Lecture 16: Stochastic Bandits, Upper Confidence Bound Algorithm

Lecturer: Kirthevasan Kandasamy *Scribed by: Leitian Tao, Xuanfei Ren*

In this lecture, we will continue our discussion on classification in a VC Class, and introduce the stochastic bandits and Upper Confidence Bound Algorithm.

1 Classification in a VC Class (Cont'd)

We now use the above results to derive a lower bound for classification in a VC class.

Theorem 1. *Let $\mathcal{P}$ be the set of all distributions supported on $\mathcal{X} \times \{0,1\}$. Let $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Y}\}$ be a hypothesis class with VC dimension $d \geqslant 8$. Let $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\} \sim_{iid} P$, where $P \in \mathcal{P}$. Then, for any estimator $\hat{h}$ that maps the dataset S to a hypothesis in $\mathcal{H}$,*

$$R^* = \inf_{\hat{h}} \sup_{P \in \mathcal{P}} \left(\mathbb{E}[F(\hat{h}, P)] - \inf_{h' \in \mathcal{H}} F(h', P) \right) \geqslant C_1 \sqrt{\frac{d}{n}}$$

for some global constant C_1.

Proof Our proof follows the standard four-step recipe when applying the Fano/Le Cam methods.

Step 1: Construct alternatives.

Let $\mathcal{X}_d = \{x_1, \ldots, x_d\}$ be a set of points shattered by $\mathcal{H}$. Let $\gamma \leqslant 1/4$ be a value to be specified later. Define

$$\mathcal{P}' = \left\{ P_\omega : P_\omega(X = x) = \frac{1}{d} \mathbb{1}\{x \in \mathcal{X}_d\},\ P_\omega(Y = 1 \mid X = x_i) = \frac{1}{2} + (2\omega_i - 1)\gamma,\ \omega \in \Omega_d \right\},$$

where Ω_d is the VG-pruned hypercube of $\{0,1\}^d$.

Remark To illustrate the above construction, consider the class of two-sided threshold classifiers with $d = 2$, i.e., $\mathcal{X}_2 = \{x_1, x_2\} \subseteq \mathbb{R}$. Let P_ω be the distribution for $\omega = (0,1)$ with $P_\omega(X = x_1) = P_\omega(X = x_2) = 1/2$. Then the conditional distribution of Y is:

$$P_\omega(Y = 1 \mid X = x_1) = \frac{1}{2} - \gamma, \qquad P_\omega(Y = 1 \mid X = x_2) = \frac{1}{2} + \gamma.$$

Step 2: Lower bound the separation $\min_{\omega,\omega'} \Delta(P_\omega, P_{\omega'})$.

We claim that for any $P_\omega, P_{\omega'} \in \mathcal{P}'$, the separation satisfies

$$\Delta(P_\omega, P_{\omega'}) \geqslant \frac{\gamma}{d} H(\omega, \omega').$$

We will prove this claim in the homework. Then, by the Varshamov-Gilbert lemma, we have

$$\min_{\omega,\omega'} \Delta(P_\omega, P_{\omega'}) \geqslant \frac{\gamma}{d} \cdot \frac{d}{8} = \frac{\gamma}{8} \triangleq \delta.$$

Step 3: Upper bound the KL divergence $\max_{\omega,\omega'} \mathrm{KL}(P_\omega, P_{\omega'})$.

We calculate the KL divergence as follows:

$$\begin{aligned}
\mathrm{KL}(P_\omega, P_{\omega'}) &= \mathbb{E}_{X,Y}\left[\log \frac{P_\omega(X,Y)}{P_{\omega'}(X,Y)}\right] \\
&= \sum_{i=1}^{d} P_\omega(x_i) \sum_{y\in\{0,1\}} P_\omega(y \mid x_i) \log \frac{P_\omega(y \mid x_i)}{P_{\omega'}(y \mid x_i)} && (\text{since } P_\omega(x) = P_{\omega'}(x)) \\
&= \sum_{i=1}^{d} \frac{1}{d} \mathbb{1}\{\omega \neq \omega'\} \underbrace{\left[(\frac{1}{2}+\gamma)\log\frac{\frac{1}{2}+\gamma}{\frac{1}{2}-\gamma} + (\frac{1}{2}-\gamma)\log\frac{\frac{1}{2}-\gamma}{\frac{1}{2}+\gamma}\right]}_{=O(\gamma^2)} \\
&\leqslant C_2 \frac{\gamma^2}{d} H(\omega, \omega').
\end{aligned}$$

Therefore, with $H(\omega, \omega') \leqslant d$, we obtain:

$$\max_{\omega,\omega'} \mathrm{KL}(P_\omega, P_{\omega'}) \leqslant C_2 \gamma^2.$$

Step 4: Conclusion.

