EE290S/CS194: ML for Sequential Decision Making under Uncertainty **Fall 2018**

Lecture 19: October 25

Instructor: Prof. Anant Sahai *Scribes: Anran Hu, Hansheng Jiang*

Today's topic: Thompson sampling and why it works.

19.1 Bayesian Setting

Suppose that the set of actions $\mathcal{A}$ is a finite set.

We interact with nature through a "box" that the stochastic nature provides us. And we interface with this "box" in the following way: at time t, we first choose some action $A_t \in \mathcal{A}$, and then the Bandit Machine (which has some uncertainty inside) will give us some observation Y_{t,A_t} which may depend on action A_t. It may happen that there is a Y_t vector generated by the machine which is the list of all observations we would have gotten from other actions. But in the multi-arm bandit setting, we will only be able to see Y_{t,A_t}.

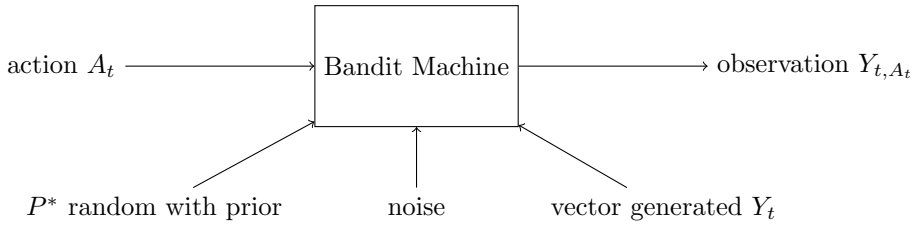

Figure 19.1: Illustration for Bayesian setting.

In this model, each observation Y_{t,A_t} is associated with a reward $R(Y_{t,A_t})$, where the reward function R is fixed and known.

Remark. In a real life setting, when we take an action, we may observe things other than rewards. Other stuff we observe may give us information of other actions which we do not take. Therefore we can not simply throw those away.

Examples:

- under bandit settings, observation (reward) can be Bernoulli, Gaussian, mixed Bernoulli, etc,
- (linear) full information, which we will see more in the next lecture.

19.1.1 Challenge and Goal

Challenge. We don't know how these observations exactly depend on the actions: $\mathbb{P}(\vec{Y}|A)$. But we may have some prior.

This prior can be used to help us learn some observations in a systematic way. The dependence on this prior will be there, but it will not be too strong as time goes by when there are a lot of data.

Denote P^* the true observation distribution that is itself randomly drawn from a family of distributions. We don't know P^* but we have a prior on P^*. After we take some action, we will get some observation based on the action, and we will learn something so that we can update the prior distribution P^*. By updating the prior distribution P^*, we actually mean to update the probability distribution with uncertainty by incorporating the sequence of observations and the prior, i.e. replace it by the posterior distribution.

Let

$$A^*_{(P^*)} = \operatorname*{argmax}_{A\in\mathcal{A}} \mathbb{E}[R(Y_{t,A})|P^*],$$

which is the action that gives us the maximum expected reward (the optimal action). Notice here we assume the model to be time independent.

The regret in the first T rounds is a random variable defined below:

$$\text{Regret}(T) = \sum_{t=1}^{T} [R(Y_{t,A^*}) - R(Y_{t,A_t})].$$

Taking expectation with respect to P^* and noise, we can define the expected regret as follows:

$$\mathbb{E}[\text{Regret}(T)] = \mathbb{E}_{P^*}\mathbb{E}_{\text{noise}}[\text{Regret}(T)].$$

Goal. Our goal is to show that the expected regret is sub-linear in T (in fact we will show it is $O(\sqrt{T})$).

Under this setting, noise is to prevent us from very quickly, perfectly learning which arm is the best. And P^* is the uncertainty of which arm is the best at the beginning.

