STATS 710 – Seq. Dec. Making with mHealth Applications Sep 08, 2016

Lecture 2: Finite-time analysis of ϵ-greedy

Instructors: Susan Murphy and Ambuj Tewari *Scribe: Julie Ghekas*

1 Recap from Lecture 1

Last time, some basic principles were discussed:

(a) basic MAB formation with stochastic (iid) rewards,

(b) regret, as defined by

$$\mathcal{R}_T(\mathcal{L}, (D_a)_{a\in\mathcal{A}}) = \mathbb{E}\left[\mu_\star T - \sum_{t=1}^{T} R_t\right] = \sum_{a:a\in\mathcal{A}} \Delta_a \mathbb{E}\left[N_T(a)\right]$$

(c) Robbins (1952) basic consistency result for the case $D_a \in \mathcal{D} := \{D : \mathbb{E}_{X\sim D}[X] < \infty\}$

1.1 Clarifications from Lecture 1

$\mathcal{L} = (\mathcal{L}_1, \mathcal{L}_2, \ldots)$ is a sequence of maps that map a new action from the previous ones:

$$\mathcal{L}_t : (a_1, r_1, a_2, r_2, \ldots, a_{t-1}, r_{t-1}) \to a_t$$

Thus, $\mathcal{L}$ is an input to the regret formulation as it selects the action that produces the reward.

Almost all algorithms begin with a circuit of sampling each arm once, resulting in $a_1, \ldots a_K$ being predetermined, where K represents the total number of arms. That way, we ensure that each arm is sampled at least once.

The optimal regret of an algorithm in the distribution-free setting with bounded (or, more generally, subgaussian) rewards scales as $O(\sqrt{T})$. However, in the case of algorithms that separate exploration from exploitation the rate is worse: $O(T^{2/3})$. For some formal lower bounds that hold for any algorithm that separates the exploration and exploitation see [DH06] (Section 6) and [GKL16]. Therefore, the minimax regret would not be achieved through separation policy, where the exploration and exploitation phases are separated.

The Robbins algorithm would still work if the rewards were not iid. All we need is the WLLN to hold for rewards. For example, we can allow for various mixing conditions that guarantee asymptotic independence, so this conditions would still permit a consistency proof to go through.

2 Algorithm Design Principles/Heuristics

2.1 Robbins (1952) strategy

One high level strategy is the exploration/exploitation discussed in Lecture 1. This strategy has different names depending on the field. Typically, electrical engineering calls this algorithm "certainty equivalence with forcing", while computer scientists use the term "phased exploration with greedy exploitation". While there may be some differences at the detailed level between these two phrases, they are closely related to each other.

Important characteristics of Robbins' strategy include that we have to be looking at an infinite sequence of time and that there is a vanishing density of any point being an exploration time. Exploration times have to be pre-selected before any process begins.

2.2 ϵ-greedy

This is similar to Robbins' method but the exploration steps are probabilistically selected rather than pre-determined.

In this randomization-based algorithm, a coin is flipped at each round. If it comes up heads, you explore, i.e. take a random action, usually based on a constant/uniform selection of the arms. If it comes up tails, you behave "greedily", i.e. go with the maximum sample mean at that point. One important point to note is that the coin's probability can change with time. A lot of papers use constant ϵ. A potential project idea is to understand mathematically/theoretically the selection of ϵ especially in the presence of non-stationarity of reward distributions.

2.3 Index-based algorithms

One example of an index-based algorithm is the upper confidence bound (UCB) that we will study later. In these algorithms, the exploration phase is not separated from the exploitation phase. Arms that are not sampled enough are giving an "exploration bonus" to entice the algorithm to sample more often from those arms.

2.4 Thompson/posterior sampling

This method follows the principles of Bayesian methodology which is best illustrated in the parametric case.

For a vector of parameters $\overrightarrow{\theta} = (\theta_0, \ldots, \theta_{K-1})$, a prior distribution is assigned to $\overrightarrow{\theta}$: $p(\overrightarrow{\theta})$. For $\overrightarrow{\theta}$, there is an optimal distribution. We can sample from the Bayesian posterior to obtain estimates of the optimal parameter:

$$P(\overrightarrow{\theta} \mid \mathcal{H}_t) \propto p(\overrightarrow{\theta}) P(\mathcal{H}_t \mid \overrightarrow{\theta})$$

In this case, rewards can be designed under a parametric family but can be analyzed outside of these parametric assumptions, landing in more optimistic range of the parameter space.

There are some who are trying to unify the index-based and Thompson algorithms, which could work into another potential project idea.

2.5 Softmax/Boltzmann exploration

This method softens the greedy component of the ϵ-greedy algorithm. Instead of pursuing the max, the softmax is used. In this case, $\mathbb{P}(a_t = a) \propto \frac{\exp(\hat{\mu}_{t-1,a})}{\eta}$, where $\hat{\mu}_{t-1,a}$ represents the estimate of the mean reward of a and η represents the "temperature" (this terminology comes from statistical physics). As $\eta \to 0$, the probability will concentrate on the greedy actions: i.e., those with maximum estimated reward, but this method is a way of softening the selection of the max.

There is no known finite time analysis of this to the best of our knowledge. A variant of Boltzmann exploration changed the softmax to the softmix, which combines softmax with another layer of uniform sampling and behaves similarly to an ϵ-greedy algorithm mixed with softmax [CBF98].

