Multi-armed Bandits

Mathematical Analysis of Machine Learning Algorithms
(Chapter 16)

Multi-armed Bandits

In the multi-armed bandit (or MAB) problem, we consider K arms. The environment generates a sequence of reward vectors for time steps $t \geq 1$ as $r_t = [r_t(1), \ldots, r_t(K)]$. Each $r_t(a)$ is associated with an *arm* $a \in \{1, \ldots, K\}$. An arm a is also referred to as an *action*.

MAB

At each time step $t = 1, 2, \ldots, T$,

- The player pulls one of the arms $a_t \in \{1, \ldots, K\}$.
- The environment returns the reward $r_t(a_t)$, but does not reveal information on any other arm $a \neq a_t$.

At each time t, a (randomized) bandit algorithm takes the historic observations observed so far, and maps it to a distribution $\hat{\pi}_{t-1}$ over actions $a \in \{1, \ldots, K\}$. We then draw a random action a_t (arm) from $\hat{\pi}_{t-1}$. Here $\hat{\pi}_{t-1}$ is referred to as policy.

Adversarial Bandit

In adversarial bandit, we are given an arbitrary reward sequence $\{[r_1(1), \ldots, r_t(K)] : t \geq 1\}$ a priori. For this reward sequence, the expected cumulative reward of a randomized bandit algorithm is

$$\mathbb{E} \sum_{t=1}^{T} r_t(a_t),$$

where $\mathbb{E}$ is over the internal randomization of the bandit algorithm.

Regret

$$\mathrm{REG}_T = \max_a \sum_{t=1}^{T} r_t(a) - \mathbb{E} \sum_{t=1}^{T} r_t(a_t). \tag{1}$$

If the algorithm pulls a single arm a_t deterministically at any time step:

$$\mathrm{REG}_T = \max_a \sum_{t=1}^{T} r_t(a) - \sum_{t=1}^{T} r_t(a_t).$$

Stochastic Bandit

In *stochastic bandit*, we assume that the reward $r_t(a)$ is drawn independently from a distribution $\mathcal{D}_a$, with mean $\mu(a) = \mathbb{E}_{r_t(a)\sim\mathcal{D}_a}[r_t(a)]$. In this setting, the goal is to find a that maximizes the expected reward $\mu(a)$.

Regret

$$\mathrm{REG}_T = T \max_a \mu(a) - \mathbb{E} \sum_{t=1}^{T} \mu(a_t). \tag{2}$$

The regret is defined for each realization of the stochastic rewards; the expectation is with respect to the randomness of the algorithm.

If the algorithm has deterministic output a_t at each time t, then

$$\mathrm{REG}_T = T \max_a \mu(a) - \sum_{t=1}^{T} \mu(a_t).$$

Expected Regret Bound for Stochastic Bandit

We can further include randomization over data into the regret of stochastic bandit, by considering the following expected regret over all possible realizations of observed data.

Expected Regret

$$\mathbb{E}\,\mathrm{REG}_T = T \max_a \mu(a) - \mathbb{E} \sum_{t=1}^{T} \mu(a_t),$$

where the expectation is with respect to both the data and the internal randomization of the learning algorithm.

Upper Confidence Bound

An important algorithm for stochastic bandit is *Upper Confidence Bound* (UCB), which is a deterministic algorithm.

In this method, we define for each $a = 1, \ldots, K$

$$\hat{n}_t(a) = \sum_{s=1}^{t} \mathbb{1}(a_s = a), \quad \hat{\mu}_t(a) = \frac{1}{\hat{n}_t(a)} \sum_{s=1}^{t} r_s(a_s)\mathbb{1}(a_s = a), \tag{3}$$

and a properly defined $c_t(a)$, so that the following upper confidence bound holds with high probability at time t.

Upper Confidence Bound

Given any optimal arm $a_* \in \arg\max_a \mu(a)$:

$$\mu(a_*) \leq \hat{\mu}_{t-1}(a_*) + \hat{c}_{t-1}(a_*).$$

UCB Algorithm

Algorithm 1: UCB Algorithm

Input: K and $T \geq K$

for $a = 1, \ldots, K$ **do**
    Let $\hat{n}_0(a) = 0$
    Let $\hat{\mu}_0(a) = 0$
for $t = 1, 2, \ldots, T$ **do**
    if $t \leq K$ **then**
        Let $a_t = t$
    else
        Let $a_t \in \arg\max_a [\hat{\mu}_{t-1}(a) + \hat{c}_{t-1}(a)]$ according to (3)
    Pull arm a_t and observe reward $r_t(a_t)$

UCB Analysis: Generic High Probability Result

Lemma 1 (Lem 16.2)

