Stat 260/CS 294-102. Learning in Sequential Decision Problems.

Peter Bartlett

1. Thompson sampling
   - Bernoulli strategy
   - Regret bounds
   - Extensions—the flexibility of Bayesian strategies

Bayesian bandit strategies

Thompson sampling:

- Simple strategy: only requires ability to sample a parameter from the posterior and maximize the expected reward under the sampled parameter.
- Applicable for more complex problems (dependent priors, complex actions, dependent rewards)
- Performs well in practice.
- Strong (frequentist) regret bounds.
- Gittins index has shortcomings:
  only applicable to infinite horizon and independent arm priors;
  complex to compute.

Bayesian Bernoulli bandits

At time t, arm j gives $X_{j,t} \sim$Bernoulli(μ_j) reward.

Suppose we have fixed, unknown parameters μ_j, and we wish to choose a sequence of arms $I_1, I_2, \ldots,$ so as to minimize regret

$$\overline{R}_n = \sum_{t=1}^{n} (\mu_{j^*} - \mu_{I_t}) .$$

Thompson's idea

Frequentist setting, Bayesian strategy:

- Uniform prior on p_j.
- Pick action j with probability that increases with the probability (under posterior distributions) that j is optimal.

Typically, 'Thompson sampling' refers to *probability matching.* That is, the probability of choosing j is set to the probability that j is optimal:

$$\Pr(I_t = j) = \Pr\left(p_j \text{ is the max}\right).$$

Thompson sampling

Given prior $\pi_{j,0} = \pi_0$:

For $t = 1, 2, \ldots,$

1. Draw $p_{1,t}, \ldots, p_{k,t}$ from posteriors $\pi_{j,t-1}$.
2. Play the arm j with maximum $p_{j,t}$.
3. Observe reward $X_{j,t}$.
4. Update posterior:

$$\pi_{j,t}(p) \propto \underbrace{p^{X_{j,t}}(1-p)^{1-X_{j,t}}}_{\text{likelihood}} \underbrace{\pi_{j,t-1}(p)}_{\text{prior}}.$$

Note that $\Pr(I_t = j) = \Pr(p_{j,t} \text{ is max})$.

Thompson sampling

For Beta(1,1) prior:

For $t = 1, 2, \ldots,$

1. Draw $p_{j,t} \sim \text{Beta}(S_{j,t} + 1, F_{j,t} + 1)$ for $j = 1, \ldots, k$.
2. Play $I_t = j$ for j with maximum $p_{j,t}$.
3. Observe reward $X_{I_t,t}$.
4. Update posterior:
   Set $S_{I_t,t+1} = S_{I_t,t} + X_{I_t,t}$.
   Set $F_{I_t,t+1} = F_{I_t,t} + 1 - X_{I_t,t}$.

Regret of Thompson sampling

Theorem: [Agrawal and Goyal]

For every $\mu_1, \ldots, \mu_k$, there is a constant C such that for all $\epsilon > 0$,

$$\overline{R}_n \leq (1+\epsilon) \sum_{j:\Delta_j > 0} \frac{\Delta_j \log n}{d(\mu_j, \mu^*)} + \frac{Ck}{\epsilon^2},$$

where $d(\mu_j, \mu^*)$ is the KL-divergence between Bernoulli distributions.

Regret of Thompson sampling: Proof

As always, $\overline{R}_n = \sum_{j=1}^{k} \mathbb{E}T_j(n)\Delta_j$, where $T_j(n) = \sum_{t=1}^{n} 1[I_t = j]$.

For UCB strategies, we were able to bound regret by showing that the upper bounds are (whp) valid, and that, after enough mistakes, they are sufficiently tight that subsequent mistakes are unlikely. Here, we don't have bounds, we're just sampling from distributions that get more concentrated as we sample more.

Regret of Thompson sampling: Proof

Fix a suboptimal arm $j \neq j^*$. Split according to $\hat{\mu}_j(t-1)$ and $p_{j,t}$:

$$\begin{aligned}
\mathbb{E}T_j(n) &= \sum_{t=1}^{n} \Pr(I_t = j) \\
&= \sum_{t=1}^{n} \left[\Pr\left(I_t = j,\, \hat{\mu}_j(t-1) \leq x_j,\, p_{j,t} \leq y_j\right) \right. \\
&\qquad + \Pr\left(I_t = j,\, \hat{\mu}_j(t-1) \leq x_j,\, p_{j,t} > y_j\right) \\
&\qquad \left. + \Pr\left(I_t = j,\, \hat{\mu}_j(t-1) > x_j\right) \right],
\end{aligned}$$

where we choose $\mu_j < x_j < y_j < \mu^*$ as follows:

Regret of Thompson sampling: Proof

$$\mu_j < x_j < \mu^* \qquad \text{s.t.} \qquad d(x_j, \mu^*) = \frac{d(\mu_j, \mu^*)}{1+\epsilon}$$

$$x_j < y_j < \mu^* \qquad \text{s.t.} \qquad d(x_j, y_j) = \frac{d(x_j, \mu^*)}{1+\epsilon}$$

$$= \frac{d(\mu_j, \mu^*)}{(1+\epsilon)^2}.$$

This ensures that the relevant divergences are only a constant factor different from $d(\mu_j, \mu^*)$.

