson jump process with mean of 7 or 10 points per year plus mean-zero normal random walk with standard deviation of 70 points per year.
- Glicko-2 model where temporal update is different for different players.
  - I do not know how to get this algorithm to use the same rating scale as Ratings Central, so did not try it.

Poisson Process

- A Poisson process is a stochastic process where an event happens at random times, e.g., customers arriving at the bank. Assume time is discrete and measured in days.
- The number of events n on a given day has a Poisson distribution with parameter λ:

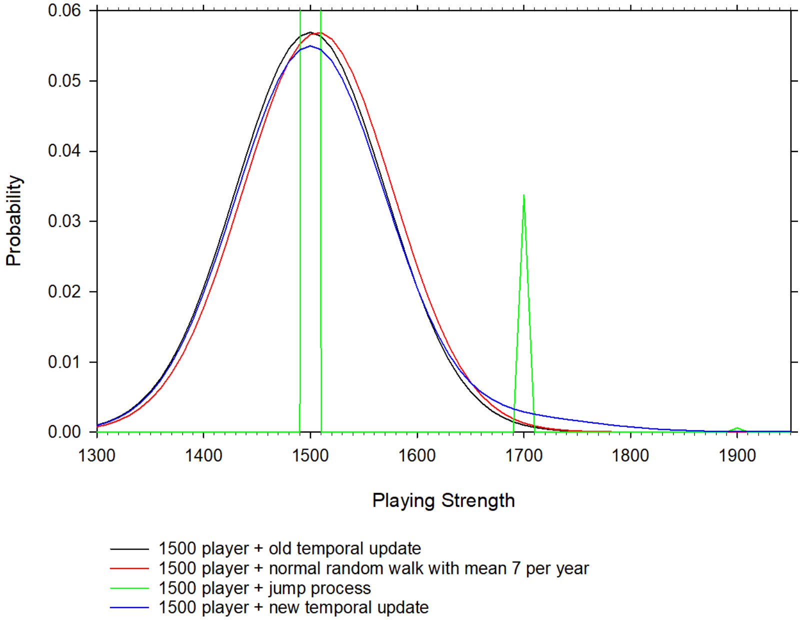
$$P[n] = \frac{\lambda^n e^{-\lambda}}{n!}.$$

- If λ is small, then the probability of no events is close to 1, and the probability of one event is a bit less than λ.
  - For λ small, Taylor's Theorem says that $e^{-\lambda} \approx 1 - \lambda$. So, $P[1] \approx \lambda(1 - \lambda) \approx \lambda$.
- The mean and the variance of a Poisson random variable are both λ.
- The sum of independent Poisson random variables is Poisson. The λ parameter for the sum is the sum of the λ parameters.

