Thompson Sampling

David S. Rosenberg

NYU: CDS

March 5, 2021

Contents

1. Bayesian updating for Gaussians
2. Thompson sampling
3. Experimental results

Bayesian updating for Gaussians

Review: Bayesian updating for Gaussian mean

- Consider $R \sim \mathcal{N}\left(q_*, \sigma^2\right)$.
- Suppose we know σ^2, but don't know q_*.
- We'll take a Bayesian approach.

Going Bayesian

When we take a Bayesian approach, we **replace all unknown parameters by unobserved random elements**, and assign a probability distribution to these randoms elements called the "**prior distribution**."

- In our case, $q_* \in \mathbb{R}$ is the only unknown parameter.

Going Bayesian

- We'll replace $q_* \in \mathbb{R}$ by the **random variable** $Q \in \mathbb{R}$.
- Put prior on Q: $Q \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$ for **known** μ_0, σ_0^2.
- Our full Bayesian model is then

$$\begin{aligned} Q &\sim \mathcal{N}\left(\mu_0, \sigma_0^2\right) \\ R_i \mid Q &\sim \mathcal{N}\left(Q, \sigma^2\right), \end{aligned}$$

  where $R_1, \ldots, R_{t-1}$ are conditionally independent given Q.

- Note that every parameter in our Bayesian model is known.

Bayesian updating

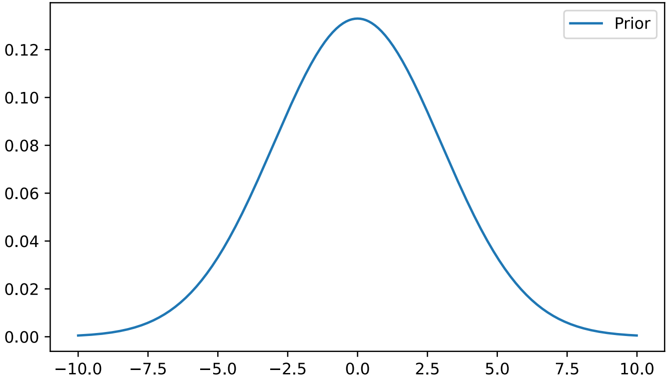
- Our prior distribution on Q is $\mathcal{N}\left(\mu_0, \sigma_0^2\right)$.
- After observing $\mathcal{D}_t = (R_1, \ldots, R_{t-1})$,
  - the posterior distribution on Q is $Q \mid \mathcal{D}_t \sim \mathcal{N}\left(\mu_t, \sigma_t^2\right)$, where

$$\begin{aligned}
\mu_t &= \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1} \left(\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\left(\frac{1}{n}\sum_{i=1}^{n} R_i\right)\right) \\
\sigma_t^2 &= \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}
\end{aligned}$$

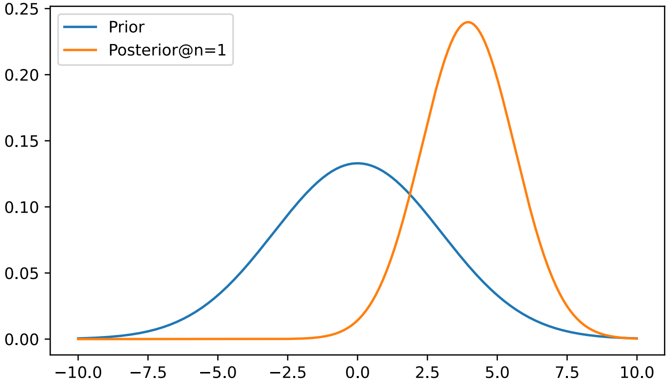
- Posterior mean μ_t is a weighted average of prior mean μ_0 and observed mean.

