CSE 547/Stat 548: Machine Learning for Big Data **Lecture**

Multi-Armed Bandits: Non-adaptive and Adaptive Sampling

Instructor: Sham Kakade

1 The (stochastic) multi-armed bandit problem

The basic paradigm is as follows:

- K Independent Arms: $a \in \{1, \dots K\}$
- Each arm a returns a random reward R_a if pulled.
  (simpler case) assume R_a is not time varying.
- Game:
  - You chose arm a_t at time t.
  - You then observe:
  $$X_t = R_{a_t}$$
  where R_{a_t} is sampled from the underlying distribution of that arm.

Critically, the distribution over R_a is not known.

1.1 Regret: an "online" performance measure

Our objective is to maximize our long term reward. We have a (possibly randomized) sequential strategy/algorithm $\mathcal{A}$, which is of the form:

$$a_t = \mathcal{A}(a_1, X_1, a_2, X_2, \dots a_{t-1}, X_{t-1})$$

In T rounds, our reward is:

$$\mathbb{E}[\sum_{t=1}^{T} X_t | \mathcal{A}]$$

where the expectation is with respect to the reward process and our algorithm.

Suppose:

$$\mu_a = \mathbb{E}[R_a]\,,$$

and let us assume $0 \leq \mu_a \leq 1$. Also, define:

$$\mu_* = \max_a \mu_a\,.$$

In T rounds and in expectation, the best we can do is obtain $\mu_* T$.

We will measure our performance by our expected *regret*, defined as follows: In T rounds, our (observed) regret is:

$$\mu_* T - \sum_{t=1}^{T} X_t | \mathcal{A}$$

and our *expected regret* is:

$$\mu_* T - \mathbb{E}\left[\sum_{t=1}^{T} X_t | \mathcal{A}\right]$$

where the expectation is with the randomness in our outcomes (and possibly our algorithm if it is randomized).

1.2 Caveat:

Our presentation in these notes will be loose in terms of $\log(\cdot)$ factors, in both K and T. There are multiple good treatments that provide improvements in terms of these factors.

2 Review: Hoeffding's bound

With N samples, denote the sample mean as:

$$\hat{\mu} = \frac{1}{N}\sum_t X_t \; .$$

Lemma 2.1. *Supposing that the X_t's have an i.i.d. distribution and are bounded between 0 and 1, then, with probability greater than $1-\delta$, we have that*

$$|\hat{\mu} - \mu| \le \sqrt{\frac{\log(2/\delta)}{2N}} \; .$$

3 Warmup: A non-adaptive strategy

Suppose we first pull each arm τ times, in an *exploration* phase. Then, for the remainder of the T steps, we pull the arm which had the best observed reward during the exploration phase.

By the union bound, with probability greater than $1-\delta$, for all actions a,

$$|\hat{\mu}_a - \mu_a| \le \mathcal{O}\sqrt{\frac{\log(K/\delta)}{\tau}} \; .$$

To see this, we simply make our error probability to be δ/K, to the total error probability is δ. Thus all the confidence intervals will hold.

During the exploration rounds, our cumulative regret is at most $K\tau$, a trivial upper bound. During the exploitation rounds, let us bound our cumulative regret for the remainder of $T - K\tau$. Note that for the arm i that we pull, we must have that:

$$\hat{\mu}_i \ge \hat{\mu}_{i_*}$$

where i_* is an optimal arm. This implies that

$$\mu_i \ge \mu_* - c\sqrt{\frac{\log(K/\delta)}{\tau}}$$

where c is a universal constant. To see this, note that by construction of the algorithm $\hat{\mu}_i \ge \hat{\mu}_{i_*}$, which implies

$$\mu_i \ge \hat{\mu}_i - |\hat{\mu}_i - \mu_i| \ge \hat{\mu}_{i_*} - |\hat{\mu}_i - \mu_i| \ge \mu_{i_*} - |\hat{\mu}_i - \mu_i| - |\hat{\mu}_{i_*} - \mu_{i_*}| \, ,$$

and the claim follows using the confidence interval bounds.

Hence, our total regret is:

$$\mu_* T - \sum_{t=1}^{T} X_t \le \tau K + \mathcal{O}\sqrt{\frac{\log(K/\delta)}{\tau}}(T - K\tau)$$

Now let us optimize for τ.

Lemma 3.1. *(Regret of the non-adaptive strategy) The total expected regret of the non-adaptive strategy is:*

$$\mu_* T - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right] \le cK^{1/3}T^{2/3}(\log T)^{1/3}$$

where c is a universal constant.

Proof. Choose $\tau = K^{2/3}T^{2/3}$ and $\delta = 1/T^2$. Note that with probability greater than $1 - 1/T^2$, our regret is bounded by $(K^{1/3}T^{2/3}(\log(KT))^{1/3})$. Also, if we 'fail', the largest regret we can pay is T, and this occurs with probability less than $1/T^2$, so the reget is:

$$\begin{aligned}\text{exp. regret} &\le \Pr(\text{no failure event}) * K^{1/3}T^{2/3}(\log(KT))^{1/3} + \Pr(\text{failure event})T \\ &\le c(1 - 1/T^2)K^{1/3}T^{2/3}(\log(KT))^{1/3} + \frac{1}{T}\,.\end{aligned}$$

This shows that the regret is bounded as $O(K^{1/3}T^{2/3}(\log(KT))^{1/3})$. For $T > K$, $\log(KT) \le 2\log T$ (and for $K < T$, the claimed regret bound is trivially true). This completes the proof (for a different universal constant). □

3.1 A (minimax) optimal adaptive algorithm

We will now provide an optimal (up to log factors) algorithms (optimal under the i.i.d. assumption for the rewards are distributed and using that the rewards are upper bounded by 1).