Regret of Thompson sampling: Proof

Consider those three probabilities:

1. $\Pr\left(I_t = j,\, \hat{\mu}_j(t-1) \le x_j,\, p_{j,t} \le y_j\right)$
   counts the times when both the empirical mean $\hat{\mu}_j(t-1)$ and the sampled $p_{j,t}$ are not too far above their expectations. For this to be small, we need to be sure that j^* is pulled a lot.

2. $\Pr\left(I_t = j,\, \hat{\mu}_j(t-1) \le x_j,\, p_{j,t} > y_j\right)$
   counts the times when $p_{j,t}$ is far above its expectation. As T_j grows, the distribution of $p_{j,t}$ gets more concentrated, so this sum is $O(\log n / d(x_j, y_j))$.

3. $\Pr\left(I_t = j,\, \hat{\mu}_j(t-1) > x_j\right)$ counts the times when the empirical mean $\hat{\mu}_j(t-1)$ is far above its expectation. Its sum is $O(1)$.

Regret of Thompson sampling: Proof

For the 3rd probability: if $\tau_{j,k}$ is the kth time when $I_t = j$, then

$$\begin{aligned}\sum_{t=1}^{n} \Pr\left(I_t = j,\ \hat{\mu}_j(t-1) > x_j\right) &\leq 1 + \mathbb{E}\sum_{k=1}^{n-1} 1\left[\hat{\mu}_j(\tau_{j,k}) > x_j\right] \\ &\leq 1 + \sum_{k=1}^{n-1} \exp\left(-kd(x_j, \mu_j)\right) \\ &\leq 1 + \frac{1}{d(x_j, \mu_j)} = O(1).\end{aligned}$$

Regret of Thompson sampling: Proof

For the 2nd probability:

$$\sum_{t=1}^{n} \Pr\left(I_t = j,\, \hat{\mu}_j(t-1) \le x_j,\, p_{j,t} > y_j\right)$$

$$\le \sum_{t=1}^{n} \Pr\left(I_t = j,\, p_{j,t} > y_j \middle|\, \hat{\mu}_j(t-1) \le x_j\right)$$

$$\le m + \mathbb{E} \sum_{t=\tau_{j,m}+1}^{n} \Pr\left(I_t = j,\, p_{j,t} > y_j \middle|\, \hat{\mu}_j(t-1) \le x_j,\, \mathcal{F}_{t-1}\right),$$

for any choice of m. (Recall that $\tau_{j,k}$ is the kth time when $I_t = j$ and $\mathcal{F}_{t-1}$ is everything up to $t-1$.)

Regret of Thompson sampling: Proof

$$\sum_{t=1}^{n} \Pr\left(I_t = j,\, \hat{\mu}_j(t-1) \le x_j,\, p_{j,t} > y_j\right)$$

$$\le m + \mathbb{E} \sum_{t=\tau_{j,m}+1}^{n} \exp\left(-T_j(t-1)d(x_j, y_j)\right)$$

$$\le \frac{\log n}{d(x_j, y_j)} + \mathbb{E} \sum_{t=\tau_{j,m}+1}^{n} \frac{1}{n}$$

$$\le \frac{\log n}{d(x_j, y_j)} + 1,$$

where we choose $m = \log n / d(x_j, y_j)$ so that

$$\exp\left(-T_j(t-1)d(x_j, y_j)\right) \le \frac{1}{n}.$$

Regret of Thompson sampling: Proof

For the 1st probability, define $\alpha_{j,t} = \Pr\left(p_{j^*,t} > y_j | \mathcal{F}_{t-1}\right)$. Then

$$\sum_{t=1}^{n} \Pr\left(I_t = j,\, \hat{\mu}_j(t-1) \le x_j,\, p_{j,t} \le y_j\right)$$

$$\textbf{(*)} \le \sum_{t=1}^{n} \mathbb{E}\left[\frac{1-\alpha_{j,t}}{\alpha_{j,t}} 1\left[I_t = j^*,\, \hat{\mu}_j(t-1) \le x_j,\, p_{j,t} \le y_j\right]\right]$$

$$\le \sum_{k=0}^{n-1} \mathbb{E}\left[\frac{1-\alpha_{j,\tau_{j^*,k}+1}}{\alpha_{j,\tau_{j^*,k}+1}} \sum_{t=\tau_{j^*,k}+1}^{\tau_{j^*,k+1}} 1\left[I_t = j^*\right]\right]$$

$$\le \sum_{k=0}^{n-1} \mathbb{E}\left[\frac{1}{\alpha_{j,\tau_{j^*,k}+1}} - 1\right] = \cdots = O(1).$$

(Recall $\tau_{j^*,k}$ is the kth time when $I_t = j^*$.)

Regret of Thompson sampling: Proof

(*) Crucial fact: For a bad arm, when its sample average and sampled parameter aren't too large, the probability it's chosen is bounded linearly by the probability that the best arm is chosen.