Let $a_* \in \arg\max_a \mu(a)$*. Let*

$$\delta_1 = \Pr\left[\exists t > K : \mu(a_*) > \hat{\mu}_{t-1}(a_*) + \hat{c}_{t-1}(a_*)\right]$$

be the probability that the upper confident bound fails on a_**. Let*

$$\delta_2 = \Pr\left[\exists t > K \;\&\; a \in \{1, \ldots, K\} \setminus \{a_*\} : \mu(a) < \hat{\mu}_{t-1}(a) - \hat{c}'_{t-1}(a)\right]$$

be the probability that the lower confident bound fails on $a \neq a_*$*. Then for Algorithm 1, we have with probability at least* $1 - \delta_1 - \delta_2$*:*

$$\mathrm{REG}_T \leq \sum_{a=1}^{K} [\mu(a_*) - \mu(a)] + \sum_{t=K+1}^{T} [\hat{c}_{t-1}(a_t) + \hat{c}'_{t-1}(a_t)] \mathbb{1}(a_t \neq a_*).$$

Proof of Lemma 1

We have with probability $1 - \delta_1 - \delta_2$, the following hold for all $t > K$:

$$\begin{aligned} \mu(a_*) \leq & \hat{\mu}_{t-1}(a_*) + \hat{c}_{t-1}(a_*), && \text{(upper confidence bound)} \\ \hat{\mu}_{t-1}(a_t)\mathbb{1}(a_t \neq a_*) \leq & [\mu(a_t) + \hat{c}'_{t-1}(a_t)]\mathbb{1}(a_t \neq a_*). \\ & && \text{(lower confidence bound)} \end{aligned}$$

It follows that for all $t > K$:

$$\begin{aligned} & \mu(a_*)\mathbb{1}(a_t \neq a_*) \\ \leq & [\hat{\mu}_{t-1}(a_*) + \hat{c}_{t-1}(a_*)]\mathbb{1}(a_t \neq a_*) && \text{(upper confidence bound)} \\ \leq & [\hat{\mu}_{t-1}(a_t) + \hat{c}_{t-1}(a_t)]\mathbb{1}(a_t \neq a_*) && \text{(UCB algorithm)} \\ \leq & [\mu(a_t) + \hat{c}'_{t-1}(a_t) + \hat{c}_{t-1}(a_t)]\mathbb{1}(a_t \neq a_*). && \text{(lower confidence bound)} \end{aligned}$$

For $t \leq K$, we have

$$\mu(a_*) = \mu(a_t) + [\mu(a_*) - \mu(a_t = t)].$$

We obtain the bound by summing over $t = 1$ to $t = T$.

Remarks

The analysis of bandits in Lemma 1 is similar to the empirical process analysis of ERM. This technique can be used in other bandit problems such as linear bandits.

- We note that although the algorithm uses an upper confidence bound, it only requires the bound to hold for the optimal arm a_*.
- The regret bound relies on both upper confidence bound, and lower confidence bound.
- The lower confidence bound needs to hold for all arms a, which implies that it holds for a_t. However, the upper confidence bound does not have to satisfy for all a_t.
- Given upper and lower confidence bounds, the regret bound for MAB becomes an estimation of the summation of the confidence bounds.

Regret for Bounded Reward

Theorem 2 (Thm 16.3)

Assume that rewards $r_t(a) \in [0,1]$. *Let* $a_* \in \arg\max_a \mu(a)$. *With a choice of*

$$\hat{c}_t(a) = \sqrt{\frac{\ln(2(\hat{n}_t(a)+1)^2/\delta)}{2\hat{n}_t(a)}},$$

we have with probability at least $1-\delta$*:*

$$\mathrm{REG}_T \leq (K-1) + \sqrt{8\ln(2KT^2/\delta)(T-K)K}.$$

Proof of Theorem 2 (I/III)

Given any integer $m \geq 1$ and $a \in \{1, \ldots, K\}$. We know that the sequence $r(a_t)\mathbb{1}(a_t = a)$ satisfies the sub-Gaussian bound in Theorem 13.3 with $\sigma_i = 0.5\mathbb{1}(a_i = a)$. This means that $\sum_{i=1}^{t} \sigma_i^2 = 0.25\hat{n}_t(a)$. By letting $\sigma = 0.5\sqrt{m}$ for a constant m, and consider the event $\hat{n}_t(a) = m$, we obtain with probability at most $0.5\delta/(m+1)^2$:

$$\exists t \geq 1 : \mu(a) > \hat{\mu}_t(a) + \hat{c}_t(a) \quad \& \quad \hat{n}_t(a) = m.$$

Similarly, with probability at most $(0.5/K)\delta/(m+1)^2$:

$$\exists t \geq 1 : \mu(a) < \hat{\mu}_t(a) - \hat{c}'_t(a) \quad \& \quad \hat{n}_t(a) = m,$$

where the lower confidence interval size is defined as

$$\hat{c}'_t(a) = \sqrt{\frac{\ln(2K(\hat{n}_t(a)+1)^2/\delta)}{2\hat{n}_t(a)}}.$$

Proof of Theorem 2 (II/III)

