Thompson Sampling: an asymptotically optimal finite-time analysis

Emilie Kaufmann, Nathaniel Korda and Rémi Munos

ALT, October 30th, 2012
1 The multi-armed bandit problem

2 From UCB to Thompson Sampling

3 Finite-time analysis of Thompson Sampling

4 A closer look at the fundamental deviation result

5 Some perspectives
The multi-armed bandit problem

The stochastic MAB with Bernoulli rewards

$K$ independent arms.

- $\mu_1, \ldots, \mu_K$ unknown parameters
- $Y_{a,t}$ is i.i.d. with distribution $B(\mu_a)$

The parameter of the best arm is $\mu^* = \max_{a=1}^K \mu_a$

- At time $t$, the forecaster chooses arm $A_t$ and gets reward $R_t = Y_{A_t,t}$.
- Goal: Design a strategy $A_t$ minimizing the cumulative regret:

$$R(T) := T\mu^* - \mathbb{E} \left[ \sum_{t=1}^T R_t \right] = \sum_{a \in A} (\mu^* - \mu_a) \mathbb{E}[N_{a,T}]$$
Asymptotically optimal bandit algorithms

- Lai and Robbins’ lower bound on the regret of a consistent policy:

\[ \mu_a < \mu^* \implies \liminf_{T \to \infty} \frac{\mathbb{E}[N_{a,T}]}{\ln T} \geq \frac{1}{K(\mu_a, \mu^*)} \]

or equivalently

\[ \liminf_{T \to \infty} \frac{\mathbb{E}[R(T)]}{\ln(T)} \geq \sum_{a : \mu_a < \mu^*} \frac{\mu^* - \mu_a}{K(\mu_a, \mu^*)} \]

with

\[ K(p, q) := p \ln \frac{p}{q} + (1 - p) \ln \frac{1 - p}{1 - q}. \]

- A bandit algorithm is **asymptotically optimal** if

\[ \mu_a < \mu^* \implies \limsup_{T \to \infty} \frac{\mathbb{E}[N_{a,T}]}{\ln T} \leq \frac{1}{K(\mu_a, \mu^*)} \]
1. The multi-armed bandit problem

2. From UCB to Thompson Sampling

3. Finite-time analysis of Thompson Sampling

4. A closer look at the fundamental deviation result

5. Some perspectives
Some successful frequentist algorithms

A family of **optimistic index policies** based on an **upper confidence bound** for the empirical mean of the rewards:

- **UCB** [Auer et al. 02] and variants:

\[
\mathbb{E}[N_{a,T}] \leq \frac{K_1}{2(\mu_a - \mu^*)^2} \ln T + K_2, \quad \text{with } K_1 > 1.
\]

- **KL-UCB** [Cappé, Garivier, Maillard, Stoltz, Munos 11] uses the index:

\[
u_{a,t} = \arg\max_{x > \frac{S_{a,t}}{N_{a,t}}} \left\{ K \left( \frac{S_{a,t}}{N_{a,t}}, x \right) \leq \frac{\ln(t) + c \ln \ln(t)}{N_{a,t}} \right\}
\]

For all \( \epsilon > 0 \), there exists a constant \( K_\epsilon \) such that:

\[
\mathbb{E}[N_{a,T}] \leq \frac{1 + \epsilon}{K(\mu_a, \mu^*)} \ln T + K_\epsilon
\]
Imagine we are given independent priors on the parameters of each arm:

- \( \mu_a \overset{i.i.d.}{\sim} \mathcal{U}([0, 1]) \)
- \((Y_{a,t})_t\) is i.i.d. conditionally to \( \mu_a \) with distribution \( \mathcal{B}(\mu_a) \)
- The posterior on arm \( a \) at time \( t \) is
  \[
  \pi_{a,t} = \text{Beta} \left( S_{a,t} + 1, N_{a,t} - S_{a,t} + 1 \right).
  \]

**Bayesian algorithms** uses this posterior \( \pi_{a,t} \) to choose \( A_t \).

\( \Rightarrow \) We still focus on frequentist guarantees (asymptotic optimality) for Bayesian algorithms.
A Bayesian Upper Confidence Bound algorithm

- Bayes-UCB [Kaufmann et al. 12] is the index policy associated with

\[ q_{a,t} := Q \left( 1 - \frac{1}{t \ln(t)c}, \pi_{a,t} \right) \]

This Bayesian algorithm is asymptotically optimal

**Figure:** UCB versus Bayes-UCB
Thompson Sampling: a new kind of optimism?

A very simple algorithm:

\[ \forall a \in \{1..K\}, \quad \theta_{a,t} \sim \pi_{a,t} \]

\[ A_t = \arg\max_{a} \theta_{a,t} \]

Recent interest for this algorithm:

- Partial analysis proposed
  [Granmo 2010][May, Korda, Lee, Leslie 2011]

- Extensive numerical study beyond the Bernoulli case
  [Chapelle, Li 2011]

- First logarithmic upper bound on the regret
  [Agrawal, Goyal 2012]
1 The multi-armed bandit problem

2 From UCB to Thompson Sampling

3 Finite-time analysis of Thompson Sampling

4 A closer look at the fundamental deviation result

5 Some perspectives
An optimal regret bound for Thompson Sampling

Assume the first arm is the unique optimal and \( \Delta_a = \mu_1 - \mu_a \).