Lemma: For suboptimal j,

$$\begin{aligned}
&\Pr\left(I_t = j,\, \hat{\mu}_j(t-1) \le x_j,\, p_{j,t} \le y_j \,|\, \mathcal{F}_{t-1}\right) \\
&\le \frac{1-\alpha_{j,t}}{\alpha_{j,t}} \Pr\left(I_t = j^*,\, \hat{\mu}_j(t-1) \le x_j,\, p_{j,t} \le y_j \,|\, \mathcal{F}_{t-1}\right),
\end{aligned}$$

where $\alpha_{j,t} = \Pr\left(p_{j^*,t} > y_j | \mathcal{F}_{t-1}\right)$.

Regret of Thompson sampling: Proof

First, notice that, conditioned on $\mathcal{F}_{t-1}$, $\hat{\mu}_j(t-1)$ is determined. So assume $\hat{\mu}_j(t-1) \leq x_j$.

Define the event $W_j(t) = \{\forall j' \neq j^*,\, p_{j,t} \geq p_{j',t}\}$ (j is the "suboptimal winner"). Then

$$
\begin{aligned}
&\Pr\left(I_t = j^* \middle|\, p_{j,t} \leq y_j,\, \mathcal{F}_{t-1}\right) \\
&\geq \Pr\left(I_t = j^*,\, W_j(t) \middle|\, p_{j,t} \leq y_j,\, \mathcal{F}_{t-1}\right) \\
&= \Pr\left(W_j(t) \middle|\, p_{j,t} \leq y_j,\, \mathcal{F}_{t-1}\right) \Pr\left(I_t = j^* \middle|\, W_j(t),\, p_{j,t} \leq y_j,\, \mathcal{F}_{t-1}\right) \\
&\geq \Pr\left(W_j(t) \middle|\, p_{j,t} \leq y_j,\, \mathcal{F}_{t-1}\right) \Pr\left(p_{j^*,t} > y_j \middle|\, W_j(t),\, p_{j,t} \leq y_j,\, \mathcal{F}_{t-1}\right) \\
&\geq \Pr\left(W_j(t) \middle|\, p_{j,t} \leq y_j,\, \mathcal{F}_{t-1}\right) \Pr\left(p_{j^*,t} > y_j \middle|\, \mathcal{F}_{t-1}\right) \\
&= \Pr\left(W_j(t) \middle|\, p_{j,t} \leq y_j,\, \mathcal{F}_{t-1}\right) \alpha_{j,t}.
\end{aligned}
$$

Regret of Thompson sampling: Proof

On the other side:

$$
\begin{aligned}
&\Pr\left(I_t = j \middle| p_{j,t} \leq y_j,\, \mathcal{F}_{t-1}\right) \\
&\leq \Pr\left(W_j(t),\, p_{j^*,t} \leq y_j \middle| p_{j,t} \leq y_j,\, \mathcal{F}_{t-1}\right) \\
&= \Pr\left(W_j(t) \middle| p_{j,t} \leq y_j,\, \mathcal{F}_{t-1}\right) \Pr\left(p_{j^*,t} \leq y_j \middle| \mathcal{F}_{t-1}\right) \\
&= \Pr\left(W_j(t) \middle| p_{j,t} \leq y_j,\, \mathcal{F}_{t-1}\right) (1 - \alpha_{j,t}).
\end{aligned}
$$

Thompson sampling for bounded rewards

For $X_{j,t} \in [0, 1]$, can use an identical approach. But the only analysis applies to Bernoulli-sampled rewards (which introduce a little extra variance):

For $t = 1, 2, \ldots,$

1. Draw $p_{j,t} \sim \text{Beta}(S_{j,t} + 1, F_{j,t} + 1)$ for $j = 1, \ldots, k$.
2. Play $I_t = j$ for j with maximum $p_{j,t}$.
3. Observe reward $X_{I_t,t}$.
4. Sample $R_t \sim \text{Bernoulli}(X_{I_t,t})$.
5. Update posterior:
   Set $S_{I_t,t+1} = S_{I_t,t} + R_t$.
   Set $F_{I_t,t+1} = F_{I_t,t} + 1 - R_t$.

Thompson sampling

One of the advantages of a Bayesian strategy (even in a frequentist setting, where the parameters are fixed but unknown) is that it is easy to generalize it to settings with rewards and (other) outcomes, where the distributions depend on parameters θ.

This allows:

- Complex actions (e.g., choose four advertisements on a web page; multi-commodity flow with random capacities).
- Limited observations (e.g., only see rewards for complex actions, not for simple ones).
- Dependent rewards.

Thompson sampling

Given parameter space Θ, action space $\mathcal{A}$, outcome space $\mathcal{Y}$, prior π_0 on Θ, likelihood $\ell(y; a, \theta)$, and reward distribution:
For $t = 1, 2, \ldots,$

1. Draw θ_t from posterior π_{t-1}.
2. Play action a_t that maximizes $\mathbb{E}_{\theta_t} R_a$.
3. Observe outcome Y_t.
4. Update posterior:

$$\pi_t(\theta) \propto \ell(Y_t; a_t, \theta)\pi_{t-1}(\theta).$$