It follows that the failure probability of upper confidence bound is given by

$$\begin{aligned}
&\Pr\left[\exists t > K : \mu(a_*) > \hat{\mu}_t(a_*) + \hat{c}_t(a_*)\right] \\
\leq &\sum_{m=1}^{\infty} \Pr\left[\exists t > K : \ \hat{n}_t(a_*) = m \ \& \ \mu(a_*) > \hat{\mu}_t(a_*) + \hat{c}_t(a_*)\right] \\
\leq &\sum_{m=1}^{\infty} 0.5\delta/(m+1)^2 \leq 0.5\delta.
\end{aligned}$$

Moreover, the failure probability of lower confidence bound is given by

$$\begin{aligned}
&\Pr\left[\exists t > K \ \& \ a \in \{1, \ldots, K\} : \mu(a) < \hat{\mu}_t(a) - \hat{c}'_t(a)\right] \\
\leq &\sum_{a=1}^{K} \Pr\left[\exists t > K : \ \mu(a) < \hat{\mu}_t(a) - \hat{c}'_t(a)\right] \\
\leq &\sum_{a=1}^{K} \sum_{m=1}^{\infty} \Pr\left[\exists t > K : \ \hat{n}_t(a) = m \ \& \ \mu(a) < \hat{\mu}_t(a) - \hat{c}'_t(a)\right] \\
\leq &\sum_{a=1}^{K} \sum_{m=1}^{\infty} \frac{\delta}{2K(m+1)^2} \leq \delta/2.
\end{aligned}$$

Proof of Theorem 2 (III/III)

The following bound follows from Lemma 1. With probability at least $1-\delta$:

$$\sum_{t=1}^{T}[\mu(a_*)-\mu(a_t)]-\sum_{a=1}^{K}[\mu(a_*)-\mu(a)] \le \sum_{t=K+1}^{T}[\hat{c}_{t-1}(a_t)+\hat{c}'_{t-1}(a_t)]$$

$$\le 2\sum_{t=K+1}^{T}\sqrt{\frac{\ln(2K(\hat{n}_{t-1}(a_t)+1)^2/\delta)}{2\hat{n}_{t-1}(a_t)}}$$

$$= 2\sum_{a=1}^{K}\sum_{t=K+1}^{T}\sqrt{\frac{\ln(2K(\hat{n}_{t-1}(a_t)+1)^2/\delta)}{2\hat{n}_{t-1}(a_t)}}\mathbb{1}(a_t=a)$$

$$\overset{(a)}{=} 2\sum_{a=1}^{K}\sum_{m=1}^{\hat{n}_T(a)-1}\sqrt{\frac{\ln(2K(m+1)^2/\delta)}{2m}} \le 2\sum_{a=1}^{K}\int_0^{\hat{n}_T(a)-1}\sqrt{\frac{\ln(2K\hat{n}_T(a)^2/\delta)}{2t}}dt$$

$$= 4\sum_{a=1}^{K}\sqrt{\frac{\ln(2K\hat{n}_T(a)^2/\delta)(\hat{n}_T(a)-1)}{2}} \le 4\sqrt{\frac{\ln(2KT^2/\delta)(T-K)K}{2}}.$$

Therefore $$\sum_{a=1}^{K}\sqrt{\hat{n}_T(a)-1} \le \sqrt{K\sum_{a=1}^{K}(\hat{n}_T(a)-1)} = \sqrt{K(T-K)}.$$

Example 3 (Multiplicative UCB)

Assume in Theorem 2, $\mu(a_*) \approx 0$. In such case, we can use the following multiplicative Chernoff bound in Theorem 13.5 to obtain an upper confidence bound. We know that with probability at most $1 - \delta$:

$$\mu(a) \leq \frac{e}{e-1}\hat{\mu}_t(a) + \frac{e\ln(1/\delta)}{(e-1)\hat{n}_t(a)} \quad \& \quad \hat{n}_t(a) = m.$$

This implies that we may use an upper confidence bound

$$\hat{c}_t(a) = \frac{1}{e-1}\hat{\mu}_t(a) + \frac{e\ln(2(\hat{n}_t(a)+1)^2/\delta)}{(e-1)\hat{n}_t(a)}$$

in Algorithm 1.

In order to obtain a better regret bound than that of Theorem 2, one also needs to use the multiplicative Chernoff bound for the lower confidence interval, and then repeat the analysis of Theorem 2 with such a multiplicative lower confidence interval bound.

Example: sub-Gaussian UCB

Example 4

Assume for each arm a, the reward $r_t(a)$ is a sub-Gaussian random variable, but different arms have different reward distributions:

$$\ln \mathbb{E} \exp(\lambda r_t(a)) \leq \lambda \mathbb{E} r_t(a) + \frac{\lambda^2}{2} M(a)^2,$$

where $M(a)$ is known. Then one can obtain a bound similar to Theorem 2, with arm dependent UCB estimate involving $M(a)$.

