Max-Min Thompson Sampling for the K-armed Dueling Bandit Problem

The University of Manchester

Joseph Mellor and Jonathan Shapiro
mellorj@cs.man.ac.uk jonathan.l.shapiro@manchester.ac.uk
School of Computer Science, University of Manchester

Motivation

The duelling bandit problem was introduced by Yue et al. (2011). It is motivated by applications where absolute rewards have no natural scale or are difficult to measure. In applications where a user has to provide feedback it is often easier for them to express a preference between two alternatives than to provide a value-judgement on a single option. For example, a user may provide preference between a set of search engine results by choosing one instead of another.

Thompson Sampling is an algorithm for the multi-armed bandit algorithm that has empirically shown impressive performance as well as strong theoretical guarantees. It is based on using a Bayesian posterior of the reward. For a given application, this should allow expert knowledge about the application domain to be more easily incorporated. It has also been suggested that the stochastic nature of the strategy provides some robustness to delayed rewards.

We propose an algorithm for the duelling-bandit problem based on a game-theory analogy that incorporates Thompson Sampling. We show empirically performance comparable, or better than state-of-the-art algorithms for the problem. We also provide a problem-dependent regret bound.

Problem Definition

In the duelling bandit problem there are K arms. There is assumed to be a preference matrix, unknown to the agent, where element $p_{i,j}$ is the probability that arm i is preferred to arm j. The preference matrix is such that there exists a unique Condercet winner. Let arm w be the Condercet winner, then the Condercet winner is defined such that $p_{w,j} > 1/2 \quad \forall j \neq w$.
At each time step the agent must select a pair of arms (i, j). The environment then returns a preference between the arms, observed by the agent, according to the probability specified in the preference matrix. The simple regret r_t at timestep t is given by

$$r_t = \frac{p_{w,i} + p_{w,j} - 1}{2}$$

The cumulative regret is the cumulative sum of the simple regret over the time the agent has been selecting arm pairs.

$$R_T = \sum_{t=0}^{T} r_t$$

The objective of the agent is to minimize the expected cumulative regret.

Other forms of simple regret are possible. The above was chosen as it was used by Zoghi et al. (2013) and so provided a suitable definition to allow comparison with previous work.

Approach

Our approach is as follows

Phase 1: Play a two-player extensive-form game to initially pick arms to play, using posterior sample estimates (like Thompson Sampling) to balance exploration.

Phase 2: Potentially alter choice of column player to further steer exploration/exploitation.

The first phase requires us to define the payoff matrix for the two-player game. In this phase the game is constructed so that it is undesirable for each player to play the same arm. After the game is played there is an arm selected by the row player and an arm selected by the column player. In the second phase the column players arm can be changed. For instance, the strategy might decide to have the column player play the same arm as the row player.

From preference matrix to payoff matrix

The preference matrix can be transformed to a payoff matrix for a two-player game.
The payoff for the row player is

$$\begin{cases} p_{i,j} - \frac{1}{2} & \text{if } i \neq j \\ -\frac{1}{2} & \text{if } i = j \end{cases}$$

This leads to a zero-sum game that is anti-symmetric.

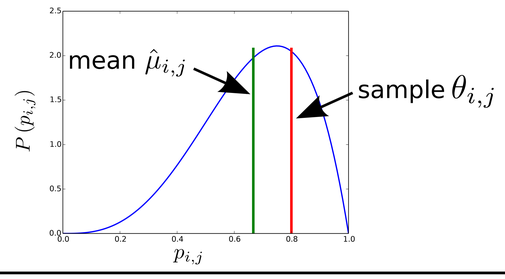
$$\begin{bmatrix} p_{1,1} & p_{1,2} & p_{1,3} & p_{1,4} \\ p_{2,1} & p_{2,2} & p_{2,3} & p_{2,4} \\ p_{3,1} & p_{3,2} & p_{3,3} & p_{3,4} \\ p_{4,1} & p_{4,2} & p_{4,3} & p_{4,4} \end{bmatrix} \rightarrow \begin{bmatrix} -\frac{1}{2} & p_{1,2}-\frac{1}{2} & p_{1,3}-\frac{1}{2} & p_{1,4}-\frac{1}{2} \\ p_{2,1}-\frac{1}{2} & -\frac{1}{2} & p_{2,3}-\frac{1}{2} & p_{2,4}-\frac{1}{2} \\ p_{3,1}-\frac{1}{2} & p_{3,2}-\frac{1}{2} & -\frac{1}{2} & p_{3,4}-\frac{1}{2} \\ p_{4,1}-\frac{1}{2} & p_{4,2}-\frac{1}{2} & p_{4,3}-\frac{1}{2} & -\frac{1}{2} \end{bmatrix}$$