19.2 Thompson Sampling

19.2.1 Description of the approach

The idea of Thompson sampling is to choose action $A_t = a$ according to the *Posterior Probability* at time t of $a \in A^*$. Here the posterior probability is conditioned on everything happened before time t, i.e.

$$\mathbb{P}_t(\cdot) := \mathbb{P}(\cdot \,|\, (A_1, Y_{1,A_1}), (A_2, Y_{2,A_2}), \cdots, (A_{t-1}, Y_{t-1,A_{t-1}})).$$

Intuition. At the beginning, we are uncertain about which is the best arm and we pull an arm according to their uncertainty. When the uncertainty is high, by nature we will explore. High uncertainty means that the probability has not concentrated on one arm. To be certain about which arm is the best means that the probability of a particular arm being the best is close to 1, while the probability of other arms being the best is close to 0. When that happens, we can only pick that arm for most of the time. In contrast, when we are uncertain, the probability will spread out and we will explore instead. **Thompson sampling will naturally explore when it is uncertain and will pick the "optimal" arm when it is certain. In short, Thompson sampling smoothly navigates this exploration-exploitation trade-off.**

Remark. To update the posterior distribution, one can simply resort to the Bayes formula/rule. For certain prior distribution families, this can be done in closed forms. In general, one may need some simulation-based approaches (e.g., Monte-Carlo methods).

Example: (Bernoulli). Suppose that there are two arms (with distributions $\text{Ber}(\mu_1)$ and $\text{Ber}(\mu_2)$ respectively). If we pull arm 1 and get reward (0 or 1) at time t, then the distribution of μ_1 will be updated while the distribution of μ_2 will stay the same as the last step. The posterior of an arm being the best one (A^*) will also be changed at the same time.

In this example, the computational efficiency of Thompson sampling comes from the ease of both the posterior updates and sampling from the posterior. More specifically, following the Bernoulli example above, $\mathbb{P}(\mu_1 > \mu_2)$ is equal to the probability that a random sample of μ_1 is greater than a random sample of μ_2. So what we do is to sample μ_1 and μ_2 under their current distributions and check which is higher.

Example: (Computing posterior distribution) [TTS] The set of actions is $\mathcal{X} = \{1, \cdots, K\}$, and rewards $R_t = Y_t$ are modeled by conditional probabilities $q_\theta(1|k) = \theta_k$ and $q(0|k) = 1 - \theta_k$. The prior distribution is encoded by vectors α and β, with probability density function given by

$$p(\theta) = \prod_{i=1}^{K} \frac{\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)} \theta_k^{\alpha_k - 1}(1 - \theta_k)^{\beta_k - 1},$$

where Γ denotes the gamma distribution. After one round, a reward is observed, and the distribution parameters are updated according to

$$(\alpha, \beta) \leftarrow (\alpha + R_t \mathbf{1}_{A_t}, \beta + (1 - R_t)\mathbf{1}_{A_t}),$$

where $\mathbf{1}_{A_t}$ is a vector with component A_t equal to 1 and all other components equal to 0.

For bandits with Guassian distribution, its conjugate prior is also Gaussian, thus we can similarly calculate the updated distribution.

Connection to Dynamic Programming. When we view the posterior distributions as states, we can interpret this problem as a Markov Decision Process (MDP), which can be solved by dynamic programming. More explicitly, define the posterior distribution of $p^\star$ based on history up to step t as P_t, then P_t forms an MDP with the transition derived from Bayes rules, *i.e.*, given P_t and A_t, one observes Y_{t,A_t} and P_{t+1} can be computed based on Y_{t,A_t} and P_t using Bayes update. The rewards of the MDP can then be defined as $r(P_t, A_t) := \mathbb{E}_t[R(Y_{t,A_t})]$. In each step, the goal is to maximize the reward to go. The equivalence between finding the optimal policy of the MDP and the original bandit problem is obvious from the correspondence in the above definition. Considering a finite-horizon scenario, then in the last step T there is no need for exploration, and in particular dynamic programming also directly chooses the action A_T that maximizes $r(P_T, A_T)$. However, Thompson sampling would still do some exploration via sampling from the posterior distribution P_T. This shows that Thompson sampling is not optimal. However, in practice, dynamic programming is not practical due to the large (or even infinite) size of the state space of all possible information states P_t. In contrast, Thompson sampling solves this issue by only sampling from the actualy posterior distribution P_t.

