CS 294-128: Algorithms and Uncertainty	Lecture 23	Date: 11/16/2016
Instructor: Nikhil Bansal		Scribe: Yiqun Chen

1 Introduction

In this lecture, we will introduce the famous Upper Confidence Bound (UCB) 1 Policy for stochastic bandits and complete our unfinished characterizations of Markov Paging problem.

2 UCB 1 Policy for Multi-arm Bandit Problem

To quickly recap the setup: we have $1, \cdots, n$ arms and for each arm i, we have an associated random variable X_i bounded in $[0, 1]$ and our goal is to minimize regret (maximize reward) where regret is defined wrt the best fixed strategy a posteriori. Every time we play arm i, the reward is picked independently and randomly with distribution X_i.

We have shown in previous lectures that if we restrict ourselves in the adversarial setting, we can do at best $O(\sqrt{nT \log T})$ so we would like to switch gears to look at stochastic inputs to minimize expected regret. The highlight of our UCB 1 policy is that we will have an almost "to good to be true" upper bound of $O(\log T)$.

To motivate the UCB-1 policy, let's consider the following naive strategy which will achieve $O(nT^{\frac{2}{3}} \ln nT)$ regret.

Definition 1 *$\forall i \in [n]$, denote $\mathbb{E}[X_i] = \mu_i$ and $\mu^* = \max_i \mu_i, \Delta_i = \mu^* - \mu_i > 0$*

Naive Algorithm: If we play each arm i for $T^{\frac{2}{3}}$ steps and play the "best" arm, i.e. the arm with highest empirical reward average, for the rest $T - nT^{\frac{2}{3}}$ steps.

Theorem 2 *The regret of the naive strategy is upper bounded by $O(nT^{\frac{2}{3}} \ln nT)$.*

Let us recall the basic Hoeffding's inequality.

Fact 1 (Hoeffding's Inequality) *Let $Z_1, \cdots, Z_n$ be independent random variables supported on $\mathbb{R}$ s.t. $a_i \leq Z_i \leq b_i$ a.s.. If $S_n = \sum_{i=1}^n Z_i$, then $\forall t > 0, P(S_n - \mathbb{E}[S_n] \geq t) \leq e^{-\frac{2t^2}{\sum_i (b_i - a_i)^2}}$ and a two-sided version is easily attained: $P(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2e^{-\frac{2t^2}{\sum_i (b_i - a_i)^2}}$. In our setting specifically, we have $a_i = 0, b_i = 1, \forall i$ which gives us $2e^{\frac{-2t^2}{n}}$.*

Proof: The high level idea of the proof is that after $T^{\frac{2}{3}}$ steps of play at each arm, our empirically determined best arm cannot be too far from the best arm. We will use $\hat{\mu}_i$ to indicate the observed empirical mean.

By Hoeffding $P(|\frac{1}{k}\sum_{i=1}^k Y_i - \mathbb{E}[Y]| \geq \epsilon) \leq e^{-2\epsilon k^2}$ where Y_i is sampled randomly from distribution Y bounded in $[0, 1]$. Picking $k = T^{\frac{2}{3}}$ and $\epsilon = \sqrt{\frac{\ln(Tn)}{k}}$ gives us $\epsilon^2 k = \ln(Tn)$ and thus $P(\hat{\mu}_i - \mu_i > \epsilon) \leq e^{-2\ln(Tn)} = \frac{1}{n^2T^2}, \forall i$. By union bound over all arms, we get $P(\exists i, \hat{\mu}_i - \mu_i > \epsilon) \leq n\frac{1}{n^2T^2} \leq \frac{1}{nT}$.

If the above event happens, we bound regret by T. Exploration phase gives regret $nT^{\frac{2}{3}}$ and the final phase conditioned that the rare event did not happen is at most ϵT. Thus the overall expected regret is $(1/nT) \cdot T + nT^{\frac{2}{3}} + \epsilon T$. Putting the bound for ϵ gives the result. □

2.1 UCB1

The problem with the above algorithm is that keeps trying a suboptimal policy even if has no chance of being the best. We now give a much more interesting policy.

Let us assume that we know T.

UCB-1 Policy: For each arm, maintain the following statistic at each round: $Index_i(t_i) = \hat{\mu}_i + \sqrt{\frac{\ln T}{t_i}}$ where t_i is the number of times we have played arm i.

At each round t, we play arm i with the highest Index.

