1 Stochastic Bandits

Suppose each arm has a distribution (say in $[0,1]$), with means $\mu_1, \ldots, \mu_k$. We need to choose which arm I_t to pull in time step t (which gives us expected reward μ_{I_t}) and minimize the expected regret against the maximum arm

$$\mathbb{E} \sum_t (\mu^* - \mu_{I_t}),$$

where $\mu^* = \max_i \mu^i$ (we may use i^* as the argmax).

We consider the algorithm UCB: Let $N_i(t-1)$ denote the number of pull of arm i by time $t-1$ (including it). For an integer s let

$$\hat{\mu}^i_s := \frac{X^1_1 + \ldots + X^1_s}{s}$$

be the empirical mean after seeing s samples from arm i, and define the upper confidence gap $g_s := \sqrt{\frac{10 \ln T}{s}}$. So at time t we have the upper confidence

$$\tilde{\mu}^i_t := \hat{\mu}^i_{N_i(t-1)} + g_{N_i(t-1)}$$

for arm i; UCB just picks the arm with highest upper confidence.

Let $\Delta_i := \mu^* - \mu_i$ be the suboptimality of arm i; since every time we pull arm i in expectation we lose Δ_i compared to pulling i^*, the regret of a strategy is

$$\mathbb{E} \sum_i N_i(T) \Delta_i.$$

Intuition: (See Figure **??**) By Chernoff bounds, we should have that with good probability

$$\hat{\mu}^i = \mu^i \pm \frac{1}{2} g_{N_i(t-1)};$$

so in this "expected state", $\tilde{\mu}^i$ is an over estimator of μ^i (i.e., even if $\hat{\mu}^i$ is at the bottom of the "confidence interval" of μ^i, the upper estimator is at the top). Let us see what happens when this estimate is too optimistic. Suppose again we are in the "expected state". Suppose further that the upper estimate of i makes it look better than i^*; this is only possible if the top of the interval $\mu^i + \frac{3}{2} g_{N_i(t-1)}$ is bigger than μ^*. In this case we will pull arm i, but this makes us refine our estimate and reduces g_s; after many mistakes, $\mu^i + \frac{3}{2} g_{s'} \leq \mu^*$, and as long as our estimates are still in the expected state it means that $\tilde{\mu}^i_{s'} \leq \tilde{\mu}^*$, and so we will not pull the arm anymore.

Theorem 1. *UCB has regret at most*

$$\textit{regret UCB} \lesssim \sum_{i:\Delta_i > 0} \left(\frac{\ln T}{\Delta_i} + \Delta_i \right).$$

To prove this theorem we make formal that, with good probability, our estimates are within what Chernoff bounds tells us. The only nontrivial part of this is that the guarantee has to hold w.r.t. a random number of samples $N_i(t-1)$, which can be correlated to the actual samples of arm i; to take care of that we will use a uniform bound via union bound. (This can also be seen as

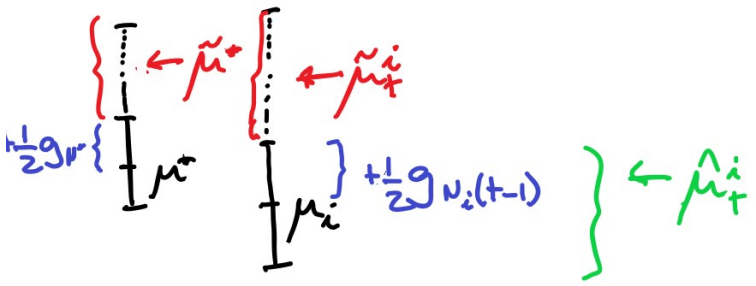

Figure 1: UCB figure

a "self-normalized" version of Chernoff, and in fact can be derived from a sort of self-normalized bound for martingales (i.e., the Freedman-type inequality from one of Rakhlin's papers), if we think about taking all the samples for the arms upfront (but not using them), and revealing them as the algorithm actually probes; then we can condition on all the samples from the arms $\neq i$, and define a stopping time over this sequence, which tells us in each scenario how many of the samples from arm i we used (up until time t); the stopped sum of the samples is precisely the (rescaled) estimate we have in each scenario, and the predictable quadratic variation is related to the square root of the number of samples, etc.)

