STATS 710 – Seq. Decision Making with mHealth Applications Sep 13, 2016

Lecture 3: Proof of theorem for ϵ-greedy algorithm

Instructors: Susan Murphy and Ambuj Tewari *Scribe: Henry Oskar Singer*

1 ϵ-greedy algorithm, theorem, and proof

First, it is convenient to introduce some notation.

- Let $\bar{R}^a_t = \frac{1}{t}\sum_{i=1}^t R^a_i$ denote the sample mean of the rewards resulting from making t pulls on arm a.
- Let $*$ denote quantities depending on a^*, the optimal action in $\mathcal{A}$. Take as example the sampling count for a^* after T rounds $N^*_T = N_T\left(a^*\right)$.

Next, we review the ϵ-greedy algorithm [ACBF02].

```
**Initialize:** Play each arm once.
**for** $t = 1$ to $T$ **do**
    With probability $1 - \epsilon_t$, play $\operatorname{argmax}_{a \in \mathcal{A}} \bar{R}^a_t$.
    Otherwise, choose uniformly at random some action $a \in \mathcal{A}$, and play this action.
**end for**
```

Next, we give a theorem that guarantees a high probability of selecting the optimal arm.

Theorem 1. *For reward distribution D_a with support in $[0,1]$, $0 < d \le \min_{a:\mu_a<\mu_*} \Delta_a$. For $c > 5$ and any suboptimal arm $a \in \mathcal{A}$, for $t > \frac{ck}{d^2}$, after t rounds,*

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

Proof. Consider a k-armed bandit and constants c and d as described in Theorem 1. Set $\epsilon_t = \min\left\{1, \frac{ck}{d^2 t}\right\}$, so for $t \ge \frac{ck}{d^2}$, $\epsilon_t = \frac{ck}{d^2 t}$. Set $x_0 = \frac{1}{2}\left(\frac{1}{k}\sum_{i=1}^t \epsilon_i\right)$. The value inside the parentheses corresponds to the expected number of times we will pull an arm due to exploration.

Now, we introduce the first inequality on the road to the final bound. Note that this must be an inequality because of the possibility of tie-breaking among multiple a's or the possibility that multiple a's majorize the empirical mean of a^* with different empirical means.

$$\mathbb{P}\left(A_t = a\right) \le \frac{\epsilon_t}{k} + \left(1 - \epsilon_t\right)\mathbb{P}\left(\bar{R}^a_{N_{t-1}(a)} \ge \bar{R}^*_{N^*_{t-1}}\right), \tag{1}$$

where A_t is the action played at time t. Since the first term on the RHS of this inequality corresponds to the first term in the final bound, we can focus on bounding the second term. Watch out for the switch from $t-1$ to t in the next line.

$$\mathbb{P}\left(\bar{R}^a_{N_{t-1}(a)} \geq \bar{R}^*_{N^*_{t-1}}\right) \;\leq\; \mathbb{P}\left(\bar{R}^*_{N^*_{t-1}} \leq \mu_* - \frac{\Delta_a}{2}\right) + \mathbb{P}\left(\bar{R}^a_{N_{t-1}(a)} \geq \mu_a + \frac{\Delta_a}{2}\right), \tag{2}$$

where μ_a is the population mean of the reward of arm a, and $\Delta_a = \mu_* - \mu_a$. This inequality comes from a union bound on the two events that result in the event inside the parentheses on the LHS. We cannot yet apply the Hoeffding-Azuma inequality because $N_t\left(a\right)$ is a random quantity in the ϵ-greedy algorithm. To account for this, we just need to show that $N_t\left(a\right)$ is "large enough" to result in a sufficiently stable empirical mean estimate. We start by using conditional probability to factor out the randomness of $N_t\left(a\right)$ from $\bar{R}^a_{N_{t-1}(a)}$.

