A Note on the Introduction to Bandit

Xuanfei Ren*

January 11, 2023

Contents

*University of Science and Technology of China; email: xuanfeiren@mail.ustc.edu.cn.

1 Introduction

This is my preparation before reading relevant literature and conducting scientific researchabout bandit. This note is for *Introduction to Multi-Armed Bandits*.

2 Probability

In this section, I will briefly note some basic probability theory knowledge, so as to facilitate the reference when reading literature.

2.1 Bayes

Lemma 2.1 (posterior for sufficient statistic)**.** *$X \sim f(x \mid \theta)$, if $T(X)$ is a sufficient statistic for θ, then*

$$\pi(\theta \mid x) = \tilde{\pi}(\theta \mid t) \tag{2.1}$$

1. $X \sim N(\theta, \sigma_S^2)$, in which σ_S^2 is known. θ has a prior $N(\mu_P, \sigma_p^2)$, then the posterior distribution of θ is:

$$\theta \sim N(\frac{\frac{x}{\sigma_S^2} + \frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_S^2} + \frac{1}{\sigma_P^2}}, (\frac{1}{\sigma_S^2} + \frac{1}{\sigma_P^2})^{-1}), \tag{2.2}$$

   which means

$$\begin{aligned} \mu &= \frac{\sigma_P^2}{\sigma_S^2 + \sigma_P^2} x + \frac{\sigma_S^2}{\sigma_S^2 + \sigma_P^2} \mu_P, \\ \sigma^2 &= \frac{\sigma_S^2 \sigma_P^2}{\sigma_S^2 + \sigma_P^2}. \end{aligned} \tag{2.3}$$

   Proof. $f(x \mid \theta) = \frac{1}{\sqrt{2\pi}\sigma_S^2} \exp\{-\frac{1}{2\sigma_S^2}(x-\theta)^2\}$

$$\begin{aligned} \pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{f(x)} &\propto f(x \mid \theta)\pi(\theta) \\ &= \exp\{-\frac{1}{2\sigma_S^2}(x-\theta)^2 - \frac{1}{2\sigma_P^2}(\theta - \mu_P)^2\} \\ &\propto \exp\{-\frac{1}{2}(\frac{1}{\sigma_S^2} + \frac{1}{\sigma_P^2})(\theta^2 - 2\frac{\frac{x}{\sigma_S^2} + \frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_S^2} + \frac{1}{\sigma_P^2}}\theta)\}. \end{aligned} \tag{2.4}$$

   □

2. Using the assumption abouve, the posterior of θ after observing $x_1, \cdots, x_n$ is a.s. Gaussian with mean μ_n and variance σ_n^2 given by:

$$\begin{aligned} \mu_n &= \frac{\sigma_S^2/n}{\sigma_S^2/n + \sigma_P^2} \mu_P + \frac{\sigma_P^2}{\sigma_S^2/n + \sigma_P^2} \bar{x}, \\ \sigma_n^2 &= \frac{\sigma_P^2 \sigma_S^2}{n\sigma_P^2 + \sigma_S^2}. \end{aligned} \tag{2.5}$$

   Proof. Note that $\bar{X} \sim N(\theta, \sigma_S^2/n)$, then subsititute x above with $\bar{x}$.
   Or you can calculate the law of $\boldsymbol{x} = (x_1, \cdots, x_n)$ and get the same result. □

3. $X \sim B(n, \theta)$, θ has a prior $Be(a, b)$, then the posterior of θ is $Be(x + a, n - x + b)$.

   Proof. $X \sim B(n, \theta)$, it has a distribution:

$$f(x \mid \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}. \tag{2.6}$$

θ satisfies:

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}. \tag{2.7}$$

then the posterior of θ can be calculated by:

$$\begin{aligned}\pi(\theta \mid x) &\propto f(x \mid \theta)\pi(\theta) \\ &= \theta^{(x+a)-1}(1-\theta)^{(n-x+b)-1}.\end{aligned} \tag{2.8}$$

□

3 Concentration Inequality

Consider random variables $X_1, X_2, \ldots$ Assume they are mutually independent, but not necessary identially distributed. Let $\bar{X}_n = \frac{X_1+\ldots+X_n}{n}$ to be the average of the first n random variables, and let $\mu_n = \mathbb{E}[\bar{X}_n]$ be its expectation. We have:

Theorem 3.1 (Hoeffding Inequality)**.**

$$\mathbb{P}\{|\bar{X}_n - \mu_n| \leq \sqrt{\alpha \log T/n}\} \geq 1 - 2T^{-2\alpha}, \alpha > 0. \tag{3.1}$$

Here T is a fixed parameter.