To conclude the proof, we set $\gamma = C_3\sqrt{\frac{d}{n}}$. Then we have:

$$\max_{\omega,\omega'} \mathrm{KL}(P_\omega, P_{\omega'}) \leqslant C_4 \frac{d}{n} \leqslant \frac{\log(2^{d/8})}{4n} \leqslant \frac{\log(|\mathcal{P}'|)}{4n},$$

where the last inequality follows from the Varshamov-Gilbert lemma. Then, by the local Fano method, we conclude:

$$R^* \geqslant \frac{\delta}{2} \geqslant C_5\sqrt{\frac{d}{n}}.$$

□

2 Stochastic Bandits

2.1 Introduction

In the next series of lectures, we will discuss sequential/adaptive decision-making problems, where there is a sequence of interactions between a learner and an environment.

Specifically, at each round t, the learner chooses an action $A_t \in \mathcal{A}$, where $\mathcal{A}$ is the set of possible actions. The environment then reveals an observation O_t, and in return, the learner receives a reward $X_t = X_t(O_t, A_t)$. The learner's objective is to maximize the sum of rewards $\sum_{t=1}^{T} X_t$. Stochastic/adversarial bandits and online learning are typical examples of sequential/adaptive decision-making problems. We will begin by focusing on stochastic bandits.

A stochastic bandit problem consists of the following components:

- Let $\nu = \{\nu_a, a \in \mathcal{A}\}$ denote a set of distributions indexed by actions in $\mathcal{A}$. This set ν is referred to as a bandit model and is a subset of some family $\mathcal{P}$.
- At each round t, the learner selects an action $A_t \in \mathcal{A}$ and observes a reward X_t sampled from ν_{A_t}.
- The learner follows a policy $\Pi = (\Pi_t)_{t\in\mathbb{N}}$, where Π_t maps the history $\{(A_s, X_s)\}_{s=1}^{t-1}$ to an action in $\mathcal{A}$.

- If Π is a randomized policy, Π_t maps the history to a probability distribution over $\mathcal{A}$, and an action is then sampled from this distribution. Π can also be a deterministic policy.
- The expected reward of action a is defined as $\mu_a = \mathbb{E}_{X\sim\nu_a}[X]$. Let $a^* \in \arg\max_{a\in\mathcal{A}} \mu_a$ be the optimal action, and let $\mu_* = \mu_{a^*}$ represent the corresponding optimal expected reward.
- The regret after T rounds of interaction is defined as

$$R_T = R_T(\Pi, \nu) = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right],$$

  where $\mathbb{E}$ is taken with respect to the distribution of the action-reward sequence $(A_1, X_1, A_2, X_2, \ldots, A_T, X_T)$ induced by the interaction between the policy Π and the bandit model ν. The quantities μ_a, a^*, and μ_* are functions of the bandit model ν, and can also be written as $\mu_a(\nu)$, $a^*(\nu)$, and $\mu_*(\nu)$, respectively.

When designing an algorithm for bandits, we require at a minimum that $R_T \in \mathcal{O}(T)$, i.e., $\lim_{T\to\infty} \frac{R_T}{T} = 0$. This condition implies that, over time, the learner is able to eventually identify and learn the optimal action (arm).

2.2 K-armed bandits

A K-armed bandit is a stochastic bandit model where the action space consists of K distinct actions, denoted by $\mathcal{A} = [K]$.

- **Stochastic Bandits:** In this setting, the action space is finite, denoted by $\mathcal{A} = [K]$, where K represents the number of distinct actions (arms).
- **Sub-Gaussian Assumption:** We assume each reward distribution ν_i associated with arm i is σ-sub-Gaussian, with the variance parameter σ known. Formally, the set of possible bandit models is given by:

$$\mathcal{P} = \{\nu = \{\nu_i \mid i \in [K]\} \mid \nu_i \text{ is } \sigma\text{-sub-Gaussian for all } i \in [K]\}.$$

- **Ordering of Expected Rewards:** Without loss of generality, we assume that the expected rewards are ordered as follows:

$$1 \geq \mu_1 \geq \mu_2 \geq \cdots \geq \mu_K \geq 0,$$

  where $\mu_i = \mathbb{E}_{X\sim\nu_i}[X]$ represents the expected reward for action i. It is important to note that the learner does not know this ordering.
- **Regret Definition:** Let $\Delta_i = \mu_1 - \mu_i$ represent the *gap* between the optimal arm (arm 1) and arm i. This quantity indicates how much worse the reward from arm i is compared to the optimal arm.