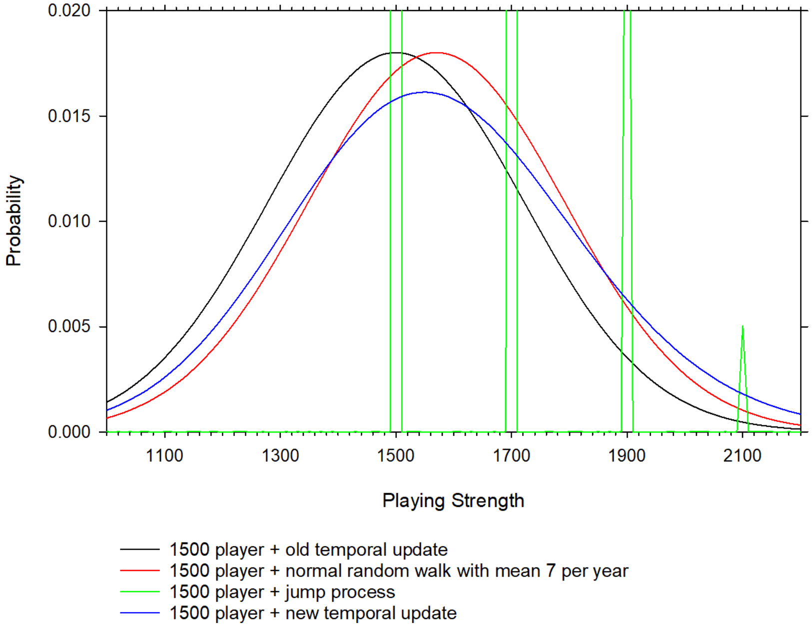
New Temporal-Update Model

- Let $A_i, i = 1, 2, 3, \ldots$ be independent Poisson random variables with mean λ.
- Let $B_t := \sum_{i=1}^{t} A_i$, i.e., the number of jumps in t days. The mean of B_t is $t\lambda$.
- Let the jump size $S := 200$ rating points.
- Let the jump process be $J_t := SB_t$, i.e., the increase in rating due to the jumps in t days. The mean of J_t is $St\lambda$. Let $\lambda := 7/(365 \cdot S)$. So, the mean for a year is 7 rating points.
- The new temporal-update model is the sum of the Poisson jump process J and the old mean-zero normal random walk with standard deviation of 70 points per year. Total standard deviation is 79 points per year.

Temporal Updates for 1 Year



Temporal Updates for 10 Years

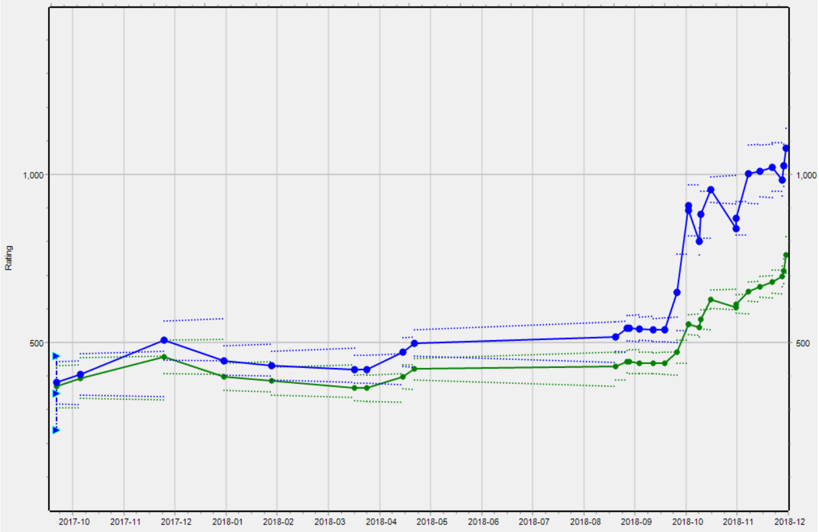


A Rapidly-Improving Player

A Rapidly-Improving Player

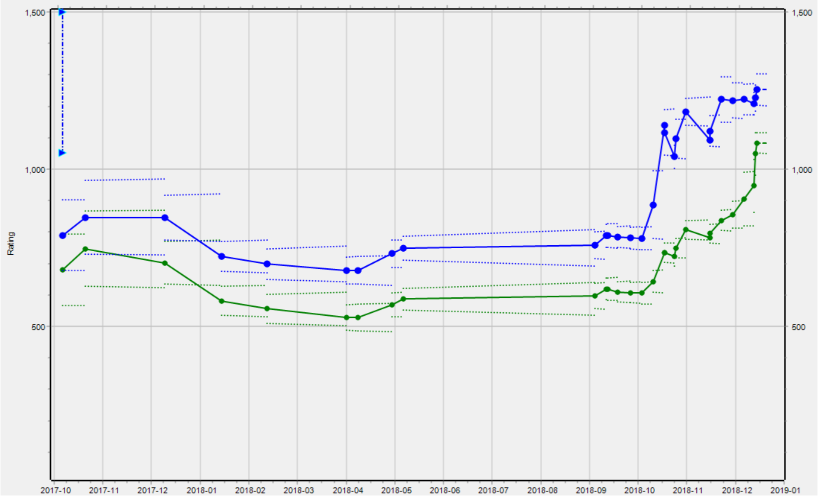
- While I was developing the new model, a parent complained that their child was underrated.
- Graphs on following slides show the child's rating history using various temporal updates and priors.
- In each case, all events for all players were calculated using the specified temporal update.