Gaussian prior distibution

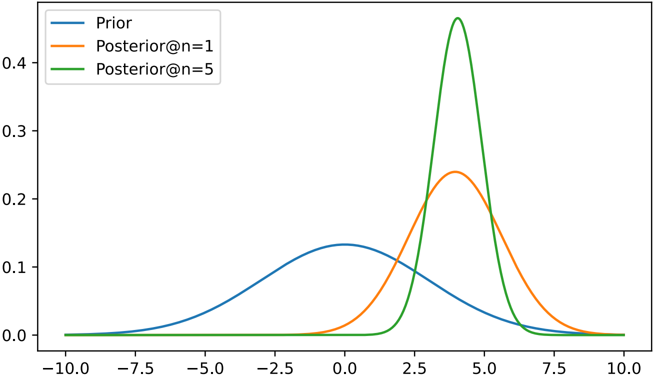
- Consider sampling from $R_1, R_2, \ldots \sim \mathcal{N}(5, \sigma = 2)$.
- Use prior $\mathcal{N}(0, \sigma = 3)$.



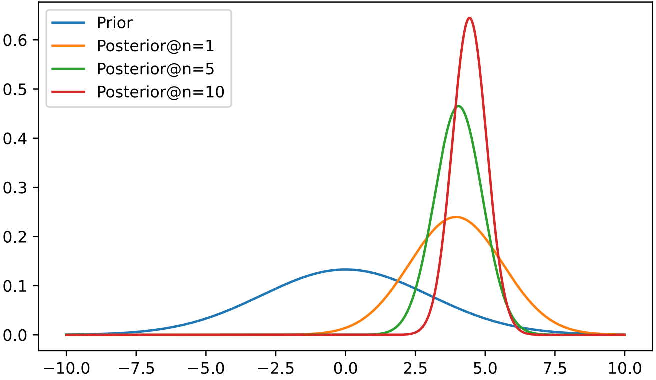
Posterior after $n = 1$ observations



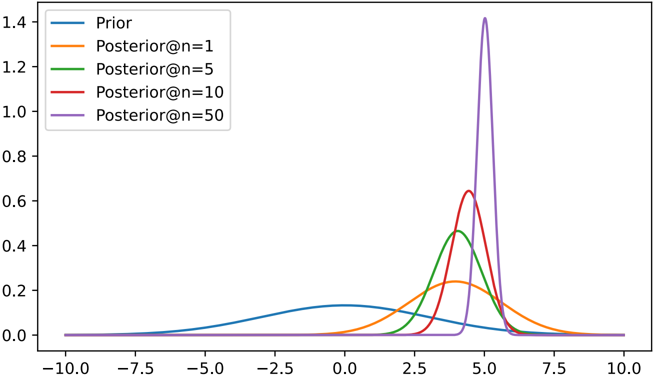
Posterior after $n = 5$ observations



Posterior after $n = 10$ observations

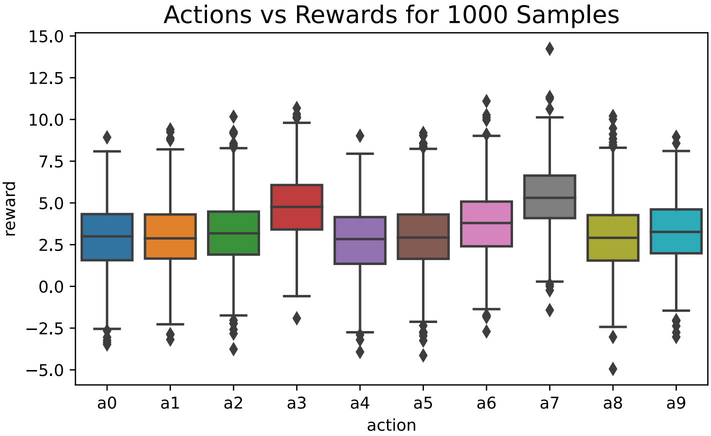


Posterior after $n = 50$ observations



Thompson sampling

Working example: 10-armed bandit



Thompson sampling

- Want to choose action with largest expected reward.
- In Thompson sampling, we take a Bayesian approach.
- We start with a prior on the reward distribution for each action (“arm”).
  - In each round t, we play an action A_t (will see how later).
- We observe reward $R_t(A_t)$.
- We update our posterior reward distribution for action A_t.
- How to choose the action we play?

Reward distribution

- The reward distribution is given by

$$R_i(a) \sim \mathcal{N}(q_*(a), \sigma = 2),$$

  for each action, where $q_*(1), \ldots, q_*(k)$ are **unknown parameters**.
- In a frequentist approach, we would
  - use data to form point estimates and confidence intervals for $q_*(1), \ldots, q_*(k)$
  - use these estimate to choose our actions using various heuristics (ε-greedy, UCB, etc.)
- We'll now take a Bayesian approach...

