Solving Multi-Armed Bandits Problem by Upper Confidence Bound Algorithms

Fan-Keng Sun, Chen-Hao Hsiao, Chi-Hsin Lo

Department of Electrical Engineering, National Taiwan University

7/5/2018

Table of Contents

Table of Contents

Introduction to Online Learning

- Online learning is a method in machine learning to solve problems where data flows in a sequential manner: a learner predicts the future by interacting with the environment online.
- Compares to batch learning:

	online learning	batch learning
source of dataset	online (function of time)	offline (fixed)
target to minimize	cumulative regret on sequence	fixed loss on dataset
weight of loss	dynamic	static
number of shots	one or few shot	multi-shot
require of prediction	on-the-fly	after training

- The reason to use online learning instead of batch learning:
  1. There is too much data to learn.
  2. Algorithm need to dynamically adapt to new patterns in data.
  3. It is simple and fast.
  4. It has theoretical guarantees.

Online Learning v.s. Reinforcement Learning

- The sequential and online characteristic of online learning is similar to reinforcement learning.
- But there are still some differences:

	online learning	reinforcement learning
data distribution	no guarantee	Markov process
environment	dynamic (adversarial)	static
reward or ground-truth	every step	no guarantee
ending	endless or predefined	agent met specific state
target at each step	minimize cumulative regret	maximive the future reward

Applications of Online Learning

- In modern world, there are many applications of online learing.
- For examples:
  - Companies maximizing advertisement click-through rate.
  - Investors making prediction of stock market.
  - Tools classifying spam emails.
  - Meteorologists forecasting future weather.
  - Website clustering users in a social network.
- Almost all applications have to deal with the impact of time (i.e. time changes everything).

Table of Contents

Introduction to Multi-Armed Bandits (MAB) Problem

- Imagine there is a gambler standing in front of a row of slot-machines (one-amred bandits). In every trial, the gambler can only pull one arm and observe its result. Now, the gambler wants to optimize the total reward in T pulls.
- Formally speaking, the gambler (learner) is going through the following process trying to maximize total reward:

Solving MAB problems for K-amrs in T trials

for $t := 1$ **to** T **do**
  Learner pulls an arm $i_t \in [K]$.
  Environment selects a reward vector $\boldsymbol{x_t} = \{x_{1,t}, \ldots, x_{K,t}\}$.
  Learner observes $x_{i_t,t}$.
end

Introduction to Multi-Armed Bandits (MAB) Problem

- Special of MAB: learner has only partial observation (only $x_{i_t,t}$).
- The original motivation of [8] for studying MAB came from clinical trials, where doctors will have to choose one medicine for a client and then observe the effects. However, in modern world, there are much more applications.
- There are three formulations of MAB problems:
  1. Stochastic: Each arm is assumed to be represented by a fixed (but unknown) probability distribution, and reward are drawn i.i.d. from the distribution. Effective strategy is UCB and its variants.
  2. Adversarial (non-stochastic): No probabilistic assumptions can be made on any arms. Effective strategy is the Exp3 randomized.
  3. Markovian: Reward from each arm follows a Markov process. Effective strategy is Gittins indices.
- We will be concentrated on stochastic MAB only.

Stochastic Multi-Armed Bandits (MAB) Problem

- The learner faces exploration v.s. exploitation dilemma: shall it continue to select the best arm observed so far (exploitation) or rather probe other arms further (exploration):
  - Exploitation takes the risk that its observation is inaccurate.
  - Exploration may be just a waste of time.
- Good algorithms must find a good balance between exploration and exploitation.

Problem Formulation of Stochastic MAB Problem

- Stochastic MAB problem is parameterized by the following:
  1. $A = \{1, \ldots, K\}$: the set of arms, and $K = |A|$ is the number of arms.
  2. $\boldsymbol{X} = \{X_1, \ldots, X_K\}$: X_i is i.i.d random variable representing the reward from arm i, and distinct arms are independent. $X_i \in [0, 1]$ if not specified.
  3. $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_K\} : \mu_i = \mathbb{E}[X_i]$ is the mean reward of arm i, and $\mu^* = \max\{\boldsymbol{\mu}\}$ is the best mean reward, $i^* \in \{i | \mu_i = \mu^*\}$ is any optimal arm.
  4. T: the number of trials (or budgets).

