Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-armed Bandit Problem with Multiple Plays

Junpei Komiyama, Junya Honda, Hiroshi Nakagawa (U-Tokyo)
at ICML2015, Lille, France.
## Optimal algorithms in MAB

List of optimal algos in the multi-armed bandits (MABs).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Optimal in Single-play MAB?</th>
<th>Optimally extendable in Multiple-play MAB?</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMED [Honda+ 2010]</td>
<td>Yes</td>
<td>?</td>
</tr>
<tr>
<td>Thompson sampling [Thompson 1930]</td>
<td>Yes</td>
<td>?</td>
</tr>
</tbody>
</table>
# Optimal algorithms in MAB

List of optimal algos in the multi-armed bandits (MABs).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Optimal in Single-play MAB?</th>
<th>Optimally extendable in Multiple-play MAB?</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMED [Honda+ 2010]</td>
<td>Yes</td>
<td>?</td>
</tr>
<tr>
<td>Thompson sampling [Thompson 1930]</td>
<td>Yes</td>
<td>Yes! (This paper)</td>
</tr>
</tbody>
</table>
Optimality in Multiple-play MAB

- Many algos optimal in Single-play MAB.
- The analysis in SP-MAB is still not sufficient in analyzing MP-MAB.
- TS has a good property in reducing regret in MP-MAB.
Main result: optimal regret of MP-TS

Multiple-play (MP) MAB: $L \geq 1$ arms among $K$ are selected at each round.

Thm: the regret of the multiple-play extension of Thompson sampling (MP-TS) is bounded as

$$\text{Regret}(T) \leq \sum_{i \in \{L+1, \ldots, K\}} \left( \frac{(\mu_L - \mu_i) \log T}{D_{KL}(\mu_i, \mu_L)} \right) + o(\log T).$$

- The $\log T$ factor matches the regret LB proven by Anantharam et al. [1987].
- MP-TS has optimal regret in MP-MAB.
Outline

- Setup: MP-MAB
  - Evaluation metric: Regret
- Multiple-play Thompson sampling (MP-TS)
- Regret lower bound
  - Decompose into two parts: factor in common with single-play / characteristic in MP.
- Regret of MP-TS
  - Satisfy both requirements for optimality.
- Experiment
- Conclusion / discussion
Input: $K$ arms indexed as $[K] = \{1, 2, \ldots, K\}$

At each round $t = 1, 2, \ldots, T$,

- Forecaster selects set of $L$ arms $I(t) \subset [K]$.
- Receives reward $X_i(t)$ of each arm $i \in I(t)$ drawn Bernoulli($\mu_i$).

Goal: maximize rewards $\sum_{t=1}^{T} \sum_{i \in I(t)} X_i(t)$. 

**MP-MAB**
Evaluation metric: Regret

- Assume $1 > \mu_1 > \mu_2 > \cdots > \mu_L > \cdots > \mu_K > 0$.
- Call arms $[L] = \{1, \ldots, L\}$ optimal, and arms $\{L + 1, \ldots, K\}$ suboptimal.
- Regret in MP-MAB:

\[
\text{Regret}(T) = \sum_{t=1}^{T} \left( \sum_{j \in [L]} \mu_j - \sum_{i \in I(t)} \mu_i \right).
\]

- Regret increases unless $I(t) = \{1,2, \ldots, L\}$.
- “Find L-best set as fast as possible”.

8
Algorithm: MP-TS

Algorithm 1 Multiple-play Thompson sampling (MP-TS) for binary rewards

Input: # of arms $K$, # of selection $L$

for $i = 1, 2, \ldots, K$ do
    $A_i, B_i = 1, 1$
end for

t ← 1.

for $t = 1, 2, \ldots, T$ do
    for $i = 1, 2, \ldots, K$ do
        $\theta_i(t) \sim \text{Beta}(A_i, B_i)$
    end for
    $I(t) =$ top-$L$ arms ranked by $\theta_i(t)$.
    for $i \in I(t)$ do
        if $X_i(t) = 1$ then
            $A_i \leftarrow A_i + 1$
        else
            $B_i \leftarrow B_i + 1$
        end if
    end for
end for
Regret lower bound

Regret LB [Anantharam et al. 1987]: the regret of a strongly consistent algorithm is lower-bounded as:

\[
\text{Regret}(T) \geq \sum_{i \in \{L+1, \ldots, K\}} \frac{(\mu_L - \mu_i) \log T}{D_{KL}(\mu_i, \mu_L)} - o(\log T)
\]

- An extension of Lai and Robbins [1985] result to \( L > 1 \).