- **Known result**: [Agrawal, Goyal, 2012]

\[
\mathbb{E}[\mathcal{R}(T)] \leq C \left( \sum_{a=2}^{K} \frac{1}{\Delta_a} \right) \ln(T) + o_{\mu}(\ln(T))
\]
An optimal regret bound for Thompson Sampling

Assume the first arm is the unique optimal and $\Delta_a = \mu_1 - \mu_a$.

- **Known result**: [Agrawal, Goyal, 2012]

  $$\mathbb{E}[R(T)] \leq C \left( \sum_{a=2}^{K} \frac{1}{\Delta_a} \right) \ln(T) + o_{\mu}(\ln(T))$$

- **Our improvement**:

  **Theorem 2** $\forall \epsilon > 0$,

  $$\mathbb{E}[R(T)] \leq (1 + \epsilon) \left( \sum_{a=2}^{K} \frac{\Delta_a}{K(\mu_a, \mu^*)} \right) \ln(T) + o_{\mu, \epsilon}(\ln(T))$$
Step 1: Decomposition

- We adapt an analysis working for optimistic index policies:

\[ A_t = \arg\max_a l_{a,t} \]

\[
E[N_{a,T}] \leq \sum_{t=1}^{T} \Pr(l_{1,t} < \mu_1) + \sum_{t=1}^{T} \Pr(l_{a,t} \geq l_{1,t} > \mu_1, A_t = a)
\]

\[ o(\ln(T)) \quad \text{and} \quad \ln(T)/K(\mu_a, \mu_1) + o(\ln(T)) \]
Step 1: Decomposition

- We adapt an analysis working for optimistic index policies:

\[ A_t = \arg\max_a l_{a,t} \]

\[ \mathbb{E}[N_{a,T}] \leq \sum_{t=1}^{T} \mathbb{P}(l_{1,t} < \mu_1) + \sum_{t=1}^{T} \mathbb{P}(l_{a,t} \geq l_{1,t} > \mu_1, A_t = a) \]

\( o(\ln(T)) \quad \text{and} \quad \ln(T)/K(\mu_a, \mu_1) + o(\ln(T)) \)

⇒ Does NOT work for Thompson Sampling
Step 1: Decomposition

- We adapt an analysis working for optimistic index policies:

\[ A_t = \arg\max_a l_{a,t} \]

\[
\mathbb{E}[N_{a,T}] \leq \sum_{t=1}^{T} \mathbb{P}(l_{1,t} < \mu_1) + \sum_{t=1}^{T} \mathbb{P}(l_{a,t} \geq l_{1,t} > \mu_1, A_t = a) \]

\[ o(\ln(T)) \quad \text{and} \quad \ln(T)/K(\mu_a, \mu_1) + o(\ln(T)) \]

⇒ Does NOT work for Thompson Sampling

- Our decomposition for Thompson Sampling is

\[
\mathbb{E}[N_{a,T}] \leq \sum_{t=1}^{T} \mathbb{P} \left( \theta_{1,t} \leq \mu_1 - \sqrt{\frac{6\ln t}{N_{1,t}}} \right) + \sum_{t=1}^{T} \mathbb{P} \left( \theta_{a,t} > \mu_1 - \sqrt{\frac{6\ln t}{N_{1,t}}}, A_t = a \right) \]

\[ (*) \]
Step 2: Linking quantiles to other known indices

We introduce the following quantile:

\[ q_{a,t} := Q \left( 1 - \frac{1}{t \ln(T)}, \pi_{a,t} \right) \]
Step 2: Linking quantiles to other known indices

- We introduce the following quantile:

\[ q_{a,t} := Q \left( 1 - \frac{1}{t \ln(T)}, \pi_{a,t} \right) \]

- And the corresponding KL-UCB index

\[ u_{a,t} := \arg \max_{x > \frac{S_{a,t}}{N_{a,t}}} \left\{ K \left( \frac{S_{a,t}}{N_{a,t}}, x \right) \leq \frac{\ln(t) + \ln(\ln(T))}{N_{a,t}} \right\} \]
Step 2: Linking quantiles to other known indices

We introduce the following quantile:

\[ q_{a,t} := Q \left( 1 - \frac{1}{t \ln(T)}, \pi_{a,t} \right) \]

And the corresponding KL-UCB index

\[ u_{a,t} := \arg\max_{x > \frac{S_{a,t}}{N_{a,t}}} \left\{ K \left( \frac{S_{a,t}}{N_{a,t}}, x \right) \leq \frac{\ln(t) + \ln(\ln(T))}{N_{a,t}} \right\} \]