Notice that the payoff, if both players choose the same arm, is artificially forced as to make it the worst possible choice. This design decision was taken because nothing is learned by the agent when choosing the same arm twice in a pair. The possibility to choose this pair is given later in phase 2.

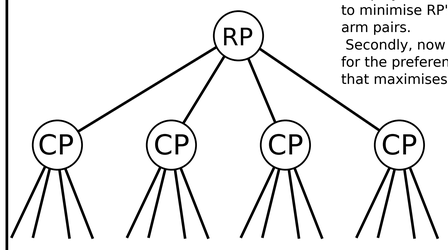
Estimating the preference matrix

Each non-diagonal probability in the preference matrix is unknown and so must be estimated from observations. Like with Thompson Sampling we quantify our uncertainty in the preference by a posterior distribution. For this problem a Beta distribution is used, since the feedback is Bernoulli. Letting $n_{i,j}$ be the number of times arm i has been preferred to arm j, the posterior is given by

$$\text{Beta}\left(n_{i,j}+1, n_{j,i}+1\right)$$



Two-player extensive-form game



Row player (RP) picks first arm.
Plays the minimax solution.
Firstly, RP estimates the preference matrix by the mean of its posterior distribution. For each arm RP can play, find the corresponding arm CP would play to minimise RP's payoff. This forms an active set of arm pairs.
Secondly, now using a Thompson sample estimate for the preference probabilities, RP chooses the arm that maximises its payoff from the active set

Column player picks second arm.
Plays best response.

Altering the column players move

The payoff matrix was formed such that the same arm will never be chosen for both arms in a pair. If the optimistic corresponding to the chosen pair exceeds 0.5, it means the row player believes they will receive a positive payoff.
This is assumed to be for 1 of 2 reasons.
1) The row player has played the Condorcet winner, in which case the column player should also play the same arm.
2) The preference matrix is poorly estimated. In this case we ensure the column player explores arms.

Cumulative regret upper-bound

We have proved a problem-dependent bound on the cumulative regret of

$$R_T = O\left(\frac{K^2}{\Delta}\log T + K^3\right)$$

This is a factor of K from the optimal.

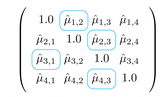
We conjecture that the algorithm has a–dependent bound on the cumulative regret of

$$R_T = O\left(\frac{K}{\Delta}\log T + K\right)$$

Algorithm and example

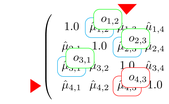
Algorithm 1 Max-Min Thompson Sampling

for $t = 1, \ldots, T$ **do**
  Calculate expected payoff, $\hat{\mu}_{i,j} = \frac{n^t_{i,j}+1}{n^t_{i,j}+n^t_{j,i}+2}$, for $i, j \in \{1, \ldots, K\}$, for $j \neq i$.
  Adjust payoffs on diagonal for exploration, $\hat{\mu}_{i,i} = 1$, for $i \in \{1, \ldots, K\}$.
  Row player finds minimum payoff for each action i, $m_i = \text{argmin}_j\, \hat{\mu}_{i,j}$ for $i \in \{1, \ldots, K\}$.
  Sample $\theta_{i,m_i} \sim \text{Beta}(n^t_{i,m_i}+1, n^t_{m_i,i}+1)$,
  Row player forms optimistic sample for minimum payoffs, $o_{i,m_i} = \max(\theta_{i,m_i}, \hat{\mu}_{i,m_i})$.
  Row player chooses max-min action, $s = \text{argmax}_i\, o_{i,m_i}$ for $i \in \{1, \ldots, K\}$.
  Let $a_1 = s$.
  if $\hat{\mu}_{s,m_s} > 0.5$ **then**
    if $\exists a_{\min} = \text{argmin}_i\, n^t_{s,i} + n^t_{i,s}$ s.t. $n^t_{s,i} + n^t_{i,s} < \frac{\log T}{(0.5-\hat{\mu}_{s,i})^2}$ **then**
      Column player plays exploratory action $a_2 = a_{\min}$
    else
      Column player plays exploitative action $a_2 = s$.
    end if
  else
    Column player plays exploratory action $a_2 = m_s$.
  end if
  Pair $c(t) = (\, a_1 \,, a_2)$.
end for