Problem Formulation of Stochastic MAB Problem

- The goal of the learner is to minimize the (cumulative) regret R_T after T trials, which is defined as:

$$R_T = \mu^* T - \sum_{i=1}^{K} \mu_i N_{i,T},$$

where $N_{i,T}$ is the number of times the learner has chosen arm i up to trial T. The expected (cumulative) regret after trial T can be written as:

$$\mathbb{E}[R_T] = \sum_{i=1}^{K} \Delta_i \, \mathbb{E}[N_{i,T}]$$

where $\Delta_i = \mu^* - \mu_i$ is the gap between the means of optimal arm and the chosen arm.

Table of Contents

Upper Confidence Bound (UCB)

- How to balance between exploration and exploitation? Something better that ϵ-greedy?
- Upper confidence bound (UCB) [4] and its variants are the most effective strategies nowadays.
- The first UCB-based algorithm: UCB1

UCB1 algorithm

for $t := 1$ **to** T **do**
  Arm selection:
  If $t \leq K$, select arm $i_t = t$. (Initialization)
  If $t > K$, select arm $i_t \in \arg\max_{i \in A}(\hat{\mu}_i^{t-1} + \sqrt{\frac{2 \log T}{N_{i,t-1}}})$.
end

where $\hat{\mu}_i^{t-1}$ is the empirical mean of reward.

Upper Confidence Bound (UCB)

- UCB is a strategy based on the "optimism in the face of uncertainty" principle.
- Intuitively, the term $(\hat{\mu}_i^{t-1} + \sqrt{\frac{2 \log T}{N_{i,t-1}}})$ is the upper bound of the reward from arm i with high confidence.
- Additionally, we can see that if $N_{i,t-1}$ is small, then we are not sure about arm i, thus arm i has larger upper bound.
- Formally, by Chernoff-Hoeffding inequality, we have

$$\mathbb{P}\left\{\hat{\mu}_i^{t-1} - \mu_i \geq \sqrt{\frac{2 \log T}{N_{i,t-1}}}\right\} \leq \exp\left(-\frac{2 \cdot N_{i,t-1}^2 \cdot \frac{2 \log T}{N_{i,t-1}}}{N_{i,t-1}}\right) = T^{-4} \tag{1}$$

Upper Confidence Bound (UCB)

- We proof the following lemma to further show the intuition formally.

Lemma 1

Fix any $i : \Delta_i > 0$. If UCB1 algorithm selects the arm i in trial t ($i_t = i$), then at least one of the following holds:

(1) $\hat{\mu}_{i^*}^{t-1} \leq \mu^* - \sqrt{\frac{2 \log t}{N_{i^*,t-1}}}$, *(2)* $\hat{\mu}_i^{t-1} \geq \mu_i + \sqrt{\frac{2 \log t}{N_{i,t-1}}}$, *or (3)* $N_{i,t-1} \leq \frac{8 \log T}{\Delta_i^2}$

- The lemma indicates that if we pull arm i in trial t, either we have
  1. a bad confidence interval constructed for any optimal arm.
  2. a bad confidence interval constructed for the pulled arm.
  3. not pulled the arm sufficiently many times.

Upper Confidence Bound (UCB)

- **Proof of lemma 1:** Prove by contradiction. If (1), (2), and (3) are all false, we have:

$$\begin{aligned}
\hat{\mu}_{i^*}^{t-1} + \sqrt{\frac{2\log t}{N_{i^*,t-1}}} &> \mu^* && \text{since (1) is false} \\
&= \mu_i + \Delta_i \\
&> \mu_i + \sqrt{\frac{8\log T}{N_{i,t-1}}} && \text{since (3) is false} \\
&\geq \mu_i + \sqrt{\frac{8\log t}{N_{i,t-1}}} \\
&> \hat{\mu}_i^{t-1} + \sqrt{\frac{2\log t}{N_{i,t-1}}}, && \text{since (2) is false}
\end{aligned}$$

which contradicts the fact that $i_t = i$.