4 Stochastic Bandit

4.1 Simple bandit algorithms

In this part I tend summarize some simple bandit algorithms and their regret bounds.

We consider the basic model with IID rewards, called stochastic bandit. An algorithm has K possible actions to choose from, and there are T rounds, for some known K and T. The mean reward of arm a is $\mu(a) := \mathbb{E}[\mathcal{D}_a]$, in which $\mathcal{D}$ is the reward distribution. The best mean reward is denoted $\mu^* = \max a \in \mathcal{A}\mu(a)$, the difference $\delta(a) := \mu^*(a) - \mu(a)$ describes how bad arm a is compared to μ^* Then we can define our main goal regret by:

Definition 4.1 (Regret)**.**

$$R(T) = \mu^* T - \Sigma_{t=1}^{T}\mu(a_t). \tag{4.1}$$

Then I can begin to summarize some simple algorithms and analysis their regret.

Algorithm 1 Explore-first algorithm

1: Exploration phase: try each arm N times;
2: Select the arm $\hat{a}$ with the highest average reward (break ties arbitrarily);
3: Exploitation phase: play arm $\hat{a}$ in all remaining rounds.

Theorem 4.2. *Explore-first algorithm achieves regret* $\mathbb{E}[R(T)] \leq T^{2/3} \times O(K \log T)^{1/3}$.

Algorithm 2 Epsilon-greedy algorithm

1: **for** each round t=1,2,… **do**
2: Toss a coin with success probability ϵ_t;
3: **if** success **then**
4: explore:choose an arm uniformly at random
5: **else**
6: exploit:choose the arm with the highest average reward so far
7: **end if**
8: **end for**

Theorem 4.3. *Epsilon-greedy algorithm with exploration probability* $\epsilon_t = t^{-1/3}(K \log t)^{1/3}$ *achieves regret bound* $\mathbb{E}[R(t)] \leq t^{2/3} O(K \log t)^{1/3}$ *for each round t.*

Remark 4.4. Explore-first and Epsilon-greedy do not adapt their exploration schedule to the history of the observed rewards. The set of all exploration rounds and the choice of arms therein is fixed before the round 1.

Definition 4.5 (Confidence interval)**.** *For each arm a and round t,*

$$\begin{aligned} r_t(a) &= \sqrt{2 \log(T)/n_t(a)} \quad \textit{(confidence radius)} \\ UCB_t(a) &= \bar{\mu}_t(a) + r_t(a) \quad \textit{(upper confidence bound)} \\ LCB_t(a) &= \bar{\mu}_t(a) - r_t(a) \quad \textit{(lower confidence bound).} \end{aligned} \tag{4.2}$$

Using concentration inequality we have *confidence interval* $[LCB_t(a), UCB_t(a)]$ and *confidence radius* $r_t(a)$. Then we have some adaptive exploration algorithms:

Algorithm 3 successive algorithm for two arms

1: Alternate two arms until $UCB_t(a) < LCB_t(a')$ after some even round t;
2: Abandon arm a, and use arm a' forever since.

For multi-armed bandit, we have:

Algorithm 4 Successive Elimination algorithm

1: All arms are initially designed as *active*
2: **loop**
3: play each active arm once
4: deactive all arms a such that, letting t be the current round, $UCB_t(a) < LCB_t(a')$ for some other arm a' deactivation rule
5: **end loop**

Theorem 4.6. *Successive Elimination algorithm achieves regret*

$$\mathbb{E}[R(t)] = O(\sqrt{Kt \log T}) for\ all\ rounds\ t \leq T. \tag{4.3}$$

Theorem 4.7. *Successive Elimination algorithms achieves regret*

$$\mathbb{E}[R(t)] \leq O(\log T) \left[\sum_{arms\ a\ with\ \mu(a)<\mu^*(a)} \frac{1}{\mu^*(a) - \mu(a)} \right] \tag{4.4}$$

Note that we can only use UCB_t to determination which arm is better, this is because an arm a can have a large $UCB_t(a)$ for two reasons (or combination thereof): because the average $\bar{mu}_t(a)$ is large, and/or because the confidence radius $r_t(a)$ is large, in which case this arm has not been explored much. So we have the algorithm:

Algorithm 5 Algorithm UCB1

1: Try each arm once
2: **for** each round t=1,...,T **do**
3: pich arm some a which maximizes $UCB_t(a)$.
4: **end for**

Theorem 4.8. *Algorithm UCB1 satisfies regret bounds in 4.3 and 4.4.*

Theorem 4.9 (lower bound)**.** *Fix time horizon T and the number of arms K. For any bandit algorithm, there exists a problem instance such that* $\mathbb{E}[R(T)] \geq \Omega(\sqrt{KT})$.