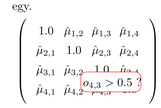


Line 4: For each row pick the column with lowest posterior mean. "Min" stage of row players minimax strategy.

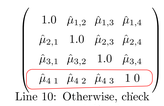
Line 5 and 6: Draw optimistic sample estimates for chosen entries and choose row with largest sample. "Max" stage of row players strategy.



Line 7: Select the entry with the largest sample estimate. Best response for Column player. Here s=4 and m_s=3.



Line 9: Now check the value of the optimistic value. If it is less than 0.5, then pull (s, m_s). Here s=4 and m_s=3.



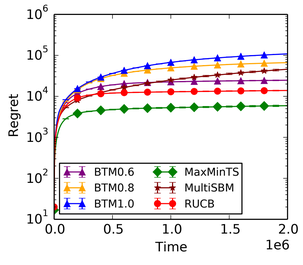
Line 10: Otherwise, check row s to check if pairs on that row have been pulled enough. If they have, pull (s, s), otherwise pull the least-pulled such pair.

Empirical evaluation

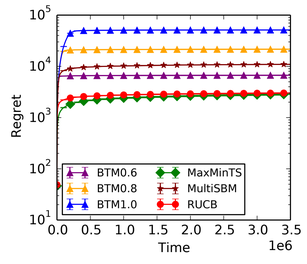
Our proposed algorithm, MaxMin Thompson Sampling (MaxMinTS) was evaluated via simulations drawn from 3 different distributions of preference matrix. The algorithm was compared to other algorithms suitable for the duelling-bandit problem. These were, RUCB, MultiSBM and Beat-the-mean (BTM).

The first set of simulations were based on the LETOR dataset. This dataset is based on an information retrieval application in which the bandit algorithm is used to find the best ranker from a set of search engine ranking strategies (such as PageRank).
The second set of simulations were based on an artificially-created environment. The goal was to test the algorithm in environments where the transitivity of preference was not preserved and instead there were cycles in the preferences. For example, such an environment could have A preferred to B, B preferred to C and C preferred to A, as long as there was one Condercet winner.
The final set of simulations were based on another artificially-created environment. The goal was to test the algorithm in an environment where there was no preference between most of the arms (there was a draw between the arms).

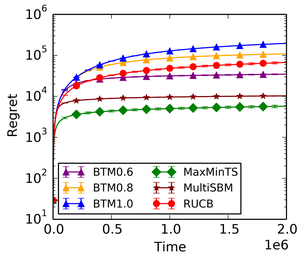
LETOR simulation



Cycles simulation



Draws simulation



Discussion

We have presented a new algorithm for the duelling-bandit problem. It does not require knowledge of the time-horizon or have any parameters to tune. The algorithm was inspired by a game-theoretic view of the problem. We believe proposing the problem in game-theoretic terms provides an interesting framework with which to design algorithms for variations of the duelling-bandit problem, for which our own proposal is just one.

Our regret bound is not optimal by a factor K. Our empirical results would suggest that the algorithm is as good as, or better, than algorithms for which optimal regret bounds have been established. Further work includes resolving this disparity. We conjecture that it is our regret bound which can be further improved.

References

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:pp. 285–294, 1933. ISSN 00063444. URL http://www.jstor.org/stable/2332286.

Y. Yue, J. Broder, J. Kleinberg, and T. Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Masrour Zoghi, Shimon Whiteson, Rémi Munos, and Maarten de Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. *CoRR*, abs/1312.3393, 2013.

Nir Ailon, Zohar Shay Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Proceedings*, pages 856–864. JMLR.org, 2014. URL http://jmlr.org/proceedings/papers/v32/ailon14.html.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 2249–2257, 2011.