Lecture 10: Multiagent Evaluation

Jun Wang
UCL

Table of Content

- Elo rating system
- Glicko
- TrueSkill
- Multidimensional Elo (mElo2k)
- α-rank

Importance of evaluation

- The third wave of AI was driven by *ImageNet* benchmark in Computer Vision
- In Natural Language Processing we have *SuperGLUE Benchmark*
- In Reinforcement learning, we have *atari* and *OpenAI Gym*
- How about multiagent learning?
  - What would be the tasks?
  - Difficult as require evaluation against another agent

The Elo Rating System[1]

- Elo rating assigns each player a numerical rating score representing their ability from historical plays
  - developed by Arpad Elo for chess competition
  - widely used for many games such as chess, football, tennis, and video games
- It is computed based on the aggregated wins and losses weighted by how likely the win or loss was based on the opponent's ability
- Elo evaluation system has various purposes:
  - match players against each other when players have similar rating score
  - use skill rating to have cut-offs for different skill tiered tournaments

[1] Arpad E Elo. *The rating of chessplayers, past and present*. Arco Pub., 1978.

The Elo Rating System

- Elo assigns a rating r to each player
- According to rating r, the predicted winning probability is:

$$\hat{p}_{ij} = \sigma(\alpha r_i - \alpha r_j) = \frac{1}{1 + e^{-\alpha(r_i - r_j)}}$$

- Minimizing cross entropy of the predicted $\hat{p}$ and the true p gives

$$L_{Elo}(p_{ij}, \hat{p}_{ij}) = -p_{ij} log(\hat{p}_{ij}) - (1 - p_{ij}) log(1 - \hat{p}_{ij})$$

- *stochastic gradient descent* up finishing a game at time t:

$$r_i^{t+1} \leftarrow r_i^t + \eta(S_{ij}^t - \hat{p}_{ij}^t) = r_i^t + \eta(S_{ij}^t - \hat{p}_{ij}^t)$$

  where S_{ij}^t is a binary value, i winning j gives 1 otherwise 0
- Elo ratings are stationary at:

$$\sum_j \hat{p}_{ij} = \sum_j \bar{p}_{ij} = \sum_j \frac{S_{ij}}{N_{ij}}$$

Explanation of super-parameters in ELO

- α is an additional parameter that represent the scaling of the skill rating being used.

$$\alpha = \frac{\ln(10)}{400}, \ \hat{p}_{ij} = \frac{1}{1 + e^{-\alpha(r_i - r_j)}} = \frac{1}{1 + 10^{-(r_i - r_j)/400}}$$

  For example, if we want to $\hat{p}_{ij} = 0.8$ then $r_i - r_j \approx 240$.

- η is a constant, frequently 32 in chess, used to control the change magnitude of ratings at every step.

$$r_i^{t+1} \leftarrow r_i^t + \eta(S_{ij}^t - \hat{p}_{ij}^t)) = r_i^t + \eta(S_{ij}^t - \hat{p}_{ij}^t)$$

- In addition, according to the above formula, the two players participating in a competition will increase or decrease the same score.

Glicko[2]

- Glicko extends Elo, considering the rating reliability
  - Suppose two players both rated 1700
  - The first defeating the second. In ELO the first player gains 16 rating points; the second loses 16 rating points
  - The first player last played at several years ago, but the second plays every week. So the first player's 1700 rating score is less reliable than the second. Intuitively,
  - The first player should add more than 16 points, because his rating is not believable and he defeated a player with a fairly precise rating of 1700.
  - The second player should decrease less than 16 points because his rating is already precisely measured to be near 1700, and he loses to a player with unreachable rating.
- Glicko adds ratings deviation(RD) to measure the uncertainty of a player's rating.

[2] Mark E Glickman. "The glicko system". In: (1995).

Glicko Formulation

- Rating
  - Rating r measures the mean skill of each player and is updated from game outcomes.
- RD(Rating Deviation)
  - High RD correspond to unreliable rating. Higher RD indicates a player has only participated small number of games, or not play frequently
  - Changes from game outcomes and from the passage of time when not playing. Playing games always decreases player's RD, and time passing without competing always increases player's RD.
- Rating changes in Glicko system are not balanced as they usually are in the Elo system.
- Player's strength with 95 % confidence in interval $[r - 2 * RD, r + 2 * RD]$.