$$\begin{aligned}
\mathbb{P}\left(\bar{R}^a_{N_t(a)} \geq \mu_a + \frac{\Delta_a}{2}\right) \;&=\; \sum_{i=1}^{t} \mathbb{P}\left(N_t\left(a\right) = i \wedge \bar{R}^a_i \geq \mu_a + \frac{\Delta_a}{2}\right) && (3)\\
&=\; \sum_{i=1}^{t} \mathbb{P}\left(N_t\left(a\right) = i \mid \bar{R}^a_i \geq \mu_a + \frac{\Delta_a}{2}\right) \mathbb{P}\left(\bar{R}^a_i \geq \mu_a + \frac{\Delta_a}{2}\right) && (4)\\
&\leq\; \sum_{i=1}^{t} \mathbb{P}\left(N_t\left(a\right) = i \mid \bar{R}^a_i \geq \mu_a + \frac{\Delta_a}{2}\right) \cdot e^{\frac{-\Delta_a i}{2}} && (5)\\
&\leq\; \sum_{i=1}^{\lfloor x_0 \rfloor} \mathbb{P}\left(N_t\left(a\right) = i \mid \bar{R}^a_i \geq \mu_a + \frac{\Delta_a}{2}\right) + \sum_{i=\lfloor x_0 \rfloor + 1}^{t} e^{\frac{-\Delta_a i}{2}}, && (6)
\end{aligned}$$

where line 5 follows from applying Hoeffding-Azuma to the second factor in each term of the sum on the RHS, made possible by the fact that i is not random, and line 6 follows from the observation that the factors we removed are upper-bounded by 1. The choice of which factor to keep in each partition of the two sums comes from the fact that the exponential factor is larger when i is small and smaller when i is large, i.e. we are keeping the smallest possible of the two factors to find a tighter bound. Now, we show a bound on the terms in the first sum for $i \leq \lfloor x_0 \rfloor$.

$$\begin{aligned}
\mathbb{P}\left(N_t\left(a\right) = i \mid \bar{R}^a_i \geq \mu_a + \frac{\Delta_a}{2}\right) \;&\leq\; \mathbb{P}\left(N_t\left(a\right) \leq x_0 \mid \bar{R}^a_i \geq \mu_a + \frac{\Delta_a}{2}\right) && (7)\\
&\leq\; \mathbb{P}\left(N^{\text{explore}}_t\left(a\right) \leq x_0 \mid \bar{R}^a_i \geq \mu_a + \frac{\Delta_a}{2}\right) && (8)\\
&\leq\; \mathbb{P}\left(N^{\text{explore}}_t\left(a\right) \leq x_0\right), && (9)
\end{aligned}$$

where $N^{\text{explore}}_t\left(a\right)$ is the number of times we have pulled arm a in an exploration round. Since this must be upper-bounded by the total number of times we have pulled arm a, line 8 follows. Line 9 follows from the fact that exploration rounds are not dependent on the empirical mean estimators. Overall, we now have

$$\mathbb{P}\left(\bar{R}^a_{N_t(a)} \geq \mu_a + \frac{\Delta_a}{2}\right) \;\leq\; x_0 \cdot \mathbb{P}\left(N^{\text{explore}}_t\left(a\right) \leq x_0\right) + \frac{2}{\Delta_a^2} \cdot e^{\frac{-\Delta_a^2 \lfloor x_0 \rfloor}{2}}. \tag{10}$$

We now bound $\mathbb{P}\left(N_t^{\text{explore}}(a) \leq x_0\right)$. To do this, we need to get an expression for the mean and a bound on the variance of $N_t^{\text{explore}}(a)$ so that we can apply Bernstein's inequality to prove that $\mathbb{P}\left(N_t^{\text{explore}}(a) \leq x_0\right)$ is very small. We start by defining

$$I_{i,a}^{\text{explore}} = \begin{cases} 1 & \text{if } a \text{ pulled at time } i \text{ for exploration} \\ 0 & \text{otherwise} \end{cases} . \tag{11}$$

We can now express $N_t^{\text{explore}}(a)$ as a sum of indicators, $N_t^{\text{explore}}(a) = \sum_{i=1}^t I_{i,a}^{\text{explore}}$. Because we have predetermined and parameterized the randomization in the ϵ-greedy algorithm, we know that $\mathbb{E}\left[I_{i,a}^{\text{explore}} = \frac{\epsilon_i}{k}\right]$. Recalling that $x_0 = \frac{1}{2}\left(\frac{1}{k}\sum_{i=1}^t \epsilon_i\right)$, we can conclude that $\mathbb{E}\left[N_t^{\text{explore}}(a)\right] = 2x_0$. Now, we bound the variance.