Gap-dependent Bound

Lemma 5 (Lem 16.6)

Let $a_* \in \arg\max_a \mu(a)$. *Let*

$$\delta_1 = \Pr\left[\exists t > K : \mu(a_*) > \hat{\mu}_{t-1}(a_*) + \hat{c}_{t-1}(a_*)\right],$$

$$\delta_2 = \Pr\left[\exists t > K, a \in \{1, \ldots, K\} \setminus \{a_*\} : \mu(a) < \hat{\mu}_{t-1}(a) - \hat{c}'_{t-1}(a)\right].$$

Define $\Delta(a) = \mu(a_*) - \mu(a)$, *and*

$$T(a) = \max\left\{m : \Delta(a) \leq \hat{c}_{t-1}(a) + \hat{c}'_{t-1}(a), \hat{n}_{t-1}(a) = m, K < t \leq T\right\} \cup \{0\}.$$

Then for Algorithm 1, we have with probability at least $1 - \delta_1 - \delta_2$*:*

$$\mathrm{REG}_T \leq \inf_{\Delta_0 > 0}\left[\sum_{a=1}^{K} T(a)\Delta(a)\mathbb{1}(\Delta(a) > \Delta_0) + (T - K)\Delta_0\right] + \sum_{a=1}^{K}\Delta(a).$$

Some Intuitions

In Lemma 5:

- δ_1: failure probability for the upper confidence bound.
- δ_2: failure probability for the lower confidence bound.
- $T(a)$: after we have pulled arm a for more than $T(a)$ times, the confidence interval for a becomes smaller than the gap $\Delta(a)$.

The definition of $T(a)$ implies that $T(a)$ is the maximum number of times that one will pull a particular arm a.

The regret caused by choosing a is upper bounded by $T(a)\Delta(a)$.

Proof of Lemma 5 (I/II)

The proof of Lemma 1 shows that with probability at least $1 - \delta$, for all $t \geq K + 1$:

$$\mu(a_*) \leq \mu(a_t) + [\hat{c}'_{t-1}(a_t) + \hat{c}_{t-1}(a_t)]\mathbb{1}(a_t \neq a_*),$$

which implies that

$$\Delta(a_t) \leq [\hat{c}'_{t-1}(a_t) + \hat{c}_{t-1}(a_t)]\mathbb{1}(a_t \neq a_*).$$

Using the assumption of the lemma, we obtain for all $a \neq a_*$ and $1 \leq t \leq T$, $\hat{n}_{t-1}(a)\mathbb{1}(a_t = a) \leq T(a)$.
It follows that for all $a \neq a_*$, let t be the last time such that $a_t = a$, then $t \leq T$ and $\hat{n}_T(a) = \hat{n}_{t-1}(a_t) + 1 \leq T(a) + 1$.

Proof of Lemma 5 (II/II)

We thus obtain the following regret bound for any $\Delta_0 \geq 0$:

$$\sum_{t=1}^{T}[\mu(a_*) - \mu(a_t)] = \sum_{a=1}^{K} \hat{n}_T(a)\Delta(a)$$

$$\leq \sum_{a=1}^{K} \Delta(a) + \sum_{a=1}^{K}(\hat{n}_T(a) - 1)\Delta_0 + \sum_{a=1}^{K}(\hat{n}_T(a) - 1)\Delta(a)\mathbb{1}(\Delta(a) > \Delta_0)$$

$$\leq \sum_{a=1}^{K} \Delta(a) + (T - K)\Delta_0 + \sum_{a=1}^{K} T(a)\Delta(a)\mathbb{1}(\Delta(a) > \Delta_0).$$

The second inequality used $\sum_{a=1}^{K} \hat{n}_T(a) = T - K$ and $\hat{n}_T(a) - 1 \leq T(a)$. This implies the bound.

Regret for Bounded Reward

Theorem 6 (Thm 16.7)

Assume that the reward $r_t(a) \in [0, 1]$. *Let* $a_* \in \arg\max_a \mu(a)$. *With a choice of*

$$\hat{c}_t(a) = \sqrt{\frac{\ln(2(\hat{n}_t(a)+1)^2/\delta)}{2\hat{n}_t(a)}},$$

we have

$$\mathrm{REG}_T \leq \inf_{\Delta_0 > 0} \left[(T-K)\Delta_0 + \sum_{a=1}^{K} \frac{2\ln(2KT^2/\delta)}{\Delta(a)} \mathbb{1}(\Delta(a) > \Delta_0) \right] + \sum_{a=1}^{K} \Delta(a).$$

Proof of Theorem 6

From the proof of Theorem 2, we know that with the choice of

$$\hat{c}'_t(a) = \sqrt{\frac{\ln(2K(\hat{n}_t(a)+1)^2/\delta)}{2\hat{n}_t(a)}},$$

the inequality $\Delta(a) \leq \hat{c}_{t-1}(a) + \hat{c}'_{t-1}(a)$ in the definition of $T(a)$ implies that

$$\Delta(a) \leq \sqrt{\frac{2\ln(2KT^2/\delta)}{\hat{n}_{t-1}(a)}}.$$

This implies that

$$\hat{n}_{t-1}(a) \leq \frac{2\ln(2KT^2/\delta)}{\Delta(a)^2}.$$

Therefore

$$T(a) \leq \frac{2\ln(2KT^2/\delta)}{\Delta(a)^2}.$$

This implies the desired result from Lemma 5.