Lemma 1 (Good estimator for 1 arm, for all times)**.** *Fix an arm i. With probability at least $1-\frac{1}{T}$, for all times t we have*

$$\hat{\mu}^i_{N_i(t-1)} = \mu^i \pm \frac{1}{2} g_{N_i(t-1)}.$$

Proof. By a union bound, it suffices to consider just one time step t. Further, it suffices to obtain the uniform bound over all possible sample sizes:

$$\Pr\left(\forall s \in [T],\, \hat{\mu}^i_s = \mu^i \pm \frac{1}{2} g_s\right) \leq \frac{1}{T^2}.$$

For that, just use Chernoff for a fixed sample size, and take a union bound. □

From the previous lemma and simple manipulations we can relate whp the upper estimates $\tilde{\mu}^i_t$, $\tilde{\mu}^*_t$, the gap Δ_i between these arms, and the size of the confidence $g_{N_i(t-1)}$ (but not dependence in the number of plays of i^*, which is crucial).

Corollary 1. *Fix an arm i. With probability at least $1-\frac{2}{T}$, for all time steps t*

$$\tilde{\mu}^i_t \leq \tilde{\mu}^*_t - \Delta_i + \frac{3}{2} g_{N_i(t-1)}.$$

Proof. Suppose the event from the previous lemma holds for both i and i^*, which happens with probability $1-\frac{2}{T}$. In this case:

$$\tilde{\mu}^i_t = \hat{\mu}^i_t + g_{N_i(t-1)} \leq \mu_i + \frac{3}{2} g_{N_i(t-1)}$$
$$\tilde{\mu}^*_t = \hat{\mu}^* + g_{N^*(t-1)} \geq \mu^* = \mu_i + \Delta_i$$

□

Proof of Theorem **??**. We just need to bound the total expected number of pulls $\mathbb{E}N_i(T)$ for each arm. Fix an arm i. Consider a scenario. If in this scenario the bound of the corollary above holds, then we stop playing arm i whenever the number of plays s is such that $g_s \leq \frac{2}{3}\Delta_i$, which means

$$s \approx \frac{\ln T}{\Delta_i^2};$$

so in such scenario, $N_i(T)$ is at most the RHS above. Otherwise, $N_i(T) \leq T$. For $1 - \frac{2}{T}$ mass of scenarios the bound of the corollary holds, so averaging over all scenarios we have

$$\mathbb{E}N_i(T) \lesssim \frac{\ln T}{\Delta_i^2} + 2. \tag{1}$$

The expected regret is then

$$\sum_i \Delta_i \mathbb{E}N_i(T) \lesssim \sum_i \frac{\ln T}{\Delta_i} + 2\sum_i \Delta_i.$$

This concludes the proof. □

We also have the following guarantee independent of the Δ_i's. Notice that this is the same as in the adversarial setting (actually the latter should be stronger, since it is not pseudo-regret).

Corollary 2. *The regret of UCB is at most*

$$\textit{regret UCB} \lesssim \sqrt{kT \ln T},$$

where k is the number of arms.

Proof. Since $N_i(T) \leq T$, so taking a geometric average with the bound from (**??**) we have $\mathbb{E}N_i(T) \leq \sqrt{\frac{\ln T}{\Delta_i^2} + 2}\sqrt{T}$, so $\Delta_i \mathbb{E}N_i(T) \leq \sqrt{\ln T + 2\Delta_i^2}\sqrt{T} \lesssim \sqrt{T \ln T}$ (recall $\Delta_i \in [0, 1]$). Adding over all arms gives regret $k\sqrt{T \ln T}$, which is slightly worse (notice no square root on k).

To improve this we need to use the info that not only $N_i(T) \leq T$, but $\sum_i N_i(T) \leq T$. So using Cauchy-Schwarz:

$$\sum_i \Delta_i \mathbb{E}N_i(T) = \sum_i \sqrt{\Delta_i^2 \mathbb{E}N_i(T)}\sqrt{N_i(T)} \overset{CS}{\leq} \sqrt{\sum_i \Delta_i^2 \mathbb{E}N_i(T)}\sqrt{\sum_i N_i(T)} \lesssim \sqrt{k \ln T}\sqrt{T}.$$

□

Observation 1. *While the previous Δ-based bound is optimal, the above one is not, one can remove the $\ln T$ term (see Audibert-Bubeck).*