Consider a simple algorithm for Bayesian bandits, called Thompson Sampling. For each round t and arm a, the algorithm computes the posterior probability that a is the best arm, and samples a with this probability.

Algorithm 6 Thompson Sampling

1: **for** each round t=1,2,... **do**
2: Observe $H_{t-1} = H$, for some feasible (t-1)-history H;
3: Draw arm a_t independently from distribution $p_t(\cdot|H)$, where $p_t(a|H) := \mathbb{P}[a^* = a|H_{t-1} = H]$ for each arm a.
4: **end for**

Algorithm 7 Thompson Sampling: equivalent version

1: **for** each round t=1,2,... **do**
2: Observe $H_{t-1} = H$, for some feasible (t-1)-history H;
3: Sample mean reward vector μ_t from the posterior distribution $\mathbb{P}_H$;
4: Choose the best arm $\tilde{a}_t$ according to μ_t.
5: **end for**

And if we have independent priors, the distribution of μ can be easily calculated by using $\mathbb{P}_H^a$ only.
Then let us analyze Bayesian regret of Thompson Sampling:

Theorem 4.10. *Bayesian Regret of Thompson Sampling is* $BR(T) = O(KT \log T)$.

This is the core theorem of the TS algorithm, and its proof is very subtle, so I want to state it in detail.
First, we can recap the definition of the confidence interval 4.2 defined before. THen we say the key lemma below hold for a more general notion of the confidence bounds, and clearly hold for 4.2. Actually, $U(a, H_t)$ and $L(a, H_t)$ can be arbitrary functions of the arm a and the t-history H_t. There are two properties we want these functions to have, for some $\gamma > 0$ to be specified later:

$$\begin{aligned} \mathbb{E}\big[[U(a, H_t) - \mu(a)]^-\big] &\le \frac{\gamma}{TK} \quad \text{for all arm } a \text{ and rounds } t, \\ \mathbb{E}\big[[\mu(a) - L(a, H_t)]^-\big] &\le \frac{\gamma}{TK} \quad \text{for all arm } a \text{ and rounds } t. \end{aligned} \tag{4.5}$$

The confidence radius can be defined as $r(a, H_t) = \frac{U(a,H_t)-L(a,H_t)}{2}$.

Lemma 4.11. *Assume we have lower and upper bound functions that satisfies properties above, for some parameter* $\gamma > 0$. *Then Bayesian Regret of Thompson Sampling can be bound as follows:*

$$BR(T) \le 2\gamma + 2\Sigma_{t=1}^{T} \mathbb{E}[r(a_t, H_T)].$$

Proof. Fix round t. As two algorithms are equivalent, we have:

$$\mathbb{P}[a_t = a|H_t = H] = \mathbb{P}[a^* = a|H_t = H] \text{for each arm } a. \tag{4.6}$$

It follows that

$$\mathbb{E}[U(a^*, H)|H_t = H] = \mathbb{E}[U(a_t, H)|H_t = H]. \tag{4.7}$$

The Bayesian Regret suffered in round t is

$$\begin{aligned} BR_t &= \mathbb{E}[\mu(a^* - \mu(a_t))] \\ &= \mathop{\mathbb{E}}_{H \sim H_t}[\mathbb{E}[\mu(a^*) - \mu(a_t)|H_t = H]] \\ &= \mathop{\mathbb{E}}_{H \sim H_t}[\mathbb{E}[U(a_t, H) - \mu(a_t) + \mu(a^*) - U(a^*, H)|H_t = H]] \\ &= \mathbb{E}[U(a_t, H_t) - \mu(a_t)] + \mathbb{E}[\mu(a^*) - U(a^*, H_t)]. \end{aligned} \tag{4.8}$$

We will use properties above to bound both summands. Note that we cannot immediately use these properties because they assume a fixed arm a, whereas both a_t and a^* are random varibales.