Example: Gap Independent Bound

Example 7

The gap dependent regret bound of Theorem 6 implies the gap independent regret bound of Theorem 2.
In fact, if we take

$$\Delta_0 = \sqrt{\frac{K \ln(KT/\delta)}{T}},$$

then we obtain a regret of

$$\mathrm{REG}_T = O\left(\sqrt{KT \ln(KT/\delta)}\right)$$

from Theorem 6.

Example: Better Regret

Example 8

We may take $\Delta_0 = 0$ in the gap dependent regret bound of Theorem 6, and obtain

$$\mathrm{REG}_T \leq \sum_{a=1}^{K} \frac{2\ln(2KT^2/\delta)}{\Delta(a)} + K.$$

This implies $\tilde{O}(1)$ regret which depends on the smallest gap, instead of $\tilde{O}(\sqrt{T})$ regret for gap-independent case.

Expected Regret (Analogy of Lemma 5)

Lemma 9 (Lem 16.10)

Let $a_* \in \arg\max_a \mu(a)$. *For* $t > K$, *define*

$$\delta_1(t) = \Pr\left[\mu(a_*) > \hat{\mu}_{t-1}(a_*) + \hat{c}_{t-1}(a_*)\right],$$
$$\delta_2(t) = \Pr\left[\exists a \in \{1, \ldots, K\} \setminus \{a_*\} : \mu(a) < \hat{\mu}_{t-1}(a) - \hat{c}'_{t-1}(a)\right],$$

and let

$$\delta = \sum_{t>K} [\delta_1(t) + \delta_2(t)].$$

Define $\Delta(a) = \mu(a_*) - \mu(a)$, $M = \sup_a \Delta(a)$, *and*

$$T(a) = \max\left\{m : \Delta(a) \leq \hat{c}_{t-1}(a) + \hat{c}'_{t-1}(a), \hat{n}_{t-1}(a) = m, K < t \leq T\right\} \cup \{0\}.$$

Then for Algorithm 1, we have

$$\mathbb{E}\,\mathrm{REG}_T \leq \inf_{\Delta_0 > 0} \left[\sum_{a=1}^{K} T(a)\Delta(a)\mathbb{1}(\Delta(a) > \Delta_0) + (T-K)\Delta_0\right] + (K+\delta)M.$$

Expected Regret for Bounded Reward

Theorem 10 (Thm 16.11)

Assume that the reward $r_t(a) \in [0, 1]$. *Let* $a_* \in \arg\max_a \mu(a)$. *With a choice of*

$$\hat{c}_t(a) = \sqrt{\frac{\alpha \ln t}{2\hat{n}_t(a)}}$$

for $\alpha > 1$. *We have*

$$\mathbb{E}\,\mathrm{REG}_T \leq \inf_{\Delta_0 > 0} \left[\sum_{a=1}^{K} \frac{2\alpha \ln(T)}{\Delta(a)} \mathbb{1}(\Delta(a) > \Delta_0) + T\Delta_0 \right] + \frac{\alpha + 1}{\alpha - 1}(K + 1).$$

Gap-dependent Lower Bound

Theorem 11 (Thm 16.12)

Assume $r_t(a) \in \{0, 1\}$. Consider an algorithm such that $\forall \beta \in (0, 1)$, $\lim_{T\to\infty}(\mathrm{REG}_T / T^\beta) = 0$, then

$$\liminf_{T\to\infty} \frac{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(a_t = a)\right]}{\ln T} \geq \frac{1}{\mathrm{KL}(\mu(a), \mu_*)}$$

for all a, where $\mu_* = \max_a \mu(a)$.

The result implies that

$$\liminf_{T\to\infty} \frac{\mathbb{E}\,\mathrm{REG}_T}{\ln T} \geq \sum_{a=1}^{K} \frac{\Delta(a)}{\mathrm{KL}(\mu(a), \mu_*)} \mathbb{1}(\mu(a) < \mu_*).$$

Compare UCB Upper Bound to Lower Bound

Since $\mathrm{KL}(\mu(a), \mu_*)^{-1} = \Omega(\Delta(a)^{-2})$, it follows that the UCB bound in Theorem 10 has the worst case optimal dependency on the gap $\Delta(a)$ up to a constant.

For Gap-independent bound, by taking $\Delta_0 = O(\sqrt{K \ln(T)/T})$ in Theorem 10, we can obtain a gap-independent expected regret bound similar to Theorem 2:

$$\mathbb{E}\,\mathrm{REG}_T = O\left(\sqrt{KT \ln(T)}\right).$$

The lower bound, stated in Theorem 12, is $\Omega(\sqrt{KT})$.