Glicko Rating Algorithm

A rating period consists of many pairwise competitions, such as tournaments.

Step 1: at beginning of a rating period:

- New player: initialize rating $= 1500$, RD$= 350$.
- Old player: $RD = \min(\sqrt{RD_{old}^2 + c^2 t}, 350)$, where c is a constant and t represents the number of rating periods without playing the game.

Step 2: at end of a rating period, updating ratings $\boldsymbol{r}$ and RD's according to game results

Updating ratings and RDs

- Suppose a player has competed with m opponents in this rating period, with ratings $r_1,\ r_2,\ \ldots\ r_m$, and rating deviations $RD_1,\ RD_2,\ \ldots\ RD_m$.
- Let $s_1,\ s_2,\ \ldots\ s_m$ are the results of m competitions, with value 0, 0.5, or 1 for a loss, draw and win.
- Updating rating and RD:

$$r' = r + \frac{q}{1/RD^2 + 1/d^2} \sum_{j=1}^{m} g(RD_j)(s_j - E(s|r, r_j, RD_j))$$

$$RD' = \sqrt{\left(\frac{1}{RD^2} + \frac{1}{d^2}\right)^{-1}}, \textit{where}$$

$$q = \frac{\ln 10}{400},\ g(RD) = \frac{1}{1 + 3q^2(RD^2)/\pi^2}$$

$$E(s|r, r_j, RD_j) = \frac{1}{1 + 10^{-g(RD_j)(r - r_j)/400}}$$

$$d^2 = \left(q^2 \sum_{j=1}^{m} g(RD_j)^2 E(s|r, r_j, RD_j)(1 - E(s|r, r_j, RD_j))\right)^{-1}$$

TrueSkill[3]

- Use Gaussian to model player's skill and performance in game which has mean and variance, not just a skill as Elo.

Suppose each player i has skill $s_i \sim \mathcal{N}(s_i; \mu_i, \sigma_i)$, performance $p_i \sim \mathcal{N}(p_i; s_i, \beta)$ with a fixed variance β. The probability of i beating can be denoted as

$$P(p_1 > p_2 | s_1, s_2) > \Phi(\frac{s_1 - s_2}{\sqrt{2}\beta}),$$

where Φ denotes the cumulative density of standard normal distribution.

- It can deal with any number of competing entities and can infer individual skills from team results

[3] Ralf Herbrich, Tom Minka, and Thore Graepel. "TrueSkill™: a Bayesian skill rating system". In: *Proceedings of the 19th international conference on neural information processing systems*. 2006, pp. 569–576.

Transitive and non-transitive games

- Elo rating assumes that relative skill is transitive
- In Rock-paper-scissors game,the maxent Nash equilibrium[4] is $\boldsymbol{p}_{\boldsymbol{A}}^{*} = (1/3, 1/3, 1/3)^{T}$ and the **Nash average** is $\boldsymbol{n}_{\boldsymbol{A}} := \boldsymbol{A} \cdot \boldsymbol{p}_{\boldsymbol{A}}^{*} = (0, 0, 0)^{T}$.

$\boldsymbol{A}$	R	P	S
R	0	-1	1
P	1	0	-1
S	-1	1	0

Table: Intransitive case, thus anti-symmetric matrix

- It's easy to check that the unique maxent Nash for $\boldsymbol{A'}$ is $\boldsymbol{p}_{\boldsymbol{A}}^{*} = (1, 0, 0)^{T}$,Nash average is $\boldsymbol{n}_{\boldsymbol{A}} = (0, -1, -2)^{T}$

$\boldsymbol{A'}$	a	b	c
a	0	1	2
b	-1	0	1
c	-2	-1	0

Table: Transitive case

[4] with greater entropy than any other Nash equilibrium

Why Elo can only solve transitive problems?