Temporal-Update Comparison With Actual Priors



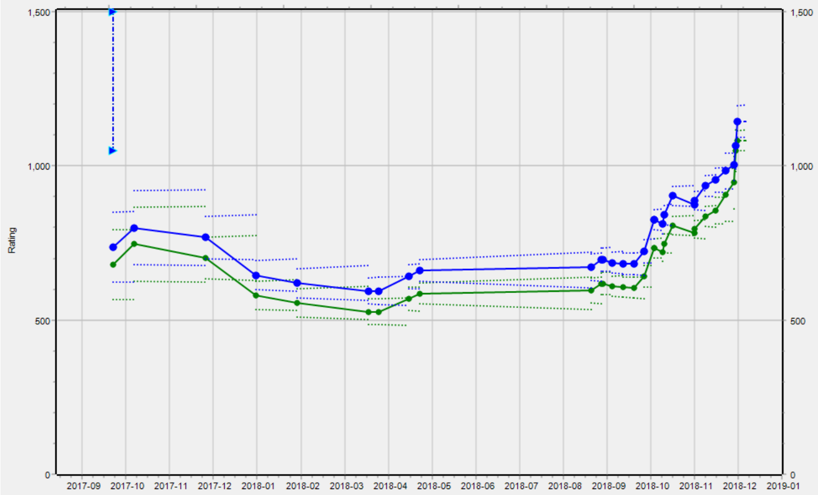
- Actual priors. Blue: New temporal update with jump process. Green: Old temporal update.

Temporal-Update Comparison With Generic Priors



- $N(1500, 450^2)$ prior for Austrian players. Blue: New temporal update with jump process. Green: Old temporal update.

Normal Random-Walk Comparison



- $N(1500, 450^2)$ prior for Austrian players. Blue: Normal random walk with mean 7 points per year and standard deviation 79 points per year. Green: Old temporal update.

Conclusion

- The graph with the jump process and generic priors is the only graph for this player where the rating does not look like it is still trying to catch up with the player.

Things That Did Not Work That Well

Simulation

- Simulated individual events:
  - For each player, generated playing strength from start-of-event law.
  - For each match, generated result using probability-of-upset function.
- Average points gained was essentially zero.
  - Not surprising because event update does calculate posterior for the after-surgery event.
- Other statistics calculated from simulation did not provide any enlightenment.
- Did not know how to usefully simulate multiple sequential events (to get rating histories of players) because ratings in each simulated event would be very different from actual ratings (because simulated match results would be different from actual match results). Maybe could have done something along these lines.
- Main purpose in doing simulation was to see if the event update was producing deflation. It did not seem to be.

Scoring Rule

- Compare models by using the logarithmic scoring rule.
  - This is a standard scoring rule.
  - Under certain conditions, the difference in logarithmic scores is the log of the Bayes factor.
- For each match, calculate the log of the probability of the match result using the start-of-event laws. Sum over all the matches.
  - This is different from the log of the probability of all the match outcomes, which is what theory says to calculate, but is hard to calculate.
- Use an optimization algorithm to select the model parameters that give the best score.
- Did not produce anything useful.
  - Sometimes picked reasonable parameters.
  - But often picked parameters that were not believable, e.g., with large nonzero means resulting in almost all players gaining hundreds of points over the years.
  - Optimum parameters for different subsets of the data were often very different.

Status

Status

- The new temporal-update model with the Poisson jump process has been live on Ratings Central since June 1, 2019.
- Ratings for all historical data were recalculated using the new model.

Future Work

Doubles Ratings

- A pickleball group runs doubles tournaments using this format: 16 players, each round 4 doubles matches, four rounds, players change doubles partners each round.
- Since partners change, want ratings for individual players rather than doubles teams.
- The pickleball people suggested team playing strength should be 2/3 of stronger-player playing strength plus 1/3 of weaker-player playing strength.
  - Do not know if this is a good model or necessary.
  - Greatly complicates the math.
  - Assume the team playing strength is average of the team members' playing strengths.

Notation

- Let P and Q be the playing strengths of players on team 1. Let R and S be for team 2.
- Suppose team 1 lost to team 2.
- Want the law of P given the match result.
- Assume P, Q, R, and S are independent.
- For a random variable Z, let L_Z be the law of Z.

Algorithm

- Want

$$U(p) := \int \pi((p+q)/2 - (r+s)/2)\, dL_Q(q) dL_R(r) dL_S(s) L_P(p).$$

- Let $A := (R+S)/2$. Can calculate L_A quickly because it is basically a convolution (trivial if R and S are normal). Then

$$U(p) = \int \pi((p+q)/2 - a)\, dL_Q(q) dL_A(a) L_P(p).$$

- Let $B := A - Q/2$. Can calculate L_B quickly because it is basically a convolution (trivial for normals). Then

$$U(p) = \int \pi(p/2 - b)\, dL_B(b) L_P(p).$$

- Can calculate this quickly because it is basically a convolution.
- Could use this with Ratings Central algorithm or with Glicko.