Upper Confidence Bound (UCB)

Theorem 2

The upper bound of total expected regret of the UCB1 algorithm until trail T is

$$8 \sum_{i \in A, \Delta_i > 0} \left(\frac{\log T}{\Delta_i} \right) + \left(1 + \frac{\pi^2}{3} \right) \sum_{i=1}^{K} \Delta_i \tag{2}$$

Proof.

$$\mathbb{E}[N_{i,T}] = \mathbb{E}[1 + \sum_{t=K+1}^{T} \mathbb{1}\{i_t = i\}] \leq \mathbb{E}[t_0 + \sum_{t=K+1}^{T} \mathbb{1}\{i_t = i, N_{i,t-1} \geq t_0\}]$$

$$\leq t_0 + \sum_{t=K+1}^{T} \left(\mathbb{P}(\hat{\mu}_{i*}^{t-1} \leq \mu^* - \sqrt{\frac{2 \log t}{N_{i*,t-1}}}) + \mathbb{P}(\hat{\mu}_i^{t-1} \geq \mu_i - \sqrt{\frac{2 \log t}{N_{i,t-1}}}) \right)$$

$$\leq t_0 + \sum_{t=K+1}^{T} \left(\mathbb{P}(\bigcup_{s=1}^{t-1} \{\hat{\mu}_{i*}^{t-1} \leq \mu^* - \sqrt{\frac{2 \log t}{s}}\}) + \mathbb{P}(\bigcup_{s'=1}^{t-1} \{\hat{\mu}_i^{t-1} \geq \mu_i - \sqrt{\frac{2 \log t}{s'}}\}) \right)$$

$$\leq t_0 + \sum_{t=1}^{T} \sum_{s=1}^{t} \sum_{s'=1}^{t} \left(\mathbb{P}(\hat{\mu}_{i*}^{t-1} \leq \mu^* - \sqrt{\frac{2 \log t}{s}}) + \mathbb{P}(\hat{\mu}_i^{t-1} \geq \mu_i - \sqrt{\frac{2 \log t}{s'}}) \right)$$

Table of Contents

Comparison of UCB Algorithms

- Let $\Delta = \min_{i:\Delta_i>0} \Delta_i$, and $\sigma = \max_i \sigma(X_i)$.
- UCB1 is the earliest UCB-based algorithm with a regret upper bound of $O(\frac{K \log T}{\Delta})$
- Improved-UCB, proposed in [1], is a round-based (pulls all arms equal number of times in each round) variant of UCB1 that eliminated bad arms that has a regret upper bound of $O(\frac{K \log(T\Delta^2)}{\Delta})$, which is better than UCB1.
- UCB-Variance (UCBV) [7] is the first algorithm that utilizes the variance to compute the confidence intervals for each arm. It has a regret upper bound of $O(\frac{K\sigma^2 \log T}{\Delta})$.
- Recently, Efficient-UCBV (EUCBV) [6] combines the elimination method of improved-UCB and variance estimation of UCBV to get a regret upper bound of $O(\frac{K\sigma^2 \log \frac{T\Delta^2}{K}}{\Delta})$.

Improved-UCB Algorithm

$\tilde{\Delta}_0 := 1$, and $B_0 := A$
for round $m := 0$ **do**

Arm selection:

If $|B_m| > 1$, choose each arm in B_m until the total number of times it has been chosen is

$$n_m := \left\lceil \frac{2\log(T\tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} \right\rceil. \tag{3}$$

Otherwise choose the single arm in B_m until step T is reached.

Arm elimination:

Delete all arms i from B_m for which

$$\left(\hat{\mu}_i + \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n_m}}\right) < \max_{j\in B_m}\left\{\hat{\mu}_j - \sqrt{\frac{\log(T\tilde{\Delta}_m^2)}{2n_m}}\right\} \tag{4}$$

to obtain B_{m+1}.