Gap-Independent Lower Bound

Theorem 12 (Thm 16.13)

Given $K \geq 2$ and $T \geq 1$. Then there exists a distribution over the assignment of rewards $r_t(a) \in [0, 1]$ such that the expected regret of any algorithm (where the expectation is taken with respect to both the randomization over rewards and the algorithm's internal randomization) is at least

$$\min(0.02T, 0.06\sqrt{KT}).$$

Recall: Assouad's Lemma for Dependent Data

Theorem 13 ($d = 1$ and $m = K$ of Thm 13.24)

Let $K \geq 2$ be integers, and let $\mathcal{P}_{\mathcal{Z}} = \{q^\tau : \tau \in \{1, \ldots, K\}\}$ contain K probability measures. Let $Q(\theta, q)$ be a non-negative loss function. Assume that there exists $\epsilon, \beta \geq 0$ such that

$$\forall \tau' \neq \tau : \quad [Q(\theta, q^\tau) + Q(\theta, q^{\tau'})] \geq \epsilon,$$

and there exists q_0 with the following property. Given any learning algorithm $\hat{q}$. If for all time step t, and $\mathcal{S}_{t-1}$:

$$\frac{1}{K} \sum_{\tau=1}^{K} \mathrm{KL}(q_0(\cdot|\hat{q}(\mathcal{S}_{t-1}), \mathcal{S}_{t-1}) || q^\tau(\cdot|\hat{q}(\mathcal{S}_{t-1}), \mathcal{S}_{t-1}))) \leq \beta_t^2,$$

then

$$\frac{1}{K} \sum_{\tau=1}^{K} \mathbb{E}_{\theta, \mathcal{S}_n \sim p(\cdot|\hat{q}, q^\tau)} Q(\theta, q^\tau) \geq 0.5\epsilon \left(1 - \sqrt{2 \sum_{t=1}^{n} \beta_t^2}\right).$$

Proof of Theorem 12 (I/II)

We would like to apply Theorem 13. We consider a family of K distributions $\mathcal{P}_{\mathcal{Z}} = \{q^\tau, \tau = 1, \ldots, K\}$, and for each arm a, the distribution $q^\tau(a)$ is a Bernoulli distribution $r \in \{0, 1\}$ with mean $\mathbb{E}[r] = 0.5 + \epsilon \mathbb{1}(a = \tau)$ with $\epsilon \in (0, 0.1]$ to be determined later. We also define $q_0(a)$ as a Bernoulli distribution $r \in \{0, 1\}$ with mean $\mathbb{E}[r] = 0.5$.
If we pull an arm θ, we have

$$Q(\theta, q^\tau) = \epsilon \mathbb{1}(\theta \neq \tau).$$

It is clear that $[Q(\theta, q^\tau) + Q(\theta, q^{\tau'})] \geq \epsilon$ for $\tau \neq \tau'$. Consider $n \leq T$ samples generated sequentially by an arbitrary bandit algorithm and $q^\tau(a)$, and let the resulting distribution that generates $\mathcal{S}_n$ as $p^\tau(\mathcal{S}_n)$, where $\mathcal{S}_n = \{a_1, r_1, \ldots, a_n, r_n\}$.

Proof of Theorem 12 (II/II)

Let $\epsilon = \min(0.1, 0.24\sqrt{K/T})$. We have for all a_t:

$$\frac{1}{K}\sum_{\tau=1}^{K} \mathrm{KL}(q_0(a_t)||q^\tau(a_t)) = \frac{1}{K}\sum_{\tau=1}^{K} \mathrm{KL}(q_0(a_t)||q^\tau(a_t))\mathbb{1}(a_t = \tau)$$

$$= \frac{1}{K}\mathrm{KL}(0.5||0.5+\epsilon) \leq \frac{0.5\epsilon^2}{K(0.5+\epsilon)(0.5-\epsilon)} \leq \frac{2.1}{K}\epsilon^2.$$

Theorem 13 with $\beta_t^2 = \frac{2.1}{K}\epsilon^2$ implies that at the end of the n-th iteration, $\hat{\theta}$ of any learning algorithm satisfies

$$\frac{1}{K}\sum_{\tau=1}^{K} \mathbb{E}_{q^\tau}\mathbb{E}_{\hat{\theta}}\, Q(\hat{\theta}, q^\tau) \geq 0.5\epsilon\left(1 - \sqrt{2 \times 2.1(n/K)\epsilon^2}\right)$$

$$\geq 0.25\epsilon = 0.25\min\left(0.1, 0.24\sqrt{K/T}\right).$$

The second inequality used $(n/K)\epsilon^2 \leq 0.24^2$. Since this holds for all steps $n \leq T$, we obtain the desired bound.

