Multi-armed bandits and reinforcement learning

Lecture 4: Introduction to Thompson Sampling

Instructor: Shipra Agrawal Scribes contributed by: Initial scribe by: Erik Waingarten

Thompson Sampling aka *Basyesian posterior sampling* is one of the oldest heuristic for the multi-armed bandit problem. It first appeared in 1933 [1].The algorithm is based on a Bayesian philosophy of learning.

1 Bayesian learning

Consider the problem of learning a parametric distribution from observations. A frequentist approach to learning considers parameters to be fixed, and uses the data learn those parameters as accurately as possible.

For example, consider the problem of learning Bernoulli distribution's parameter (a random variable is distributed as Bernoulli(μ) is 1 with probability μ and 0 with probability $1-\mu$). We are given 10 independent samples:

$$0, 0, 1, 1, 0, 1, 1, 1, 0, 0$$

A frequentist would guess that μ is close to 0.5 with some confidence (probability).

On the other hand, a Bayesian learner maintains a probability distribution to represent his uncertainty about the parameter. The probability distribution represents the chance that the parameter is of a certain value. At the beginning (before seeing the data), this distribution is called the *prior* and it encodes the initial belief a learner has about the value of the parameter. Upon seeing the data, the learner adjusts his/her beliefs using Bayes's rule. This updated distribution is called the *posterior* distribution.

Let's continue with the example from above where we try to learn the parameter μ for a Bernoulli distribution. The learner starts with a prior $p(x)$ representing the learner's prior belief (probability) that μ takes value x:

$$p(x) = \Pr[\mu = x].$$

After observing observe data D (e.g., samples $0, 0, 1, 1, 0, 1, 1, 1, 0, 0$), the learner obtains a posterior distribution, using Bayes rule:

$$\begin{aligned} \Pr[\mu = x|D] &= \frac{\Pr[D|\mu = x] * \Pr[\mu = x]}{\Pr[D]} \qquad (1) \\ &\propto \Pr[D|\mu = x] * p(x) \qquad (2) \end{aligned}$$

Here, $\Pr[D|\mu = x]$ is the probability of generating data D from the Bernoulli distribution with parameter x. This is also called *likelihood.*

1.1 Posterior calculations for Beta priors and Bernoulli i.i.d. observations

We can sometimes solve for closed form solutions of the posterior distribution given the prior and the observations. In particular, for Bernoulli IID samples, if prior is Beta distribution, then the posterior distribution is also given by a Beta distribution.

Definition 1. *A Beta distribution has support* $(0, 1)$ *with two parameters,* (α, β) *with probability density function*

$$f(x : \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}$$

Note $\Gamma(x)$ is called the Gamma function. For integers $x \geq 1$, $\Gamma(x) = (x-1)!$.

Some useful facts about Beta distributions

- $\mathbf{E}[x] = \dfrac{\alpha}{\alpha+\beta}$, where x is distributed from the Beta distribution with parameters (α, β). (Mode is close to expected value. Mode $= \frac{\alpha-1}{\alpha+\beta-2}$ for $\alpha, \beta > 1$)
- $\mathbf{Var}[x] = \dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

Suppose we have a Beta prior with parameters (α, β). We observe one sample $r \in \{0 \text{ (w.p } 1-\mu), 1 \text{ (w.p } \mu)\}$. Then, the calculations below show that the posterior distribution is $\text{Beta}(\alpha + r, \beta + 1 - r)$.

$$\begin{aligned}
\Pr[\mu = \theta | r] &\propto \Pr[r | \mu = \theta] \Pr[\mu = \theta] && (3)\\
&= \text{Bernoulli}_\theta(r)\text{Beta}_{\alpha,\beta}(\theta) && (4)\\
&= \theta^r (1-\theta)^{1-r} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} && (5)\\
&\propto \theta^{\alpha+r-1}(1-\theta)^{\beta-r} && (6)\\
&\propto \text{Beta}_{\alpha+r,\beta+1-r}(\theta) && (7)
\end{aligned}$$