- Elo rating cannot solve intransitive problems.
- For example, rock-paper-scissors (e.g. paper beats rock with $p = 1$,but the predicted $\hat{p} = 0.5$)

Hodge decomposition[5]**:** any antisymmetric matrix decomposes as

$$\boldsymbol{A} = \textit{transitive component} + \textit{cyclic component} = grad(\boldsymbol{r}) + rot(\boldsymbol{A})$$

- The divergence of $\boldsymbol{A}$ is $\boldsymbol{r} = div(\boldsymbol{A}) = \frac{1}{n}\boldsymbol{A} \cdot 1$. $grad(\boldsymbol{r}) = \boldsymbol{r}1^{\boldsymbol{T}} - 1\boldsymbol{r}^{\boldsymbol{T}}$
- The rotation of $\boldsymbol{A}$ is $rot(\boldsymbol{A})_{ij} = \frac{1}{n}\sum_{k=1}^{n} curl(\boldsymbol{A})_{ijk}$ $curl(\boldsymbol{A})_{ijk} = \boldsymbol{A}_{ij} + \boldsymbol{A}_{jk} - \boldsymbol{A}_{ik}$
- There is an Elo rating that generates probabilities $\boldsymbol{P}$ iff curl(logit $\boldsymbol{P}$) $= 0$

[5] David Balduzzi Karl Tuyls Julien Perolat Thore Graepel. "Re-evaluating Evaluation". In: (NIPS 2018); X. Jiang. "Statistical ranking and combinatorial Hodge theory". In: (2011).

Multidimensional Elo(mElo$_{2k}$)[6]

mElo$_{2k}$ can solve intransitive problems because it considers the low-rank(2k) approximation of the cyclic component of the matrix logit(P).

- $C^T \Omega C$ is an approximation to the **Schur decomposition** of the cyclic component $\tilde{A}$.

$$\mathbf{A}_{n\times n} = \text{grad}(\mathbf{r}) + \tilde{\mathbf{A}} \approx \text{grad}(\mathbf{r}) + \mathbf{C}^\intercal \begin{pmatrix} 0 & 1 & & \\ -1 & 0 & & \\ & & \ddots & \end{pmatrix} \mathbf{C} =: \text{grad}(\mathbf{r}) + \mathbf{C}^\intercal_{n\times 2k} \mathbf{\Omega}_{2k\times 2k} \mathbf{C}_{2k\times n}$$

- Let mElo2k assign each player Elo rating r_i and 2k-dimensional vector c_i.

$$\textbf{mElo}_{2k}\textbf{:}\ \hat{p}_{ij} = \sigma\Big(r_i - r_j + \mathbf{c}_i^\intercal \cdot \mathbf{\Omega}_{2k\times 2k} \cdot \mathbf{c}_j\Big) \text{ where } \mathbf{\Omega}_{2k\times 2k} = \sum_{i=1}^{k} (\mathbf{e}_{2i-1}\mathbf{e}_{2i}^\intercal - \mathbf{e}_{2i}\mathbf{e}_{2i-1}^\intercal).$$

[6] David Balduzzi Karl Tuyls Julien Perolat Thore Graepel. "Re-evaluating Evaluation". In: (NIPS 2018).

α-rank[7]

α-rank provides an alternative solution concept for multi-player normal form game

		II		
		L	C	R
	U	$2,1$	$1,2$	$0,0$
I	M	$1,2$	$2,1$	$1,0$
	D	$0,0$	$0,1$	$2,2$

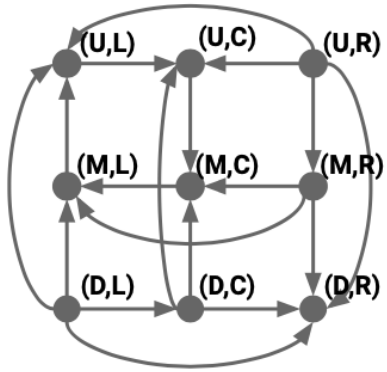
- In this case,there are two players,player 1's strategy set is U,M,D,player 2's strategy set is L,C,R,and there are 3*3=9 strategy profiles (U,L),(U,C),(U,R).....

[7] Shayegan Omidshafiei et al. "α-rank: Multi-agent evaluation by evolution". In: *Scientific reports* 9.1 (2019), pp. 1–29.

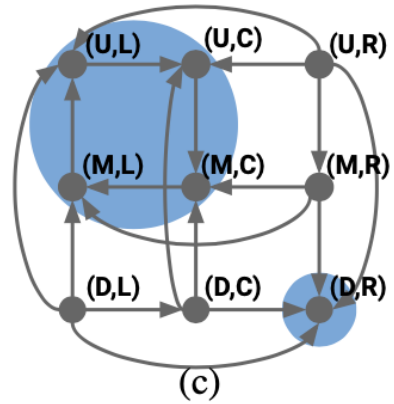
α-rank

		II		
		L	C	R
I	U	$2,1$	$1,2$	$0,0$
	M	$1,2$	$2,1$	$1,0$
	D	$0,0$	$0,1$	$2,2$

(a)



(b)



(c)

- According to the payoff table M in Fig(a), construct the Response graph
- In response graph nodes correspond to strategy profiles, and directed edges if the deviating player's new strategy is a better-response.[8] For example in Fig(b), from (U,L) to (U,C) player 2's payoff increases from 1 to 2.
- The sink strongly connected components (SSCC) of the response graph are illustrated in Fig(c) in blue. Desirable strategy profiles are all in SSCC, because the edges always point to the increasing direction.