Stochastic Linear Bandit

In MAB, the UCB algorithm has regret scale with K. This might not be desirable for problems that contain many arms.

In order to deal with such problems, we need to impose additional structures that model correlations among arms. A popular model for such problems is stochastic linear bandits, which allow large action (arm) space. In this section, we assume that the set of arms (or actions) is $\mathcal{A}$, which is finite: $|\mathcal{A}| = K$.

Stochastic Linear Bandit

Each time, we pull one arm $a \in \mathcal{A}$. We also know a feature vector $\phi(a) \in \mathcal{H}$ (where $\mathcal{H}$ is an inner product space) so that the expected reward is a linear function

$$\mu(a) = \theta_*^\top \psi(a)$$

with an unknown parameter $\theta_* \in \mathcal{H}$ to be estimated.

Arm Elimination for Stochastic Linear Bandits

Algorithm 2: Arm Elimination for Stochastic Linear Bandit

Input: $\mathcal{A}$, $\{\psi(a) : a \in \mathcal{A}\}$

Let $\mathcal{A}_0 = \mathcal{A}$

for $\ell = 1, 2, \ldots, L$ **do**

  Set parameters $\lambda_\ell, T_\ell, \beta_\ell, n_\ell, \eta_\ell$

  Use Algorithm 9.1 (with $\lambda = \lambda_\ell, n = n_\ell, \eta = \eta_\ell$) to obtain a policy $\sum_{i=1}^{m_\ell} \pi_{\ell,i} \mathbb{1}(a = a_{\ell,i})$ with $m_\ell \le n_\ell$ examples $\{a_{\ell,s} \in \mathcal{A}_{\ell-1}\}$

  For each $i = 1, \ldots, T_\ell$, pull $a_{\ell,i}$ for $J_\ell = \lceil T_\ell \pi_{\ell,i} \rceil$ times, and observe rewards $r_{\ell,i,j} \in [0, 1]$ $(j = 1, \ldots, J_\ell)$

  Let $\theta_\ell = \arg\min_\theta \sum_{i=1}^{m_\ell} \sum_{j=1}^{J_\ell} \left[(\theta^\top \psi(a_{\ell,i}) - r_{\ell,i,j})^2 + \lambda_\ell \|\theta\|_2^2\right]$

  Let $a_\ell = \arg\max_{a \in \mathcal{A}_{\ell-1}} \theta_\ell^\top \psi(a)$

  Let $\mathcal{A}_\ell = \{a \in \mathcal{A}_{\ell-1} : \theta_\ell^\top \psi(a) \ge \theta_\ell^\top \psi(a_\ell) - \beta_\ell\}$

Theorem 14 (Regret Bound, Thm 16.14)

Assume that we know $\|\theta_*\|_2 \leq B$. *Let* $a_* \in \arg\max_{a \in \mathcal{A}} \mu(a)$. *Given* $\eta > 0$. *For each* $\ell \geq 1$ *we set*

$$\lambda_\ell = \alpha^2/(B^2 T_\ell), \qquad T_\ell = 2^{\ell-1},$$
$$n_\ell = \lceil 8 \text{entro}(\lambda_\ell, \psi(\mathcal{A})) \rceil, \quad \eta_\ell = \min(0.1, 0.1/\dim(\lambda_\ell, \psi(\mathcal{A}))),$$
$$\beta_\ell = 2\left(\alpha + \sqrt{\frac{\ln(2K(\ell+1)^2/\delta)}{2}}\right)\sqrt{\frac{4\dim(\lambda_\ell, \psi(\mathcal{A}))}{T_\ell}}$$

in Algorithm 2. We also define $\beta_0 = 0.5$. *It follows that* $\forall \ell \geq 0$, $a_* \in \mathcal{A}_\ell$ *and*

$$\sup\{\mu(a_*) - \mu(a) : a \in \mathcal{A}_\ell\} \leq 2\beta_\ell.$$

This implies that after iteration L, and we have pulled total number of $T \leq (2^L - 1) + \sum_{\ell=1}^{L} n_\ell$ *arms, with probability at least* $1 - \delta$*:*

$$\text{REG}_T \leq 2\sum_{\ell=1}^{L}(n_\ell + T_\ell)\beta_{\ell-1}.$$

Interpretation of Theorem 14

For finite dimensional linear bandits, we can take $\dim(\mathcal{H}) = d$ and $\alpha = 1$ in Theorem 14. Proposition 15.8 implies that

$$\dim(\lambda_\ell, \psi(\mathcal{A})) \leq d,\ \mathrm{entro}(\lambda_\ell, \psi(\mathcal{A})) \leq d \ln(1 + (BB')^2/(\lambda_\ell d)) = O(d\ell),$$

where $B' = \sup_a \|\psi(a)\|_2$. We thus have $n_\ell = O(d \ln T)$.

