STATS 710 – Seq. Dec. Making with mHealth Applications Sep 20, 2016

Lecture 5: Distribution Independent Bounds and Thompson Sampling with Beta Priors and Bernoulli Rewards

Instructors: Susan Murphy and Ambuj Tewari *Scribe: Hyesun Yoo*

1 Small detour into distribution-independent bounds

1.1 Upper bound

We want to know what happens when the gap Δ_a goes to zero. Suppose $\mathbb{E}[N_T(a)] \leq c\frac{\log T}{\Delta_a^2}$. It seems like as $\Delta_a \to 0$, this bound would blow up. However, we can derive distribution-independent bounds that do not depend on Δ_a. For any threshold $\delta > 0$,

$$\begin{aligned}
\text{expected regret} &= \sum_a \Delta_a \mathbb{E}[N_T(a)] \\
&= \sum_{a:\Delta_a \leq \delta} \Delta_a \mathbb{E}[N_T(a)] + \sum_{a:\Delta_a > \delta} \Delta_a \mathbb{E}[N_T(a)] \\
&\leq \delta \cdot \mathbb{E}\left[\sum_{a:\Delta_a \leq \delta} N_T(a)\right] + \sum_{a:\Delta_a > \delta} \Delta_a c \frac{\log T}{\Delta_a^2} \\
&\leq \delta T + \frac{1}{\delta} K c \log T
\end{aligned}$$

We don't have δ in our algorithm so we can tune it. If we set $\delta = \sqrt{\frac{cK \log T}{T}}$, which makes both terms equal, and plug in, then we would get

$$\text{expected regret} \leq 2\sqrt{cKT \log T}.$$

1.2 Lower bound

Theorem 1. *Fix any learning algorithm, K and T. Then, there exists a MAB problem with Bernoulli rewards such that the expected regret is at least $c'\sqrt{KT}$ for some universal constant c'.*

Proof. **(informal)**. With n iid samples, we cannot estimate the mean to an accuracy better than $\frac{1}{\sqrt{n}}$. Since there must be an arm played at most $\frac{T}{K}$ times, it will be difficult to distinguish the quality of this arm from any other one whose mean is within $\frac{1}{\sqrt{T/K}}$. For example, consider a K-armed bandit problem. Suppose there are $K-1$ Bernoulli(1/2) arms and one Bernoulli($\frac{1}{2} + \sqrt{\frac{K}{T}}$) arm. You can see how the lower bound works here. The learning algorithm will not be able to figure out which arm is optimal and will suffer regret $T \times \sqrt{\frac{K}{T}} = \sqrt{TK}$. □

The actual proof is much longer than the proof sketch given here. The proof of this theorem can be found in [ACBFS02] (see the theorem on p. 13 and its proof on pp. 31-33). We also discussed the lower bound in [LR85] where they used idea of hypothesis testing. Also, there was discussion

on the use of KL divergence. If you want to control how random quantities associated with the learning algorithm behave under a change of measure, KL divergence is a good tool for that.

2 Thompson sampling

2.1 Algorithms

The beta distribution is useful for Bernoulli rewards because, in a Bayesian framework, if the prior is Beta(α, β), then after a single Bernoulli trial, the posterior is either Beta$(\alpha+1, \beta)$ or Beta$(\alpha, \beta+1)$. Thompson sampling (with Bernoulli rewards) keeps track of the number of "successes", S_a, and "failures", F_a, for each arm and updates a beta distribution for the selected arm in each round. Algorithm 2 is a general version of Algorithm 1 in which the rewards are generated from unknown distributions with support $[0, 1]$.

Algorithm 1 Thompson Sampling for Bernoulli Bandits

Using beta distribution as prior.
For each arm a $\in \mathcal{A}$, $S_a = F_a = 0$.
for t=1,2, $\cdots$, **do**
  For each a$\in \mathcal{A}$, sample $\theta_{a,t} \sim Beta(S_a + 1, F_a + 1)$.
  Play arm $A_t = \text{argmax}_{a\in\mathcal{A}}\ \theta_{a,t}$ and collect reward R_t.
  If $R_t = 1$, $S_{A_t} = S_{A_t} + 1$. Else if $R_t = 0$, $F_{A_t} = F_{A_t} + 1$.
end for

Algorithm 2 Thompson Sampling for General Stochastic Bandits

Using beta distribution as prior.
For each arm a $\in \mathcal{A}$, $S_a = F_a = 0$.
for t=1,2, $\cdots$, **do**
  For each a$\in \mathcal{A}$, sample $\theta_{a,t} \sim Beta(S_a + 1, F_a + 1)$.
  Play arm $A_t = \text{argmax}_{a\in\mathcal{A}}\ \theta_{a,t}$ and collect reward $\tilde{R}_t$ ($\tilde{R}_t \in$ [0,1])
  Perform a Bernoulli trial with success probability $\tilde{R}_t$ and observe output R_t.
  If $R_t = 1$, $S_{A_t} = S_{A_t} + 1$. Else if $R_t = 0$, $F_{A_t} = F_{A_t} + 1$.
end for

Comments:

- Note that $S_{a,t}$ and $F_{a,t}$ depend on time, but we suppressed the subscript t.
- For exploration, UCB uses an explicit formula while Thompson sampling seems to explore using its posterior distribution. The search for unifying principles behind UCB and TS algorithms continues. See, e.g., the work of Russo and Van Roy [RVR14].
- As you get more samples, the posterior distributions for suboptimal arms will be eventually get separated from those for the optimal arm.
- If we simply use the posterior mean instead of drawing a sample from the posterior distribution, we might not explore enough.

2.2 Theorem for Thompson Sampling

This algorithm dates back to 1933 [Tho33] but the first finite time regret analysis appeared only in 2012 [AG12]. Our presentation of TS and its proof will follow a later paper [AG13].

Theorem 2. *Assume Bernoulli reward distributions for the K arms with means $\mu_0, \cdots, \mu_{K-1}$, where $\mu_0 > max_{a\geq 1}\mu_a$. Then for any ε,*

$$R_T(\textit{TS with Beta}, (\mathcal{D})_{a\in\mathcal{A}}) \leq (1+\varepsilon)\sum_{a=1}^{K-1}\frac{\log T}{d(\mu_a,\mu_0)}\Delta_a + O(\frac{K}{\varepsilon^2})$$

where $d(\mu_a,\mu_0) := \mu_a \log\frac{\mu_a}{\mu_0} + (1-\mu_a)\log\frac{1-\mu_a}{1-\mu_0}$.

Comments:

- Note that by Pinsker's inequality, $d(\mu_a,\mu_0) \geq 2(\mu_a-\mu_0)^2 = 2\Delta_a^2$.
- The last term in this inequality actually involves μ and Δ_a, which make our bound complicated.
- $d(\mu_a,\mu_0)$ can blow up if the optimal arm is deterministic, i.e. equal to zero or one with probability one.
- At the end of class, there was a discussion on KL-UCB [GC11] whose finite time bound matches, up to constants, the asymptotic lower bound of Lai and Robbins [LR85]

References

[ACBFS02] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The non-stochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[AG12] Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *COLT*, pages 39–1, 2012.

[AG13] Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *AISTATS*, pages 99–107, 2013.

[GC11] Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *COLT*, pages 359–376, 2011.

[LR85] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[RVR14] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[Tho33] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.