3 Explore-then-Commit

We have stated this algorithm formal in Algorithm 1.

Let $\mathcal{P}$ be the class of σ-sub-Gaussian K-armed bandit models. For all $\nu \in \mathcal{P}$, the regret of the ETC algorithm π_m^{ETC} satisfies:

$$R_T(\pi_m^{\text{ETC}}, \nu) \leq m \sum_{i,\Delta_i>0} \Delta_i + (T - mK) \sum_{i,\Delta_i>0} \Delta_i \exp\left(-\frac{m\Delta_i^2}{4\sigma^2}\right).$$

If we choose $m = K^{-1/3}T^{1/3}$, then

$$\sup_{\nu\in\mathcal{P}} R_T\left(\pi_{K^{-1/3}T^{1/3}}^{\text{ETC}}, \nu\right) \in \tilde{O}\left(K^{1/3}T^{2/3}\right).$$

Algorithm 1 Explore-then-Commit Algorithm

Data: time horizon T, number of exploration rewards m $(\leq T/K)$
- *Exploration phase*: Pull each arm m times in the first mK rounds.
- Let

$$A = \arg\max_{T \in [K]} \widehat{\mu}_i, \quad \text{where } \widehat{\mu}_i = \frac{1}{m} \sum_{s=1}^{mK} \mathbb{1}(A_s = i) x_s$$

- *Commit phase*: Pull arm A for the remaining $T - mK$ rounds

The regret cannot be improved (via tighter analysis and/or better choice of m), and we have

$$\inf_{m \in [T]} \sup_{\nu \in \mathcal{P}} R_T(\pi_m^{\mathrm{ETC}}, \nu) \in \Omega\left(K^{1/3} T^{2/3}\right).$$

4 The Upper Confidence Bound (UCB) Algorithm

The UCB algorithm is based on the principle of *optimism in the face of uncertainty*, where, in each round, we act as though the bandit model is as favorable as is *statistically plausible.* To state the algorithm formally, we first define the upper confidence bounds for each arm at the end of round t as follows:

$$N_{i,t} = \sum_{s=1}^{t} \mathbb{1}(A_s = i),$$

$$\hat{\mu}_{i,t} = \frac{1}{N_{i,t}} \sum_{s=1}^{t} \mathbb{1}(A_s = i) X_s \quad \text{(undefined if } N_{i,t} = 0\text{)},$$

$$e_{i,t} = \sigma \sqrt{\frac{2 \log(1/\delta_t)}{N_{i,t}}} \quad \text{where} \quad \delta_t = \frac{1}{T^2 t} \quad \text{(undefined if } N_{i,t} = 0\text{)}.$$

Thus, $\hat{\mu}_{i,t} + e_{i,t}$ is an upper confidence bound for μ_i, and $\hat{\mu}_{i,t} - e_{i,t}$ is a lower confidence bound for μ_i.

The UCB algorithm works by choosing the arm with the highest upper confidence bound $\hat{\mu}_{i,t-1} + e_{i,t-1}$ at each round. Intuitively, maximizing $\hat{\mu}_{i,t-1} + e_{i,t-1}$ balances *exploitation* (through $\hat{\mu}_{i,t-1}$) and *exploration* (through $e_{i,t-1}$). The formal statement of the UCB algorithm is given below in Algorithm 2.

Algorithm 2 The Upper Confidence Bound Algorithm

Data: time horizon T, number of exploration rounds $m(\leq T/K)$
for $t = 1, \ldots, K$ **do**
| Pull arm t, i.e., set $A_t = t$ and observe $X_t \sim \nu_t$
end
for $t = K+1, \ldots, T$ **do**
| Pull $A_t = \arg\max_{i \in [K]} \hat{\mu}_{i,t-1} + e_{i,t-1}$ and observe $X_t \sim \nu_{A_t}$ ▷ break ties arbitrarily
end

4.1 Theoretical Results

Theorem 2. *Let $\mathcal{P}$ denote the class of σ-sub-Gaussian bandit models, and let $\nu \in \mathcal{P}$. Then the UCB policy satisfies:*

$$R_T(\nu) \leq 3K + \sum_{i:\Delta_i > 0} \frac{24\sigma^2 \log(T)}{\Delta_i}.$$

Moreover, we have the gap-independent bound:

$$\sup_{\nu \in \mathcal{P}} R_T(\nu) \leq 3K + \sigma\sqrt{96KT\log(T)} \in \tilde{\mathcal{O}}\left(\sqrt{KT}\right).$$

The first bound is a *gap-dependent bound*, whereas the second is a *gap-independent bound* or a *worst-case bound*. If the gaps $\Delta_i = \mu_1 - \mu_i$ are large, then $R_T \in \mathcal{O}(\log(T))$; otherwise, $R_T \in \tilde{\mathcal{O}}(\sqrt{KT})$.