$$\begin{aligned}\mathbb{E}[\mu(a^*) - U(a^*, H_t)] &\leq \mathbb{E}[(\mu(a^*) - U(a^*, H_t))^+] \\ &\leq \mathbb{E}[\sum_{armsa} (\mu(a^*) - U(a^*, H_t))^+] \\ &= \Sigma_{armsa}\mathbb{E}[(U(a, H_t) - \mu(a))^-] \\ &\leq K\frac{\gamma}{KT} = \frac{\gamma}{T}.\end{aligned} \tag{4.9}$$

$$\mathbb{E}[U(a_t, H_t) - \mu(a_t)] = \mathbb{E}[2r(a_t, H_t) + L((a_t, H_t) - \mu(a_t))] \quad = \mathbb{E}[2r(a_t, H_t)] + \mathbb{E}[L((a_t, H_t) - \mu(a_t))] \tag{4.10}$$

$$\begin{aligned}\mathbb{E}[L((a_t, H_t) - \mu(a_t))] &\leq \mathbb{E}[(L((a_t, H_t) - \mu(a_t)))^+] \\ &\leq \mathop{\mathbb{E}}_{armsa}[(L((a_t, H_t) - \mu(a_t)))^+] \\ &= \mathop{\mathbb{E}}_{armsa}[(\mu(a_t) - L((a_t, H_t)))^-] \\ &\leq K\frac{\gamma}{KT} = \frac{\gamma}{T}.\end{aligned} \tag{4.11}$$

Thus, $BR_t(T) \leq 2\frac{\gamma}{T} + 2\mathbb{E}[r(a_t, H_t)]$, the lemma follows by summing up over all rounds t. □

Now we can proof the main theorem.

Proof. By lemma,

$$BR(T) \leq O(\sqrt{\log T})\sum_{t=1}^{T}\mathbb{E}\Big[\frac{1}{\sqrt{n_t(a_t)}}\Big]$$

Moreover,

$$\begin{aligned}\sum_{t=1}^{T}\frac{1}{\sqrt{n_t(a_t)}} &= \sum_{armsa}\sum_{roundst:a_t=a}\frac{1}{\sqrt{n_t(a)}} \\ &= \sum_{armsa}\sum_{j=1}^{n_{T+1}(a)}\frac{1}{\sqrt{j}} = \sum_{armsa}O(\sqrt{n(a)}).\end{aligned}$$

It follows that

$$BR(T) \leq O(\sqrt{\log T})\sum_{armsa}\sqrt{n(a)} \leq O(\sqrt{logT})\sqrt{K\sum_{armsa}n(a)} = O(\sqrt{KT\log T})$$

. □

4.2 Lipschitz Bandit

In a special case, arms correspond to point in the interval $X = [0, 1]$, and expected rewards obey a Lipschitz condition:

$$|\mu(x) - \mu(y)| \leq L|x - y| \text{for any two arms } x, y \in X. \tag{4.12}$$

We can see this Lipschitz condition describes "similar" arms have similar expected rewards. As this bandit has infinity many arms, a simple solution to this is use finite arms to approximate the whole model. And such Lipschitz condition can guarantee these finite arms have enough information.

Theorem 4.12. *Consider continuum-armed bandits with Lipschitz constant L and time horizon T . Uniform discretization with algorithm ALG satisfying Lipschitz and discretization step $\epsilon = (TL^2/\log T)^{-1/3}$ attains*

$$\mathbb{E}[R(T)] \leq L^{1/3}T^{2/3}(1 + c_{ALG})(\log T)^{1/3} \tag{4.13}$$

The main take-away here is the $\tilde{O}(L^{1/3}T^2/3)$ regret rate. The explicit constant and logarithmic dependence are less important.

Actually, uniform discretization is optimal in the worst case: we have an $\Omega(L^{1/3}T^{2/3})$ lower bound on regret.

Theorem 4.13. *Let* ALG be any algorithm for continuum-armed bandits with time horizon T and Lipschitz constant L. There exists a problem instance $L = L(x^*, \epsilon)$, for some $x \in [0, 1]$ and $\epsilon > 0$, such that

$$\mathbb{E}[R(T)|L] \geq (L^{1/3}T^{2/3}). \tag{4.14}$$

In a more general case, the Lipschitz condition can be stated as:

$$|\mu(x) - \mu(y)| \leq \mathcal{D}(x, y) \text{for any arms } x, y \tag{4.15}$$

where $\mathcal{D}$is a metric.

Definition 4.14. *A subset $S \in X$ is called an $\epsilon - mash, \epsilon > 0$, if every point $x \in X$ is within distance ϵ from S, in the sense that $\mathcal{D}(x, y) \leq \epsilon$ for some $y \in S$.*

Then we have regret bound for this more general case:

Theorem 4.15. *Consider Lipschitz bandits with time horizon T. Optimzing over the choice of an $\epsilon - mesh$, uniform discretization with algorithm* ALG attains regret

$$\mathbb{E}[R(T)] \leq \inf_{\epsilon > 0, \epsilon - mesh\ S} \epsilon T + c_{\text{ALG}} \sqrt{|S| T \log T} \tag{4.16}$$

In the $d - dimension$ Eulidean space with l_p metric, we can get a explicit form. And in more general cases, this form can be attained through the definition of "covering dimension". But all of these are about optimizing over the choice of all subset. We leave out these unimportant details to get into our main algorithm quickly.