2.6 Bootstrap

Also called the "poor man's Bayes", the bootstrap can be used by subsampling to create a number of parameter estimates. We use the histogram created from the subsampling parameter estimates to get an idea of the parameter distribution. We are unlikely to discuss this method in class but see [EK14, BMM14, OVR15] for related ideas.

3 Finite-time regret analysis of ϵ-greedy

For this section, we assume that all rewards are bounded, say in $[0, 1]$. After sampling each arm once, the following algorithm is executed:

1: **for** $t = 1$ to T **do**
2: Learner selects random arm $a \in \mathcal{A}$ with probability ϵ_t
3: Learner selects $a_t \in \mathcal{A}$ where $a_t := \text{argmax}_{a \in \mathcal{A}} \hat{\mu}_{t-1,a}$ with remaining probability $1 - \epsilon_t$.
4: **end for**

The coin flips used to decide whether the learning algorithm explores by taking a random action are assumed to occur independent of nature.

3.1 Regret bound

Theorem 1. *Suppose D_a has support in $[0, 1], \forall a \in \mathcal{A}$. Define $\epsilon_t = \min\{1, \frac{cK}{d^2 t}\}$, where K is the number of actions possible, t is the time, $c > 5$, and $0 \leq d \leq min_{a:\mu_a<\mu_\star} \Delta_a$ Run ϵ_t-greedy algorithm under these constraints, and fix a suboptimal arm $a \in \mathcal{A}$. Then the probability that ϵ_t-greedy chooses a at t, for $t \geq \frac{cK}{d}$, is at most $\frac{c}{d^2 t} + o(\frac{1}{t})$. That is to say, jointly over the algorithm and nature for every t large enough,*

$$\mathbb{P}(A_t = a) \leq \frac{c}{d^2 t} + o\left(\frac{1}{t}\right)$$

One question associated with this particular theorem is how to select d. One hopes to select the largest d possible, but d also must be less than the unknown gap between a and the optimal arm. The algorithm is sensitive to the choice of d. If $d = 0$, then the bound blows up and becomes useless (especially if the reward is assumed to be bounded).

The rate of growth of the regret in this case is $\log(t)$.

3.2 Corollary

$$\mathcal{R}_T(\epsilon_t\text{-greedy}, (D_a)_{a \in \mathcal{A}}) \leq \frac{c}{d^2} \log(T)(\sum_{a \in \mathcal{A}} \Delta_a)$$

Here, $\sum_{a \in \mathcal{A}} \Delta_a$ represents the suboptimality gap. One does not need to condition the addition on the fact that $\mu_a < \mu_\star$ as before, since if $\mu_a = \mu_\star$, then $\Delta_a = 0$.

We need two facts (concentration inequalities) before we begin the proof (see next lecture).

3.3 Fact 1: Hoeffding-Azuma concentration inequality

Let $X_1, X_2, \ldots$ be random variables with range $[0, 1]$ and $\mathbb{E}(X_t|X_{1:t-1}) = \mu$, where $X_{1:t-1}$ represents $X_1, X_2, \ldots, X_{t-1}$. Then

$$\forall \delta > 0, \mathbb{P}\left(\frac{\sum_{i=1}^{t} X_i}{t} - \mu \geq \delta\right) \leq \exp(-2t\delta^2)$$

and

$$\forall \delta > 0, \mathbb{P}\left(\mu - \frac{\sum_{i=1}^{t} X_i}{t} \geq \delta\right) \leq \exp(-2t\delta^2)$$

Our random variables in this case could be iid, although this is not a requirement for this fact.

3.4 Fact 2: Bernstein concentration inequality

As before, let $X_1, X_2, \ldots$ be random variables with range $[0, 1]$ and $\sum_{i=1}^{t} \mathrm{Var}(X_i \mid X_{1:i-1}) = \sigma_t^2$. Then

$$\forall \delta > 0, \mathbb{P}\left(\sum_{i=1}^{t} X_i - \mathbb{E}\left[\sum_{i=1}^{t} X_i\right] \geq \delta\right) \leq \exp\left(\frac{-\delta^2/2}{\sigma_t^2 + \delta/2}\right)$$

and

$$\forall \delta > 0, \mathbb{P}\left(\mathbb{E}\left[\sum_{i=1}^{t} X_i\right] - \sum_{i=1}^{t} X_i \geq \delta\right) \leq \exp\left(\frac{-\delta^2/2}{\sigma_t^2 + \delta/2}\right)$$

The random variables we will apply Bernstein's inequality to will be independent but not identically distributed.

References

[BMM14] Akram Baransi, Odalric-Ambrym Maillard, and Shie Mannor. Sub-sampling for multi-armed bandits. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases-Volume 8724*, pages 115–131. Springer-Verlag New York, Inc., 2014.

[CBF98] Nicolò Cesa-Bianchi and Paul Fischer. Finite-time regret bounds for the multiarmed bandit problem. In *ICML*, pages 100–108. Citeseer, 1998.

[DH06] Varsha Dani and Thomas P Hayes. Robbing the bandit: Less regret in online geometric optimization against an adaptive adversary. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 937–943. Society for Industrial and Applied Mathematics, 2006.

[EK14] Dean Eckles and Maurits Kaptein. Thompson sampling with the online bootstrap. *arXiv preprint arXiv:1410.4009*, 2014.

[GKL16] Aurélien Garivier, Emilie Kaufmann, and Tor Lattimore. On explore-then-commit strategies. *arXiv preprint arXiv:1605.08988*, 2016.

[OVR15] Ian Osband and Benjamin Van Roy. Bootstrapped thompson sampling and deep exploration. *arXiv preprint arXiv:1507.00300*, 2015.