4.2 Regret Decomposition

Before proving Theorem 2, we first state the following regret decomposition lemma.

Lemma 1 (Regret decomposition)**.** *This applies to any policy, not just UCB:*

$$R_T(\nu) = \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[N_{i,T}],$$

where the expectation $\mathbb{E}$ *is taken with respect to the action-reward sequence* $(A_1, X_1, A_2, X_2, \ldots, A_T, X_T)$.

Proof

$$\begin{aligned}
R_T &= \sum_{t=1}^{T} (\mu_1 - \mathbb{E}[X_t]) \\
&= \sum_{t=1}^{T} \left(\mu_1 - \mathbb{E}\left[\sum_{i=1}^{K} \mathbb{1}(A_t = i)X_t\right]\right) \\
&= \sum_{t=1}^{T}\sum_{i=1}^{K} \mathbb{E}\left[(\mu_1 - X_t)\mathbb{1}(A_t = i)\right] \\
&= \sum_{i=1}^{K}\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(A_t = i)\mathbb{E}[(\mu_1 - X_t) \mid A_t]\right] \\
&= \sum_{i=1}^{K}\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(A_t = i)(\mu_1 - \mu_i)\right] \\
&= \sum_{i=1}^{K}\sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}(A_t = i)\Delta_i\right] \\
&= \sum_{i=1}^{K} \Delta_i \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}(A_t = i)\right] \\
&= \sum_{i=1}^{K} \Delta_i \mathbb{E}[N_{i,T}].
\end{aligned}$$

□

4.3 Proof of Theorem 2

We assume without loss of generality (w.l.o.g.) that each arm samples rewards $\{y_{i,r}\}_{r\in\mathbb{N}}$, and we observe these samples one-by-one as we pull each arm. Hence, we can write:

$$\hat{\mu}_{i,t} = \frac{1}{N_{i,t}} \sum_{r=1}^{N_{i,t}} y_{i,r}.$$

We now define the following "good" events G_1 and G_i (for all i such that $\Delta_i > 0$):

$$G_1 = \{\forall t > K,\ \mu_1 < \hat{\mu}_{1,t} + e_{1,t}\},$$

$$G_i = \{\forall t > K,\ \mu_i > \hat{\mu}_{i,t} - e_{i,t}\}.$$

Here, G_1 indicates that the true mean is below the UCB, and G_i indicates that the true mean is above the LCB.

Claim 1. *We have* $\mathbb{P}(G_1^c) \leq \frac{1}{T}$, *and* $\mathbb{P}(G_i^c) \leq \frac{1}{T}$.

Proof

$$\begin{aligned}
\mathbb{P}(G_1^c) &= \mathbb{P}\left(\exists t > K \text{ such that } \mu_1 \geq \hat{\mu}_{1,t} + e_{1,t}\right) \\
&\leq \sum_{t>K} \mathbb{P}\left(\mu_1 > \hat{\mu}_{1,t} + e_{1,t}\right) \\
&= \sum_{t>K} \mathbb{P}\left(\mu_1 > \frac{1}{N_{1,t}} \sum_{r=1}^{N_{1,t}} y_{1,r} + \sigma\sqrt{\frac{2\log(1/\delta_t)}{N_{1,t}}}\right) \\
&\leq \sum_{t>K} \sum_{s=1}^{t-K+1} \mathbb{P}\left(\frac{1}{s}\sum_{r=1}^{s}(y_{1,r} - \mu_1) < -\sigma\sqrt{\frac{2\log(1/\delta_t)}{s}}\right) \\
&\leq \sum_{t>K} \sum_{s=1}^{t-K+1} \exp\left(-\frac{s}{2\sigma^2} \cdot \sigma^2 \cdot \frac{2\log(1/\delta_t)}{s}\right) \\
&= \sum_{t>K} \sum_{s=1}^{t-K+1} \frac{1}{T^2 t} \quad (\text{since } \delta_t = \frac{1}{T^2 t}) \\
&\leq \sum_{t>K} \frac{1}{T^2} \leq \frac{1}{T}.
\end{aligned}$$

□

Remark The trick we used in the fourth and fifth steps only works in K-armed bandits. For other bandit models, we typically use martingales.

Acknowledgements

These notes are based on scribed lecture materials prepaared in Fall 2023 by Ransheng Guan and Yamin Zhou