We know from previous work [Kaufmann et al.] that

\[ q_{a,t} < u_{a,t} \]
Step 2: Linking quantiles to other known indices

- Introducing the quantile $q_{a,t}$:

$$\sum_{t=1}^{T} \mathbb{P} \left( \theta_{a,t} > \mu_1 - \sqrt{\frac{6 \ln t}{N_{1,t}}}, A_t = a \right)$$

$$\leq \sum_{t=1}^{T} \mathbb{P} \left( q_{a,t} > \mu_1 - \sqrt{\frac{6 \ln t}{N_{1,t}}}, A_t = a \right) + \underbrace{\sum_{t=1}^{T} \mathbb{P} (\theta_{a,t} > q_{a,t})}_{\leq 2}$$
Step 2: Linking quantiles to other known indices

- Introducing the quantile $q_{a,t}$:

  $$
  \sum_{t=1}^{T} \mathbb{P} \left( \theta_{a,t} > \mu_1 - \sqrt{\frac{6\ln t}{N_{1,t}}}, A_t = a \right) 
  \leq 
  \sum_{t=1}^{T} \mathbb{P} \left( q_{a,t} > \mu_1 - \sqrt{\frac{6\ln t}{N_{1,t}}}, A_t = a \right) + \sum_{t=1}^{T} \mathbb{P} \left( \theta_{a,t} > q_{a,t} \right) 
  \leq 2
  $$

- Then the KL-UCB index $u_{a,t}$:

  $$
  \sum_{t=1}^{T} \mathbb{P} \left( \theta_{a,t} > \mu_1 - \sqrt{\frac{6\ln t}{N_{1,t}}}, A_t = a \right) 
  \leq 
  \sum_{t=1}^{T} \mathbb{P} \left( u_{a,t} > \mu_1 - \sqrt{\frac{6\ln t}{N_{1,t}}}, A_t = a \right) + 2
  $$
The final decomposition is:

\[
\mathbb{E}[N_{a,t}] \leq \sum_{t=1}^{T} \mathbb{P} \left( \theta_{1,t} \leq \mu_1 - \sqrt{\frac{6 \ln t}{N_{1,t}}} \right) + \sum_{t=1}^{T} \mathbb{P} \left( u_{a,t} > \mu_1 - \sqrt{\frac{6 \ln t}{N_{1,t}}}, A_t = a \right) + 2
\]
Step 3: One extra ingredient for bounding term $A$ and $B$

- We state a fundamental deviation result:

**Proposition 1** There exists constants $b = b(\mu_1, \mu_2) \in (0, 1)$ and $C_b < \infty$ such that:

$$\sum_{t=1}^{\infty} \mathbb{P} \left( N_{1,t} \leq t^b \right) \leq C_b.$$
1. The multi-armed bandit problem

2. From UCB to Thompson Sampling

3. Finite-time analysis of Thompson Sampling

4. A closer look at the fundamental deviation result

5. Some perspectives
A closer look at the fundamental deviation result

Understanding the deviation result

- Recall the result

There exist constants $b = b(\mu_1, \mu_2) \in (0, 1)$ and $C_b < \infty$ such that

$$\sum_{t=1}^{\infty} \mathbb{P}(N_{1,t} \leq t^b) \leq C_b.$$  

- Where does it come from?

$$\{N_{1,t} \leq t^b\} = \{\text{there exists a time range of length at least } t^{1-b} - 1 \text{ with no draw of arm 1}\}$$
A closer look at the fundamental deviation result.

Assume that:

- on $\mathcal{I}_j = [\tau_j, \tau_j + [t^{1-b} - 1]]$ there is no draw of arm 1
- there exists $\mathcal{J}_j \subset \mathcal{I}_j$ such that $\forall s \in \mathcal{J}_j, \forall a \neq 1, \theta_{a,s} \leq \mu_2 + \delta$

Then:

- $\forall s \in \mathcal{J}_j, \theta_{1,s} \leq \mu_2 + \delta$

$\Rightarrow$ This only happens with small probability.
1. The multi-armed bandit problem

2. From UCB to Thompson Sampling

3. Finite-time analysis of Thompson Sampling

4. A closer look at the fundamental deviation result

5. Some perspectives
Conclusion and perspectives

Thompson Sampling in the Bernoulli setting:
- has the same theoretical guarantees than known optimal algorithms (KL-UCB, Bayes-UCB)
- and displays excellent empirical performance

The proof we give:
- is close to the analysis of optimistic bandit algorithms
- also gives a deviation result on the number of draws of optimal arms

Can Thompson Sampling be extended to more general settings?
- Contextual bandit ([Agrawal, Goyal, Thompson Sampling for Contextual Bandits with Linear Payoffs, sept 2012])
- Model-based Bayesian reinforcement learning
Figure: Regret as a function of time (on a log scale) for a 10 arms problem

**Thompson Sampling outperforms other optimal algorithms**
Any question?