Remark. In terms of its implementation and the practical usage, Thompson sampling is great in that it reduces the problem of balancing the exploration-exploitation trade-off into two sub-problems: **posterior updates** and **sampling**.

What you are sampling is the parameter that decides the expected reward. You are **NOT** sampling the reward itself. Instead, You pick the best arm based on the the parameters sampled.

19.2.2 Outline of the regret bound proof

The key insight of the proof is that, **as we learn more, we will have less regret in the future.**

We begin by noticing that Thompson sampling matches the action selection distribution to the posterior distribution of the optimal action, in the sense that $\mathbb{P}_t(A^* = a) = \mathbb{P}_t(A_t = a)$ for all $a \in \mathcal{A}$. Therefore we can see that the regret at time t

$$\begin{aligned}
&\mathbb{E}_t[R(Y_{t,A^*}) - R(Y_{t,A_t})] \\
=&\sum_{a\in\mathcal{A}} \mathbb{P}_t(A^* = a)\mathbb{E}_t[R(Y_{t,a})|A^* = a] - \sum_{a\in\mathcal{A}} \mathbb{P}_t(A_t = a)\mathbb{E}_t[R(Y_{t,a})|A_t = a] \\
=&\sum_{a\in\mathcal{A}} \mathbb{P}_t(A^* = a)(\mathbb{E}_t[R(Y_{t,a})|A^* = a] - \mathbb{E}_t[R(Y_{t,a})]).
\end{aligned}$$

The last equality holds because of the fact that the action A_t is selected based on past observations and independent noise, which means that, conditioned on history, A_t is independent of the observation vector $Y_t = (Y_{t,a})_{a\in\mathcal{A}}$.

Notice that here if the distribution of A^* conditioned on P^* is δ_{A^*}. If the distribution of A_t is also δ_{A^*}, the regret will be 0. If the distribution of A_t is very close to δ_{A^*}, then the regret will be close to 0. Thus the regret implicitly defines a distance between distributions on A^*. In fact, it measures the average difference (in the sense of $\mathbb{P}_t(A^*)$) between rewards generated from $\mathbb{P}_t(Y_{t,a})$ (the posterior predictive distribution at a) and $\mathbb{P}_t(Y_{t,a}|A^* = a)$ (the posterior predictive distribution at a conditioned on a being the optimal action).

To understand why Thompson sampling works and how it works, we need to understand the distance between the two distributions: the exact optimum and what we think is the optimum, how these are getting close to each other and what that means for the regret.

We will use entropy and K-L divergence. Recall that entropy is used to measure uncertainty, while K-L divergence is a natural distance between probabilities.

Below we list the five steps to prove the regret bound.

1) Connect the reward-induced distance on distributions to a ℓ_1-type distance on distributions;

2) Connect the ℓ_1-type distance on distributions to K-L divergence distance (Pinsker Inequality);

3) Define an information ratio;

4) Low regret ($O(\sqrt{T})$) if the information ratio is bounded;

5) Show that the information ratio is bounded.

We remark that here 1) + 2) show the consistency between low regret and small distance between distributions.

Step 1) Connect regret to l_1 distance between distributions.

$$\begin{aligned}
&|\mathbb{E}_P[g(X)] - \mathbb{E}_Q[g(X)]| \qquad \text{for the case } g \in [0,1] \\
=&|\sum_a g(a)P(a) - \sum_a g(a)Q(a)| \\
\leqslant&\sum_a |g(a)| \cdot |P(a) - Q(a)| \\
\leqslant&\sum_a |P(a) - Q(a)| \\
=&\|P - Q\|_1.
\end{aligned}$$

This means that the regret is bounded by ℓ_1 norm. So small ℓ_1 distance leads to low regret.