Therefore, if we start with the prior Beta$(1,1)$ which is uniform over $(0,1)$. Then, after observing n i.i.d. samples from Bernoulli distribution, the posterior is Beta$(S_n + 1, F_n + 1)$ where S_n is the number of 1s in the n samples and F_n is the number of 0s. Let $\hat{\mu} = \frac{S_n+1}{n+2}$. Beta$(S_n+1, F_n+1) = $ Beta$(\hat{\mu}(n+1), (1-\hat{\mu}(n+1)))$, with mean $\hat{\mu}$, and variance $\frac{\hat{\mu}(1-\hat{\mu})}{n+2}$. As n increases, the variance of our posterior distribution decreases; and $\hat{\mu}$ converges to the mean μ of the Bernoulli distribution; which means that for large n we have a tighter distribution with mean close to μ.

1.2 Posterior calculations for Gaussian priors and Gaussian i.i.d. observations

We consider the special case when observations are generated from $\mathcal{N}(\mu, 1)$, where μ is the *single unknown parameter* we are trying to learn.

Start with $\mathcal{N}(0,1)$ prior. Then, we show that after observing n i.i.d. samples from $\mathcal{N}(\mu, 1)$, the posterior distribution for parameter μ is $\mathcal{N}(\hat{\mu}_n, \frac{1}{n+1})$. Here $\hat{\mu}_n$ is empirical average of the n samples. $\hat{\mu}_0$ is defined as 0.

We prove by induction. Note that the base case is trivially true. Suppose the posterior after n samples, $\mathcal{N}(\hat{\mu}_n, \frac{1}{n+1})$ We observe another sample $r \sim \mathcal{N}(\mu, 1)$. Then, we show that posterior by Bayes rule is $\mathcal{N}(\hat{\mu}_{n+1}, \frac{1}{n+2})$, where $\hat{\mu}_{n+1} = \frac{n\hat{\mu}+r}{n+1}$.

$$
\begin{aligned}
\Pr(\theta|r) &\propto \Pr(r|\theta)\Pr(\theta) \\
&\propto \exp\{-\frac{1}{2}((r-\theta))^2\} \cdot \exp\{-\frac{(n+1)}{2}(\theta-\hat{\mu}_n)^2\} \\
&\propto \exp\{-\frac{1}{2}(\theta^2(n+2) - 2\theta(r+\hat{\mu}_n(n+1)))\} \\
&\quad \text{(we can ignore all terms inside exponent that do not involve } \theta\text{)} \\
&= \exp\{-\frac{1}{2}(\theta^2(n+2) - 2\theta\hat{\mu}_{n+1}(n+2)\} \\
&\propto \exp\{-\frac{1}{2}(\theta^2(n+2) - 2\theta\hat{\mu}_{n+1}(n+2) + (\hat{\mu}_{n+1})^2(n+2)\} \\
&\quad \text{(we can add any term inside exponent that does not involve } \theta\text{)} \\
&\propto \exp\{-\frac{(n+2)}{2}(\theta-\hat{\mu}_{n+1})^2\}
\end{aligned}
$$

The term on rhs is proportial to pdf of $\mathcal{N}(\hat{\mu}_{n+1}, \frac{1}{n+2})$

2 Thompson sampling algorithm

We present the Thompson sampling algorithm. The algorithm solves the Stochastic IID MAB problem. Recall that an instance is given by a distributions for each arm. We will describe the algorithm and state performance bounds for some special cases, as well as for the general problem.

The Algorithm: Suppose that for each arm reward is generated from some parameteric distribution ν_i. Then, the overall structure of the algorithm is as follows:

- For every arm, start with a prior belief on parameters of the distribution
- On making observations from an arm, update to posterior belief.
- At time t, play every arm with its posterior probability of being the best arm.