Let $N_{a,t}$ be the number of times we pulled arm a up to time t. The question is what arm should pull a time $t + 1$?

3.2 Confidence bounds

If we don't care about $\log$ factors, then the following is a straightforward argument to see that our confidence bounds will simultaneously hold for *all* times t (from 0 to ∞) and all K arms.

Lemma 3.2. *With probability greater than $1 - \delta$, we will have that for all times $t \ge K$, all $a \in [K]$,*

$$|\hat{\mu}_{a,t} - \mu_a| \le c\sqrt{\frac{\log(t/\delta)}{N_{a,t}}}$$

where c is a universal constant.

Proof. We will actually prove a stronger statement: suppose that we observe the outcome of *every* arm, we will first provide a probabilistic statement for the confidence intervals of all the arms (and for all sample sizes). Let us apply Hoeffding's bound with an error probability of $\delta/(K\tau^2)$. Specifically, for the arm a with τ samples, we have that with probability greater than $1 - \delta/(K\tau^2)$:

$$|\hat{\mu}_{a,\tau} - \mu_a| \le c\sqrt{\frac{\log(\tau K/\delta)}{\tau}}$$

(by a straightforward application of Hoeffding's bound). Now that the total error probability over all arms an over sample size τ is:

$$\sum_a \sum_{\tau=0}^{\infty} \frac{\delta}{K\tau^2} = \delta\pi^2/6$$

(the $\pi^2/6$ is from Basel's problem). Note the sum is finite, which means the error total probability for all of these confidence intervals is less than a constant $* \,\delta$.

We have thus shown the following (note the quantifiers): with probability greater than $1-\delta$, that for *all* arms a and *all* sample sizes $\tau \geq 1$ that:

$$|\hat{\mu}_{a,\tau} - \mu_a| \leq c\sqrt{\frac{\log(\tau K/\delta)}{\tau}}\,,$$

(for a possibility different constant c). Observe that the confidence bounds that any algorithm uses at time t is due to having $N_{a,t}$ samples, so we can now apply the above bound in this case, where:

$$c\sqrt{\frac{\log(N_{a,t}K/\delta)}{N_{a,t}}} \leq c\sqrt{\frac{\log(tK/\delta)}{N_{a,t}}}$$

since $N_{a,t} \leq t$. This shows that these confidence bounds are valid for all times t and all arms a. The proof is completed by nothing for $t \geq K$, $\log(Kt) \leq 2\log t$. □

3.3 The Upper Confidence Bound (UCB) Algorithm

- At each time t,
  - Pull arm:

    $$\begin{aligned} a_t &= \arg\max \hat{\mu}_{a,t} + c\sqrt{\frac{\log(t/\delta)}{N_{a,t}}} \\ &:= \arg\max \hat{\mu}_{a,t} + \text{ConfBound}_{a,t} \end{aligned}$$

    (where $c \leq 10$ is a constant).
  - Observe reward X_t.
  - Update $\mu_{a,t}$, $N_{a,t}$, and $\text{ConfBound}_{a,t}$.

With probability greater than $1-\delta$ all the confidence bounds will hold for all arms and all times t.

3.4 Analysis of UCB

If pull arm a at time t, what is our instantaneous regret, i.e. what is:

$$\mu_* - \mu_{a_t} \leq ?$$

Let i_* be an optimal arm. Note by construction of the algorithm we have, if we pull arm a at time t, then:

$$\hat{\hat{\mu}}_{a,t} + \text{ConfBound}_{a,t} \geq \hat{\mu}_{i_*} + \text{ConfBound}_{i_*} \geq \mu_{i_*}\,,$$

the last step follows due to that μ_{i_*} is contained within the confidence interval for i_*.

Using this we have that:

$$\begin{aligned} \mu_{a_t} &\geq \hat{\mu}_{a,t} - \text{ConfBound}_{a,t} \\ &\geq \hat{\mu}_{i_*} - 2\text{ConfBound}_{a,t} \end{aligned}$$

Theorem 3.3. *(UCB regret) The total expected regret of UCB is:*

$$\mu_* T - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right] \leq c\sqrt{KT\log T}$$

for an appropriately chosen universal constant c.

Proof. The expected regret is bounded as:

$$\begin{aligned}\mu_* T - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right] &\leq 2\sum_t \text{ConfBound}_{a,t} \\ &\leq 2c\sum_t \sqrt{\frac{\log(t/\delta)}{N_{a,t}}} \\ &\leq 2c\sqrt{\log(T/\delta)N_{a,T}}\,. \qquad (1)\end{aligned}$$

Note the following constraint on the $N_{a,T}$'s must hold:

$$\sum_a N_{a,T} = T$$

One can now show the worst case setting of $N_{a,T}$ that makes Equation 1 as large as possible (subject to this constraint on the $N_{a,T}$'s) is when $N_{a,t} = T/K$. Finally, to obtain the expected regret bound, the proof is identical to that of the previous argument (in the non-adaptive case, where we choose $\delta = 1/T^2$). □