Set $\tilde{\Delta}_{m+1}$:

$\tilde{\Delta}_{m+1} := \frac{\tilde{\Delta}_m}{2}$

end

Table of Contents

Definition: Consistent algorithm

We call a algorithm $\mathcal{A}$ a consistent algorithm if for any sub-optimal arm $i : \Delta_i > 0$ and any $a > 0$

$$\mathbb{E}[N_{i,T}] = o(T^a)$$

i.e. The algorithm eventually samples any sub-optimal arm a vanishing fraction of times.

Lai & Robbin's (1985)[3]

If algorithm $\mathcal{A}$ is consistent and reward $X_i \sim \text{Bernoulli}(\mu_i), \forall i$ then

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R_T[\mathcal{A}]]}{\log T} \geq \sum_{i:\Delta_i>0} \frac{\Delta_i}{KL(X_i, X^*)}$$

KL-UCB algorithm [2]

for $t := 1$ **to** T **do**

Arm selection:

If $t \leq K$, select arm $i_t = t$. (Initialization).

If $t > K$, select arm $i_t \in$

$$\arg\max_{i \in A} \max_{q \in [0,1]} \{ N_{i,t-1} \cdot \underbrace{d(\hat{\mu}_i^{t-1}, q)}_{d(p,q) = p \log_2 \frac{p}{q} + (1-p) \log_2 \frac{1-p}{1-q}} \leq \log t + 3 \log \log t \} .$$

end

Theorem 3

Using KL-UCB, for any sub-optimal arm $i : \Delta_i > 0$ and any $\epsilon > 0$:

$$\mathbb{E}[N_{i,T}] \leq \frac{\log T}{d(\mu_i, \mu^*)}(1+\epsilon) + C_1 \log \log T + \frac{C_2(\epsilon)}{n^{\beta(\epsilon)}},$$

where C_1 is a positive constant and $C_2(\epsilon) = O(\epsilon^{-2}), \beta(\epsilon) = O(\epsilon^2)$ are both positive function of ϵ.

Remarks

- Theorem 3 implies $\liminf_{T\to\infty} \frac{\mathbb{E}[N_{i,T}]}{\log T} \leq \frac{1}{d(\mu_i,\mu^*)}$, which shows that KL-UCB algorithm fits the lower bound when rewards are Bernoulli.
- By Pinsker's Inequality, $d(\mu_i, \mu^*) > 2(\mu_i - \mu^*)^2 = 2\Delta_i^2$, we can see that KL-UCB is strictly better than UCB1.
- Burnetas & Katehakis [5] derived a more general lower bound on consistent algorithms, and other variants of UCB algorithm are designed to approach the bound asymptotically.

References

[1] P. Auer and R. Ortner. "UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem". In: *Periodica Mathematica Hungarica* (2010), pp. 55–65.

[2] Garivier Aurélien and Cappé Olivier. "The KL-UCB algorithm for bounded stochastic bandits and beyond". In: *Proceedings of the 24th annual Conference On Learning Theory*. 2011, pp. 359–376.

[3] Lai Tze Leung and Robbins Herbert. "Asymptotically efficient adaptive allocation rules". In: *Advances in applied mathematics* 6.1 (1985), pp. 4–22.

[4] Auer P.; Cesa-Bianchi N.; and Fischer P. "Finite-time analysis of the multiarmed bandit problem". In: *Machine Learning* (2002), pp. 235–256.

[5] Burnetas Apostolos N. and Katehakis Michael N. "Optimal adaptive policies for sequential allocation problems". In: *Advances in Applied Mathematics* 17.2 (1996), pp. 122–142.

[6] Mukherjee S.; Naveen K.P.; Sudarsanam N.; and Ravindran B. "Efficient-UCBV: An Almost Optimal Algorithm Using Variance Estimates". In: *AAAI Conference on Artificial Intelligence* (2018).

[7] Audibert J.-Y.; Munos R.; and Szepesvari C. "Exploration–exploitation tradeoff using variance estimates in multi-armed bandits". In: *Theoretical Computer Science* (2009), pp. 1876–1902.

[8] W. Thompson. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". In: *Bulletin of the American Mathematics Society* (1933), pp. 285–294.