Step 2) Connect $\|P - Q\|_1$ to $D_{\mathrm{KL}}(P\|Q)$. (Pinsker's inequality)

It turns out that $\|P - Q\|_1 \leqslant \sqrt{2D_{\mathrm{KL}}(P\|Q)}$, i.e.

$$\|P - Q\|_1^2 \leqslant 2D_{\mathrm{KL}}(P\|Q). \tag{19.1}$$

Notice that although the K-L divergence is non-symmetric, here you can choose any order.

The intuition behind this inequality (quadratic-like relation) comes from two aspects:

- $D_{\mathrm{KL}}(P\|Q) \geqslant 0$,
- $D_{\mathrm{KL}}(P\|Q) = 0$ when $P = Q$.

These two observation may indicate that $D_{\mathrm{KL}}(P\|Q)$ will look like a quadratic function in the small neighborhood. The following figure shows the possible "nice bowl shape" of the distance between p and q, $D_{\mathrm{KL}(p\|q)}$, where we fix p and vary q.

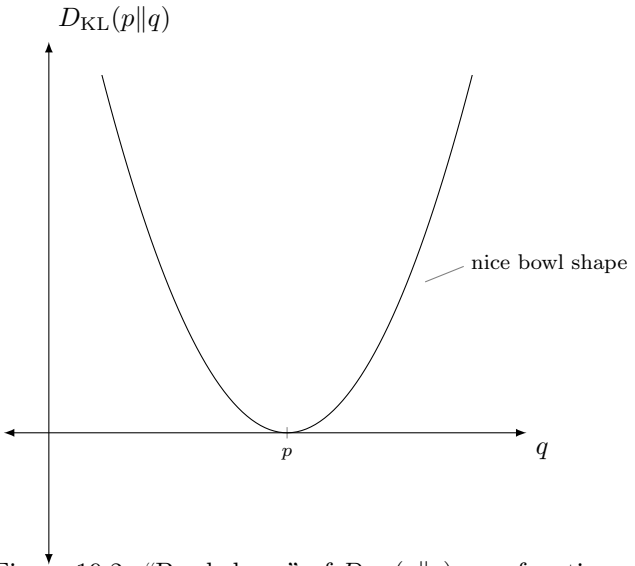

Figure 19.2: "Bowl shape" of $D_{\mathrm{KL}}(p\|q)$ as a function of q.

Proof.

1. For $|\mathcal{A}| = 2$ case.

Without loss of generality, we assume the two distributions P, Q are Bernoulli with $B(p)$ and $B(q)$ respectively, and $p \geqslant q$. To prove (19.1), we need to show that

$$p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q} \geqslant \frac{1}{2}(|p-q| + |1-p-1+q|)^2 = 2(p-q)^2,$$

i.e.

$$f(q,p) := p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q} - 2(p-q)^2 \geqslant 0.$$

Note that equality $f(q,p) = 0$ holds if $q = p$.

For any fixed p, differentiating $f(q,p)$ with respect to q, we obtain

$$-p/q + (1-p)/(1-q) + 4(p-q) = (p-q)(4 - \frac{1}{q(1-q)}) \leqslant 0,$$

where the last inequality follows from $q(1-q) \leqslant 1/4$ for $q \in [0,1]$, i.e. $\frac{\partial f(q,p)}{\partial q} \leqslant 0$ when $q \in [0,p]$.

Thus $f(q,p) \geqslant f(p,p) = 0$ for any $q \in [0,p]$ by monotonicity of $f(q,p)$.

2. Next we show that the general case can be proved by reducing it to $|\mathcal{A}| = 2$.

Let $\Omega = \{\mathcal{A}_1, \mathcal{A}_2\}$, with $\mathcal{A}_1 := \{x \in \mathcal{A} \,|\, P(x) \geqslant Q(x)\}$ and $\mathcal{A}_2 := \{x \in \mathcal{A} \,|\, P(x) < Q(x)\}$.