$$\begin{aligned} Var\left(N_t^{\text{explore}}(a)\right) &= \sum_{i=1}^t Var\left(I_{i,a}^{\text{explore}}\right) && (12) \\ &\leq \sum_{i=1}^t \frac{\epsilon_i}{k} && (13) \\ &= 2x_0, && (14) \end{aligned}$$

where line 12 follows from the independence of the indicators, and line 13 follows from the decision to explore or exploit being a Bernoulli random variable, which has variance upper-bounded by mean. Applying Bernstein's inequality, we get

$$\begin{aligned} \mathbb{P}\left(N_t^{\text{explore}}(a) \leq x_0\right) &= \mathbb{P}\left(N_t^{\text{explore}}(a) \leq 2x_0 - x_0\right) && (15) \\ &= \mathbb{P}\left(N_t^{\text{explore}}(a) \leq \mathbb{E}\left[N_t^{\text{explore}}(a)\right] - x_0\right) && (16) \\ &\leq e^{\frac{-x_0^2/2}{2x_0 + x_0/2}} && (17) \\ &= e^{\frac{-x_0}{5}}. && (18) \end{aligned}$$

Now, plugging this bound into the inequality 10, we get

$$\mathbb{P}\left(\bar{R}_{N_t(a)}^a \geq \mu_a + \frac{\Delta_a}{2}\right) \leq x_0 \cdot e^{\frac{-x_0}{5}} + \frac{2}{\Delta_a^2} \cdot e^{\frac{-\Delta_a^2 \lfloor x_0 \rfloor}{2}} . \tag{19}$$

Expanding x_0, we get the bound

$$x_0 \geq \frac{c}{d^2} \cdot \log\left(\frac{td^2 e}{2ck}\right) . \tag{20}$$

Note that we bounded just one of the two probabilities from the RHS of equation (2). But the other probability can be bounded in exactly the same way.

Plugging the bounds (19) and (20) back into inequalities (1) and (2) and invoking our chosen expression for the value of ϵ_t, we get the following.

$$\begin{aligned}
\mathbb{P}\left(A_t = a\right) \;&\leq\; \frac{c}{d^2 t} + 2\left(\frac{c}{d^2}\ln\frac{(n-1)\cdot d^2 e^{1/2}}{ck}\right)\left(\frac{ck}{(t-1)\cdot d^2 e^{1/2}}\right)^{c/(5d^2)} \qquad (21)\\
&\quad + \frac{4e}{d^2}\cdot\left(\frac{ck}{(t-1)\cdot d^2 e^{1/2}}\right)^{c/2}. \qquad (22)
\end{aligned}$$

Note that the last two terms are $o(1/t)$ if $c > 5$, and Theorem 1 follows. ■

2 Additional Discussion

2.1 The Sub-Gaussian Property

Theorem 1 still applies to rewards that are distributed with unbounded support as long as the random variables corresponding to the rewards possess the *sub-Gaussian* property, which is defined below.

Definition 1. *A random variable X is sub-Gaussian if and only if $\exists c_1, c_2$ such that*

$$\mathbb{P}\left(|X| > t\right) \leq c_1 e^{-c_2 t^2}.$$

This property ensures that the density of a reward random variable decays like a Gaussian random variable (or faster) as the reward value deviates from the random variable's mean. The key concentration inequality used in the proof above holds not just for sums of iid bounded random variables but also hold for sums of iid subgaussian random variables. See, for example, Proposition 5.10 (Hoeffding type inequality) in Prof. Roman Vershynin's "Introduction to the non-asymptotic analysis of random matrices":

`https://arxiv.org/pdf/1011.3027v7.pdf`

2.2 Why $\epsilon_t = \frac{1}{t}$?

It was asked in lecture whether a better rate could be attained by choosing a different setting of ϵ_t. The answer was that this is not possible because even for Bernoulli rewards the $O(\log T)$ distribution-dependent rate is known to be not improvable. See, for example, Theorem 2.2 (distibution-dependent lower bound) in Bubeck and Cesa-Bianchi's survey [BCB+12].

However, in the case where the rewards are distributed with heavier tails, it may require a different choice of ϵ_t (and perhaps even a different estimator of the mean instead of the sample mean) to result in the optimal rates.

References

[ACBF02] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[BCB$^+$12] Sébastien Bubeck, Nicolò Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.