Theorem 3 *$\forall i \in [n]$, denote $\mathbb{E}[X_i] = \mu_i$ and $\mu^* = \max_i \mu_i, \Delta_i = \mu^* - \mu_i > 0$, UCB-1 policy achieves regret $\leq \sum_{i \neq i^*} (\frac{4 \ln T}{\Delta_i} + 2\Delta_i)$ where i^* is the index of r.v. with maximum expectation.*

The key observation behind the proof is the following.

Fact 2 *$P(Index_i \leq \mu_i,$ after t_i rounds of playing arm $i)$ is*

$$P(|\hat{\mu}_i - \mu_i| \geq \sqrt{\frac{lnT}{t_i}}) \overset{Hoeffding}{\leq} e^{-2t_i \frac{lnT}{t_i}} = \frac{1}{T^2}$$

. By union bound over all rounds T, $P(Index_i \leq \mu_i,$ at all rounds of playing arm $i) \leq \frac{1}{T}$.

So, let us condition on the event that each index stays above μ_i.

Claim 4 *Let i^* denote the index of optimal arm and $\forall i \neq i^*$, define $Q_i = \sum_{j=1}^{T} 1(t_j = 1) =$ number of times we play arm i, then $\mathbb{E}[Q_i] \leq \frac{4 \ln T}{\Delta_i^2} + 2$*

Proof: We know from the above fact that our index never falls below μ_i. Now let $k_i = \frac{4 \ln T}{\Delta_i^2}, \forall i \neq i^*$, then after k_i steps on arm i we have $Index_i(k_i) = \hat{\mu}_i + \frac{\Delta_i}{2}$.

We are interested in mis-specification of best arm. As index of best arm is at least μ^* (by our conditioning) it suffices to bound

$$P(Index_i(k_i) \geq \mu^*) = P(\hat{\mu}_i + \frac{\Delta_i}{2} \geq \mu^*)$$

$$= P(\hat{\mu}_i - \mu_i \geq \mu^* - \mu_i - \frac{\Delta_i}{2}) = P(\hat{\mu}_i - \mu_i \geq \frac{\Delta_i}{2}) \overset{Hoeffding}{\leq} e^{-2k_i \frac{\Delta_i^2}{4}} = \frac{1}{T^2}$$

. Now this means that w.p. $\geq 1 - \frac{2}{T}$, we will not pick $i \neq i^*$ after k_i trials which gives us $\mathbb{E}[Q_i] \leq (1 - \frac{2}{T})k_i + \frac{2}{T}T = \frac{4 \ln T}{\Delta_i^2} + 2$. □

The above directly gives the claimed bound for UCB-1.

Remark 5 *Note that one caveat in our analysis is that our regret are in effect expected regret, i.e. the regret we get in expectation w.r.t. the best fixed strategy in expectation.*

Corollary 6 *In the worst case, overall all random variables X_i's, we have Regret $\approx \sqrt{nT \log T}$.*

Proof: Fix $\epsilon > 0$ and let $S_1 = \{i : \Delta \leq \epsilon\}$ and $S_2 = \{i : \Delta > \epsilon\}$ be the partition of $[n]$, then applying our UCB-1 theorem on $X_i, \forall i \in S_2$ gives an upper bound on regret $\leq \sum_{i \neq i^*, i \in S_2} (\frac{4 \ln T}{\epsilon} + 2)$ (2 follows from the X_i bounded in $[0, 1]$ assumption). Now on set S_1, we are off by at most ϵ at each round so we can loosely bound the regret by ϵT and $Regret_{[n]} = Regret_{S_1} + Regret_{S_2} \leq \frac{4n \ln + -}{\epsilon} + 2n + \epsilon T$ and minimizing the sum by setting $\epsilon = \sqrt{\frac{4n \ln T}{T}}$ gives us the desired upper bound. □