Now we can describe the algorithm. It contains two parts: activation step and selection step. We consider the confidence ball defined as $B_t(x) = \{y \in X : \mathcal{D}(x, y) \leq r_t(x)\}$. In the activation step, we find an arm which is not in any confidence ball of all the active arms: we can see in intuition that if some arm lies very close to some active arm, we don't need to explore it because of the Lipschitz condition.

Then in the selection step, we select an arm which is active. The idea is just like UCB1. If an arm x is active at time t, we define

$$index_t(x) = \bar{\mu}_t(x) + 2r_t(x). \tag{4.17}$$

The selcetion rule is very simple: play an active arm with the largest index.

Algorithm 8 Zooming algorithm for adaptive discretization.

```
Initialize: set of active arms S ← ∅
for each round t=1,2,… do
    if some arm y is not covered by the confidence balls of active arms then
        pick any such arm y and "activate" it: S ← S ∪ {y}.
    end if
    Play an active arm x with the largest index_t(x).
end for
```

Definition 4.16. *The smallest number of subsets in an $\epsilon -$covering is called the covering number and denoted $N_\epsilon(X)$.*

For any instance of Lipschitz MAB, the zooming dimension with multiplier $c > 0$ is

$$\inf_{d \geq 0} \{N_{r/3}(X) \leq cr^{-d}\} \tag{4.18}$$

Theorem 4.17. *Consider Lipschitz bandits with time horizon T . Assume that realized rewards take values on a finite set. For any given problem instance and any $c > 0$, the zooming algorithm attains regret*

$$\mathbb{E}[R(T)] \leq O(T^{(d+1)/(d+2)}(c \log T)^{1/(d+2)}), \textit{where } d \textit{ is the zooming dimension with multiplier } c. \tag{4.19}$$

5 Adversarial Bandits

To catch up quickly, let's run through this section.

First, we need to have a basic knowledge about adversarial cost. It may be influenced by algorithms and mean to "fool" the algorithm. The main set up of this part is at each step, we can choose one arm, suffer the cost and view the cost of every arm. Now we can desctibe two main algorithms and alalysis their effectiveness.

Algorithm 9 Weighted Majority Algorithm

Parameter: $\epsilon \in [0, 1]$
for each round t **do**
  Make predictions using weighted majority vote based on ω.
  for each expert i **do**
    if the i-th expert' s prediction is correct **then**
      ω_i stays the same
    else
      $\omega_i = \omega_i(1 - \epsilon)$
    end if
  end for
end for

Theorem 5.1. *The number of mistakes made by WMA with parameter $\epsilon \in [0, 1]$ is at most*

$$\frac{2}{1-\epsilon} cost^* + \frac{2}{\epsilon} \ln K$$

Algorithm 10 Hedge algorithm for online learning with experts

Initialize the weights as $\omega_1(a) = 1$ for each arm a.
for each round t **do**
  Let $p_t(a) = \frac{\omega_t(a)}{\Sigma_{a'=1}^{K} \omega_t(a')}$.
  Sample an arm a_t from distribution $p_t()$.
  Observe cost $c_t(a)$ for each arm a.
  For each arm a, update its weight $\omega_{t+1}(a) = \omega_t(a)(1 - \epsilon)^{c_t(a)}$
end for

Below we analyze Hedge, and prove $O(\sqrt{T \log K})$ bound on expected regret, the best possible for regret.

6 Contextual Bandits

The problem set of this so-called contextual bandit is every time before we make a choice among arms a_t, algorithm could observe a "context" x_t. As usual, reward $r_t \in [0, 1]$ is realized.

For small number of context, we can just apply algorithms we have developed before. Here is an easy algorithm which describe this process.

Algorithm 11 Contextual bandit algorithm for a small number of contexts

Initialization: For each context x, create an instance ALG_x of algorithm ALG
for each round t **do**
  invoke algorithm ALG_x with $x = x_t$
  "play" action a_t chosen by ALG_x, return reward r_t to ALG_x.
end for

Theorem 6.1. *Algorithm above has regret* $\mathbb{E}[R(T)] = O(\sqrt{KT|\mathcal{X}| \ln T})$.

To handle contextual bandit with a large $|\mathcal{X}|$, we either assume some structure such as Lipschitz condition or the setting of linear contextual bandits, or change the objective.

Let's stop here for the moment. It's time to read the literature.