We give precise details of this algorithm for special case of a) Bernoulli samples, and b) Gaussian samples.

2.1 Thompson Sampling for Bernoulli MAB

Assume that we have a Bernoulli Multi Armed Bandit (Bernoulli MAB) instance. That is for arm i, every time it is pulled, the reward is generated from $Bernoulli(\mu_i)$. The aim is to learn model parameters $\mu_i, i = 1, ..., N$ for all arms to find the best arm.

For every arm i, the algorithms starts with a uniform prior belief $Beta(1,1)$ about its mean. After $n_{i,t}$ pulls in time $1, \ldots, t$, the algorithm updates its belief to $Beta(S_{i,t}+1, F_{i,t}+1)$, where
$S_{i,t}$: number of 1s in $n_{i,t}$ pulls of arm i
$F_{i,t}$: number of 0s in $n_{i,t}$ pulls of arm i
The initial values of these variables (before any pulls) are set to 0.

(Observe that mean of posterior $Beta(S_{i,t}+1, F_{i,t}+1)$ at time t is same as empirical mean $\hat{\mu}_{i,t}$)

The algorithm, at time t, plays arm i with its probability of being the best. That is, if X_i is a random variable distributed as $Beta(S_{i,t}+1, F_{i,t}+1)$. Then, it plays i with probability $\Pr(X_i > \max_{j \neq i} X_j)$. Note that a quick way to implement this is to generate a sample from $Beta(S_{i,t}+1, F_{i,t}+1$ for each i. And, pull the arm whose sample is largest.

This algorithm achieves the following problem-dependent (or instance-dependent) bound for Bernoulli MAB:

Algorithm 1: Thompson Sampling for Bernoulli MAB using Beta priors

```
foreach $t = 1, 2, \dots,$ do
    For each arm $i = 1, \dots, N$, independently sample $\theta_{i,t} \sim \text{Beta}(S_{i,t-1} + 1, F_{i,t-1} + 1)$
    Play arm $I_t := \arg\max_i \theta_{i,t}$
    Observe $r_t$.
end
```

Theorem 2. *For any instance $\Theta = \{\mu_1, ..., \mu_N\}$ of Bernoulli MAB,*

$$R(T, \Theta) \leq (1 + \epsilon) \sum_{i \neq I^*} \frac{\ln(T)\Delta_i}{KL(\mu_i, \mu^*)} + O(N/\epsilon^2)$$

Recall that we have $\lim_{T\to\infty} \dfrac{R(T,\Theta)}{\ln(T)} \geq \sum_{i \neq I^*} \dfrac{\Delta_i}{KL(\mu_i, \mu^*)}$. Above theorem says that Thompson Sampling matches this lower bound. We also have the following problem independent regret bound for this algorithm.

Theorem 3. *For all Θ,*

$$R(T) = \max_{\Theta} R(T, \Theta) \leq O(\sqrt{NT \log T} + N)$$

For proofs of above theorems, refer to [2].

2.2 Thompson Sampling for Gaussian MAB

Consider instance $\Theta = (\nu_1, \dots, \nu_i)$ of the stochastic MAB problem, where reward r_t on pulling arm i is generated i.i.d. from the Gaussian distribution $\nu_i = \mathcal{N}(\mu_i, 1)$; μ_i is unknown.

We present a Thompson Sampling algorithm using Gaussian priors. As proved earlier, in this case, we can compute a closed form of the posterior. Specifically, if the prior is the Gaussian $\mathcal{N}(0, 1)$, the posterior after time t will be the Gaussian $\mathcal{N}(\hat{\mu}_{i,t}, \frac{1}{n_{i,t}+1})$. Here $n_{i,t}$ is the number of plays of arm i in time $[1, \dots, t]$ and $\hat{\mu}_{i,t} = \frac{1}{(n_{i,t}+1)} \sum_{\tau=1:I_\tau=i}^{t} r_t$. We set $\hat{\mu}_{i,0} = 0$ for all i.