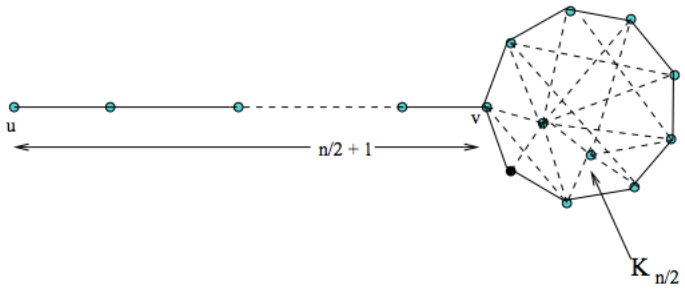

Figure 1: Lollipop Graph

3 Markov Paging

Now let's look at a different stochastic setting with Markov Paging: the set up is a sequence of pages requests and a cache of fixed size and we incur loss 1 whenever we have a miss. Now Markov Paging problem models the incoming sequence characterized by a Markov Chain. In the competitive analysis setting, we know that the coupon collector sequence, i.e. offline evict the page that's farthest in the future gives us $\Omega(\log n)$ competitive ratio where n is the size of our cache.

Claim 7 *Marking algorithm is* $\Omega(\log n)$ *competitive.*

Proof: Our proof heavily relies on the "Lollipop Graph" which is a line segment attached with a complete graph at the end. A phase of our page requests starts with a request at some vertex of the graph, and ends just before all nodes in the graph have been requested because we have exactly one more pages than cache size. The last node to be requested starts a new phase. The important property of the Lollipop Markov Chain is that once a node in the clique is requested, with high probability all the nodes in the clique will be requested before the end node u is requested; thus almost half of our phases will begin with u. Now we will state a series of hitting time claims (H_{vu} denotes the hitting time of starting MC at v and reach u):

Claim 8 *If we start our MC at vertex* v*, it takes* $\Theta(n^3)$ *to reach* u*, i.e.* $H_{vu} = \Theta(n^3)$

Claim 9 $P($*visit* u *before visiting all clique vertices starting from* $v) \leq c\frac{\log n}{n^2}$

Proof: A nice proof is linked in reference 4 using the idea of resistance. □

Now consider our paging problem with cache size n and pages coming from these $n+1$ vertices, if a phase starts on u, then any marking algorithm will incur $\Omega(\log k)$ misses in the clique before the phase ends by a standard coupon collector type of argument but alternatively one can just evict u once we reaches v.

□

Theorem 10 *There exists a* $O(Poly(n))$ *algorithm that is* 4*-competitive of the best offline algorithm.*

Proof: We start with the following high level idea: every time we need to evict a page P to take in a newly requested page R, if P is also missing from OPT's cache, we can just evict without any problem and if P is indeed in OPT's cache, we will look for some Q that is not in OPT's cache s.t. $charge(P) = Q$ where P(Q being requested before P is requested again) $\geq \frac{1}{c}$ and no Q is charged multiple times.

Lemma 11 *The above charging scheme is $c+1$ competitive if our algorithm is deterministic and $2c$ competitive if our algorithm is randomized.*

Lemma 12 (Combinatorial Lemma for LP Formulation) *Suppose we have k pages in cache, define $w(q,p) = P(p$ is requested before q is requested, $\forall p, q \in$ Pages). Then it's obvious that $w(q,p) + w(p,q) = 1$ and we define $w(q,q) = 0$. The lemma states that $\exists$ a dominating distribution $x(p)$ on pages p in cache s.t. $\forall q, \sum_p w(q,p)x(p) \leq \frac{1}{2}$.*

Proof: The high level idea is that if we can prove the existence of such a distribution, we can always solve the following LP to get $x(p)$:

$$\min c$$
$$s.t. \sum_p w(q,p)x(p) \leq c$$
$$\sum_p x(p) = 1$$

.

We refer interested readers to proof of existence outlined under Theorem 3.4 of reference 3. □

Finally we claim the following **dominating distribution algorithm** is 4-competitive (interested readers, again, should check out the proofs in reference 3): every time in the algorithm when we need to evict a page, we evict p w.p. dominating strategy $x(p)$.

□

References

[1] Auer, Peter and Cesa-Bianchi, Nicolò and Fischer, Paul. Machine Learning (2002) 47: 235. doi:10.1023/A:1013689704352

[2] Anna R. Karlin, Steven J. Phillips, Prabhakar Raghavan, Markov Paging, SIAM J. Comput., vol. 30 (2000), pp. 906-922.

[3] Carsten Lund, Steven Phillips, Nick Reingold, Paging against a Distribution and IP Networking, Journal of Computer and System Sciences, Volume 58, Issue 1, February 1999, Pages 222-231

[4] Lecture Notes for CS 271, Alistair Sinclair, https://people.eecs.berkeley.edu/ sinclair/cs271/n23.pdf