Regret Bound of Theorem 14

$$\mathrm{REG}_T = O\left(\sqrt{Td \ln(KT)}\right).$$

If we ignore the log-factors, this generalizes the result of MAB. In fact, we note that MAB can be regarded as a linear bandit with $\psi(a) = e_a \in \mathbb{R}^K$, where e_a is the vector with value 1 at component a, and value 0 elsewhere. Therefore with $d = K$, we recover MAB results up to logarithmic factors.

Thompson Sampling

Thompson sampling[1] is another popular algorithm for bandit problems with a Bayesian interpretation.

The basic idea is to consider a prior distribution on the mean of the reward distribution of every arm, and at any time step, sample a mean from the posterior for each arm, then pick the arm with the highest sampled mean.

In practice, it will be convenient to use a model so that the posterior is simple. This can be achieved with conjugate priors.

[1] W. R. Thompson (1933). "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". In: *Biometrika* 25.3/4, pp. 285–294.

Thompson Sampling (Gaussian)

We can assume a Gaussian prior and reward likelihood:

$$\mu(a) \sim N(0,1), \quad r_t(a) \sim N(\mu(a),1).$$

Then the posterior for arm a after time step $t-1$ is given by the normal distribution

$$N(\hat{\mu}_{t-1}(a), \hat{V}_{t-1}(a)),$$

where

$$\hat{\mu}_{t-1}(a) = \frac{\sum_{s=1}^{t-1} \mathbb{1}(a_s = a) r_s(a_s)}{1 + \sum_{s=1}^{t-1} \mathbb{1}(a_s = a)}, \quad \hat{V}_{t-1}(a) = \frac{1}{1 + \sum_{s=1}^{t-1} \mathbb{1}(a_s = a)}. \tag{4}$$

This leads to Algorithm 3.

Algorithm

Algorithm 3: Thompson Sampling (Gaussian)

Input: K and T

for $t = 1, 2, \ldots, T$ **do**
  for $a = 1, \ldots, K$ **do**
    Sample $\tilde{\mu}_t(a) \sim N(\hat{\mu}_{t-1}(a), \hat{V}_{t-1}(a))$ according to (4)
  Let $a_t = \arg\max_a \tilde{\mu}_t(a)$
  Pull arm a_t and observe reward $r_t(a_t)$

Regret bound[2]

$$\mathrm{REG}_T = O(\sqrt{TK \ln K})$$

[2] S. Agrawal and N. Goyal (2013). "Further optimal regret bounds for Thompson sampling". In: *Artificial intelligence and statistics*. Ed. by C. M. Carvalho and P. Ravikumar. Vol. 31. PMLR, pp. 99–107.

EXP3 for Adversarial MAB

Algorithm 4: EXP3

Input: K, T, $\gamma \in (0, 1]$

for $a = 1, \ldots, K$ **do**
  Let $w_0(a) = 1$

for $t = 1, 2, \ldots, T$ **do**
  Let $w_{t-1} = \sum_{a=1}^{K} w_{t-1}(a)$
  for $a = 1, \ldots, K$ **do**
    Let $\hat{\pi}_t(a) = (1 - \gamma) w_{t-1}(a) / w_{t-1} + \gamma / K$
  Sample a_t according to $\hat{\pi}_t(\cdot)$
  Pull arm a_t and observe reward $r_t(a_t) \in [0, 1]$
  for $a = 1, \ldots, K$ **do**
    Let $\hat{r}_t(a, a_t) = r_t(a_t) \mathbb{1}(a = a_t) / \hat{\pi}_t(a_t)$
    Let $w_t(a) = w_{t-1}(a) \exp(\gamma \hat{r}_t(a, a_t) / K)$

Regret Bound

The following regret bounds hold for the EXP3 algorithm, with regret defined in (1).

Theorem 15 (Thm 16.17)

Consider Algorithm 4. Let $G_* = \max_a \sum_{i=1}^{T} r_t(a)$. *We have the following bound for the adversarial regret* (1)*:*

$$\mathrm{REG}_T \leq (e-1)\gamma G_* + \frac{K \ln K}{\gamma}.$$

Interpretation of the Bound

If we take

$$\gamma = \sqrt{\frac{K \ln K}{(e-1)g}}$$

for some $g \geq \max(G_*, K \ln K)$ in Theorem 15, then we obtain a bound

$$\mathrm{REG}_T \leq 2\sqrt{e-1}\sqrt{gK \ln K}.$$

In particular, with $g = T$, we have

$$\mathrm{REG}_T \leq 2\sqrt{e-1}\sqrt{TK \ln K}.$$

Summary (Chapter 16)

- Multi-armed Bandits
  - adversarial MAB
  - stochastic MAB
- UCB algorithm for stochastic MAB
- High probability analysis
  - upper confidence bound for a_*
  - lower confidence bound for all a
  - regret bound is summation of the confidence bounds
- Expected UCB bound
  - Gap dependent versus Gap independent bounds
  - Near optimal regret but not optimal
- Thompson Sampling
- EXP3 for Adversarial MAB