Algorithm 2: Thompson Sampling using Gaussian priors

```
foreach $t = 1, 2, \dots,$ do
    For each arm $i = 1, \dots, N$, sample $\theta_i$ independently from distrinution $\mathcal{N}(\hat{\mu}_{i,t-1}, \frac{1}{n_{i,t-1}+1})$.
    Play arm $I_t := \arg\max_i \theta_i$
    Observe reward $r_t$.
end
```

We prove that this algorithm achieves logarithmic problem-dependent regret bound, *not only for Gaussian MAB*, but for general stochastic MAB!

Further reading: It is not known if this algorithm achieves the asymptotic lower bound on regret for Gaussian MAB, as the previous algorithm did for Bernoulli MAB. [3] provides a Thompson Sampling algorithm for all single parameter exponential family of distributions (which includes the Gaussian MAB we defined above, because we consider onlu μ_i as the "single" unknown parameter) using Jeffrey priors, and prove that it achives aysmptotic lower bounds for this family.

2.3 Thompson Sampling for general stochastic MAB

In general, we may not be able to assume that the distributions ν_i for arm i may not be Bernoulli or Gaussian or even parametric. However Algorithm 2 is still *a valid online algorithm* that can be used even if rewards do not have a Gaussian distrinution. Stricly speaking, this is not a *Bayeisan posterior sampling algorithm* for general stochastic MAB, because the posterior calculations (which were done for Gaussian reward distributions) are no longer valid. However, it *is* an online algorithm, which we prove achieves following regret bounds for general stochastic MAB.

Theorem 4. *For any N armed stochastic MAB instance $\Theta = \{\nu_1, \dots, \nu_N\}$, Algorithm 2 achieves following regret bound:*

$$\boldsymbol{E}[R(T, \Theta)] \le \sum_{i \neq I^*} \frac{18 \log(T\Delta_i^2)}{\mu_i} + \frac{25}{\Delta_i} + O(1)$$

Here, mean of distribution ν_i is μ_i and $\Delta_i = \mu^* - \mu_i$.

We will prove this theorem for two-armed bandits, and provide outline of the proof for N-armed case. Refer to [2] for a complete proof.

3 Bayesian Regret

So far, in our definition of regret, we did not use any priors. If we have a good prior, we would hope we could get better performance. To take that into account, another definition of regret called *Bayesian regret* is often considered in literature, especially when analyzing algorithms like Thompson Sampling which are based on Bayesian posterior sampling.

Given a prior $P(\Theta)$ over instances of the stochastic MAB problem, Bayesian regret is expected regret over instances sampled from the this prior P:

$$\text{Bayesian regret in time } T = \mathbf{E}_{\Theta \sim P}[\mathbf{E}[R(T, \Theta)]],$$

where the first expectation is over the instance and the second expectation is over the algorithm/reward generation.

In comparison, recall that the worst-case regret bounds we have been proving are either of the form

$$\begin{aligned} \forall \Theta, R(T, \Theta) \le \dots \quad & \text{(Problem-dependent } \log(T) \text{ bounds), or} \\ R(T) = \max_\Theta R(T, \Theta) \le \dots \quad & \text{(Problem-independent } \sqrt{T} \text{ bounds)} \end{aligned}$$

Note that the problem-independent bounds on worst-case regret automatically imply the same bound on Bayesian regret. It may be possible to achieve better Bayesian regret bounds given an informative prior.

References

[1] Thompson, William R. “On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.” Biometrika 25.3/4 (1933): 285-294.

[2] S. Agrawal, N. Goyal, ”Further optimal regret bounds for Thompson Sampling”, In Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS), 2013.

[3] Nathaniel Korda, Emilie Kaufmann, and Remi Munos, Thompson Sampling for 1-Dimensional Exponential Family Bandits. NIPS 2013