Let P_Ω and Q_Ω denote the distributions on Ω induced by P and Q respectively, then

$$\begin{aligned}
\|P - Q\|_1 &= \sum_{x \in \mathcal{A}} |P(x) - Q(x)| \\
&= \sum_{x \in \mathcal{A}_1} |P(x) - Q(x)| + \sum_{x \in \mathcal{A}_2} |P(x) - Q(x)| \\
&= \sum_{x \in \mathcal{A}_1} (P(x) - Q(x)) + \sum_{x \in \mathcal{A}_2} (Q(x) - P(x)) \\
&= (P_\Omega(1) - Q_\Omega(1)) + (Q_\Omega(2) - P_\Omega(2)) \\
&= \|P_\Omega - Q_\Omega\|_1.
\end{aligned}$$

We need a math lemma called log-sum inequality, see proof in [LSI].

Log-Sum Inequality: Let $a_1, \cdots, a_n$ and $b_1, \cdots, b_n$ be non-negative numbers, and $a = \sum_{i=1}^n a_i$ and $b = \sum_{i=1}^n b_i$, then

$$\sum_{i=1}^n a_i \ln \frac{a_i}{b_i} \geqslant a \ln \frac{a}{b},$$

with equality holding if and only if $\frac{a_i}{b_i} = c$ for all i, where c is some constant.

By log-sum inequality,

$$\begin{aligned}
D_{\mathrm{KL}}(P\|Q) &= \sum_{x \in \mathcal{A}_1} P(x) \ln \frac{P(x)}{Q(x)} + \sum_{x \in \mathcal{A}_2} P(x) \ln \frac{P(x)}{Q(x)} \\
&\geqslant P_\Omega(1) \ln \frac{P_\Omega(1)}{Q_\Omega(1)} + P_\Omega(2) \ln \frac{P_\Omega(2)}{Q_\Omega(2)} \\
&= D_{\mathrm{KL}}(P_\Omega \| Q_\Omega).
\end{aligned}$$

By results on $|\mathcal{A}| = 2$, we have

$$\|P_\Omega - Q_\Omega\|_1^2 \leqslant 2D_{\mathrm{KL}}(P_\Omega \| Q_\Omega),$$

thus

$$2D_{\mathrm{KL}}(P\|Q) \geqslant 2D_{\mathrm{KL}}(P_\Omega\|Q_\Omega) \geqslant \|P_\Omega - Q_\Omega\|_1^2 = \|P - Q\|_1^2.$$

□

Step 3) Define an information ratio.

We define the information ratio as the following:

$$\Gamma_t := \frac{(\mathbb{E}_t[R(Y_{t,A^*}) - R(Y_{t,A_t})])^2}{I_t(A^*; (A_t, Y_{t,A_t}))},$$

which is the ratio between the square of expected regret and the information gain in period t [RVR16]. Recall that the mutual information between X and Y

$$I(X;Y) = D_{\mathrm{KL}}(P(x,Y)\|P(X)P(Y))$$

is the K-L divergence between the joint distribution of X and Y and the product of the marginal distributions.

Lemma 19.1 *$I(X;Y) = H(X) - H(X|Y)$, where $H(X)$ is the Shannon entropy of X defined as $H(X) = -\sum_{x\in\mathcal{X}} \mathbb{P}(X = x) \log \mathbb{P}(X = x)$.*

The above lemma tells us that the denominator of the information ratio is exactly $H_t(A^*) - H_t(A^*|A_t, Y_{t,A_t})$, which characterizes the information gained by choosing the action A_t.

The information ratio measures how much the incremental regret squared changes with the information gained on the optimal action by choosing the particular arm. It turns out that this ratio is bounded by a constant. We will explain more about it in the next lecture.

References

[RVR16] Daniel Russo and Benjamin Van Roy, An Information-Theoretic Analysis of Thompson Sampling. *The Journal of Machine Learning Research* 17.1 (2016)

[LSI] `https://en.wikipedia.org/wiki/Log_sum_inequality`

[TTS] Russo D J, Van Roy B, Kazerouni A, et al. A tutorial on thompson sampling[J]. Foundations and Trends in Machine Learning, 2018, 11(1): 1-96.