[8] Karl Tuyls Mark Rowland Shayegan Omidshafiei. "Multiagent Evaluation under Incomplete Information". In: (NIPS 2019).

α-rank

- In order to be able to find SSCC and its stationary distribution-Rank creates a so-called, Markov-Conley chain, where the edges are "soft".

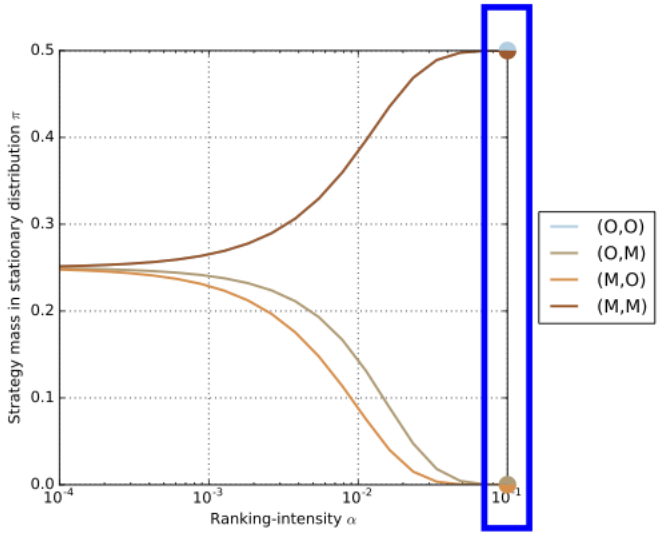
$$\mathbf{C}_{s,\sigma} = \begin{cases} \eta \frac{1-\exp\left(-\alpha\left(\mathbf{M}^k(\sigma)-\mathbf{M}^k(s)\right)\right)}{1-\exp(-\alpha m(\mathbf{M}^k(\sigma)-\mathbf{M}^k(s)))} & \text{if } \mathbf{M}^k(\sigma) \neq \mathbf{M}^k(s) \\ \frac{\eta}{m} & \text{otherwise}, \end{cases} \quad \text{and} \quad \mathbf{C}_{s,s} = 1 - \sum_{\substack{k \in [K] \\ \sigma | \sigma^k \in S^k \setminus \{s^k\}}} \mathbf{C}_{s,\sigma},$$

- The stationary distribution computed by α-Rank corresponds to a ranking of strategy profiles:

$$\pi = C^T \pi$$

α-rank results - Battle of the Sexes

- As α increases, two sink chain components emerge corresponding to strategy profiles (O, O) and (M, M).



(b) Ranking-intensity sweep.

Agent	Rank	Score
(O,O)	1	0.5
(M,M)	1	0.5
(O,M)	2	0.0
(M,O)	2	0.0

(c) α-Rank results.

	O	M
O	(3,2)	(0,0)
M	(0,0)	(2,3)

References I

Arpad E Elo. *The rating of chessplayers, past and present*. Arco Pub., 1978.

Mark E Glickman. "The glicko system". In: (1995).

David Balduzzi Karl Tuyls Julien Perolat Thore Graepel. "Re-evaluating Evaluation". In: (NIPS 2018).

Ralf Herbrich, Tom Minka, and Thore Graepel. "TrueSkill™: a Bayesian skill rating system". In: *Proceedings of the 19th international conference on neural information processing systems*. 2006, pp. 569–576.

References II

X. Jiang. "Statistical ranking and combinatorial Hodge theory". In: (2011).

Karl Tuyls Mark Rowland Shayegan Omidshafiei. "Multiagent Evaluation under Incomplete Information". In: (NIPS 2019).

Shayegan Omidshafiei et al. "α-rank: Multi-agent evaluation by evolution". In: *Scientific reports* 9.1 (2019), pp. 1–29.