Question: how can MP-TS achieve this bound?
Regret lower bound

\[
\text{Regret}(T) \geq \sum_{i \in \{L+1, \ldots, K\}} \frac{(\mu_L - \mu_i) \log T}{D_{KL}(\mu_i, \mu_L)} - o(\log T).
\]

- \(\frac{\log T}{D_{KL}(\mu_i, \mu_L)}\): # of samples (for making sure that arm \(i\) is suboptimal).

- \((\mu_L - \mu_i)\): minimum regret increase per selecting arm \(i\): characteristic in MP.

- Optimal algorithm: both blue and green parts must be optimal.
Blue part: consistency

- Good algorithm = algorithm that works with any set of arms.
  - An algorithm is strongly consistent if for any arm set and $a > 0$, $\text{Regret}(T) = o(T^a)$.
  - In MP-MAB, the strong consistency requires each suboptimal arm $i > \{L + 1, ..., K\}$ to be drawn
    \[
    \log T \quad \frac{\log T}{D_{KL}(\mu_i, \mu_L)}
    \]
    times.

($D_{KL} = the \ KL \ divergence \ between \ two \ Bernoulli \ dists$)
Q. Does MP-TS has *optimal number of draws on suboptimal arms*?

A. Yes!

Based on existing techniques of TS for single-play MAB [Agrawal & Goyal 2012, 2013, Honda & Takemura 2014], MP-TS is optimal in the blue part.
Green part: regret per exploration

\[
\text{Regret}(T) \geq \sum_{i \in \{L+1, \ldots, K\}} \frac{(\mu_L - \mu_i) \log T}{D_{KL}(\mu_i, \mu_L)} - o(\log T) .
\]

- \(\mu_L - \mu_i\): difference of expected rewards of arm \(L\) (optimal arm with smallest expectation) and \(i\) (suboptimal arm).
- Push out arm \(L\) when we explores arm \(i\).
Optimality in **green** part implies: an algorithm cannot select two (or more) suboptimal arms in the same round more than $O(\log T)$ times.

$(\mu_2 > \mu_L$: more regret per an exploration, if we push out $\mu_2$)
Green part: regret per exploration

- The necessary condition for an optimal (green part): # of simultaneous draw of two suboptimal arms is $o(\log T)$.
- We show that MP-TS satisfies this condition.
Algorithm: MP-TS

**Algorithm 1** Multiple-play Thompson sampling (MP-TS) for binary rewards

Input: # of arms $K$, # of selection $L$

```plaintext
for $i = 1, 2, \ldots, K$ do
    $A_i, B_i = 1, 1$
end for

$t \leftarrow 1.$

for $t = 1, 2, \ldots, T$ do
    for $i = 1, 2, \ldots, K$ do
        $\theta_i(t) \sim \text{Beta}(A_i, B_i)$
    end for

$I(t) =$ top-$L$ arms ranked by $\theta_i(t)$.

for $i \in I(t)$ do
    if $X_i(t) = 1$ then
        $A_i \leftarrow A_i + 1$
    else
        $B_i \leftarrow B_i + 1$
    end if
end for
```

Key property: Sample of each arm is independently drawn from associated posterior distribution.
Analysis: two suboptimal arms are rarely drawn simultaneously

- MP-TS draws a suboptimal arm with prob $O \left( \frac{1}{t} \right)$.
- prob that two suboptimal arms are simultaneously drawn is $O \left( \frac{1}{t} \times \frac{1}{t} \right) = O \left( \frac{1}{t^2} \right)$, and $\sum_{t=1}^{T} O \left( \frac{1}{t^2} \right) = O(1) = o(\log T)$.

- MP-TS is optimal in green part.
- As MP-TS is optimal in both blue and green parts, its regret is optimal.
Optimality of other algorithms.

- KL-UCB [Lai 1987, Garivier+ 2011] is optimal in single-play MAB.
- Q. is extension of KL-UCB for MP-MAB is also optimal?
- A. probably it is, but hard to prove since the UCB score of each arm is dependent each other.
Experimental result 1: 20-armed bandits, L=3, synthetic arms

- 20-armed bandits, L=3, synthetic arms

- Dotted line: log(T) term of regret lower bound

- MP extension of KL-UCB

- MP-TS (proposed)

- Slope of MP-TS quickly converges to the one. → optimal regret
100-armed MAB with $L = 3$, parameters $\{\mu_i\}$ are estimated from real search engine ads (soso.com: Tencent)

Experiment result 2: search engine ads

MP extension of KL-UCB
MP-TS (proposed)

About a half regret, compared to KL-UCB
Conclusion and discussion

- MP-TS is optimal in MP-MAB, and empirically outperforms the other algos.
- Two bad events rarely occur simultaneously in TS.
- The independence of posterior samples can also be useful for more general combinatorial bandit problems.

Thx for listening!