Gaussian priors

- We will now go Bayesian.
- Need to replace unknown parameter vector $q_* = (q_*(1), \ldots, q_*(k)) \in \mathbb{R}^k$
- Replace with random vector $Q = (Q(1), \ldots, Q(k)) \in \mathbb{R}^k$.
- Our prior distribution is
$$Q(1), \ldots, Q(k) \text{ i.i.d. } \mathcal{N}(0, \sigma = 5).$$
- The full Bayesian distribution is given by
$$\begin{aligned} Q(a) &\sim \mathcal{N}(0, \sigma = 5) \\ R_i(a) \mid Q &\sim \mathcal{N}(Q(a), \sigma = 2), \end{aligned}$$

where $R_1(a), R_2(a), \ldots,$ are conditionally independent given Q, for each a.

Posteriors

- In each round t, we observe $R_t(A_t)$.
- At the beginning of round t, we have observed

$$\mathcal{D}_t = \left((A_1, R_1(A_1)), \ldots, (A_{t-1}, R_{t-1}(A_{t-1}))\right).$$

- Although we never observe $Q = (Q(1), \ldots, Q(k))$,
  - the data $\mathcal{D}_t$ gives us information about it.
- As we gather data, we can update our posterior on Q.
- This is exactly the Gaussian updating we described in the first section,
  - applied separately to $Q(1), \ldots, Q(k)$.

Action choice

.

- We want the reward with the largest expected value.
- If we knew $Q=(Q(1),\ldots,Q(k))$, we would always select action a, where

$$\begin{aligned} a &= \underset{a}{\arg\max}\,\mathbb{E}\left[R(a)\mid Q\right] \\ &= \underset{a}{\arg\max}\,Q(a). \end{aligned}$$

- But we don't observe $Q(a)$.

Bayesian pure exploitation

- At the beginning of round t, a reasonable guess for $Q(a)$ is

$$\mathbb{E}\left[Q(a) \mid \mathcal{D}_t\right],$$

which is the posterior mean of $Q(a)$ conditioned on all our observations so far.

- One possible action strategy would be to choose

$$A_t = \underset{a}{\arg\max}\, \mathbb{E}\left[Q(a) \mid \mathcal{D}_t\right].$$

- This would be **pure exploitation**, since we make no attempt to improve our certainty (i.e. reduce the variance in our posterior) for $Q(a')$, $a' \neq a$.

Probability that an action is the best

- Action a is the best if

$$a = \operatorname*{arg\,max}_a \mathbb{E}\left[R(a) \mid Q\right] = \operatorname*{arg\,max}_a Q(a).$$

- Although we don't know Q, we have a distribution for Q (the posterior).
- Let p_a be the posterior probability that a is the best action:

$$p_a := \mathbb{P}\left(a = \operatorname*{arg\,max}_a (Q(a)) \mid \mathcal{D}_t\right)$$

- If there are ties in the $\arg\max$, we'll choose the numerically smallest action.

Thompson sampling action choice

Thompson sampling action choice

At round t, randomly select action A_t with probability $\mathbb{P}(A_t = a) = p_a$, where

$$p_a := \mathbb{P}\left(a = \arg\max_a (Q(a)) \mid \mathcal{D}_t\right).$$

In words, select action a with probability equal to the posterior probability that action a has the highest expected reward.

- The more certain we are that a is the best in terms of $\mathbb{E}[Q(a) \mid \mathcal{D}_t]$, the more likely we are to select a.
- Thompson sampling is a **heuristic** approach to the explore/exploit tradeoff.
- How to sample from this particular distribution?

The Thompson sampling trick

- Calculating $p_a = \mathbb{P}\left(a = \arg\max_a (Q(a)) \mid \mathcal{D}_t\right)$ may be difficult.

Thompson sampling recipe

1. For each a, draw $Q_t(a) \sim p(q(a) \mid \mathcal{D}_t)$ from the posterior distribution of $Q(a) \mid \mathcal{D}_t$.
2. Choose action $A_t = \arg\max_a Q_t(a)$.

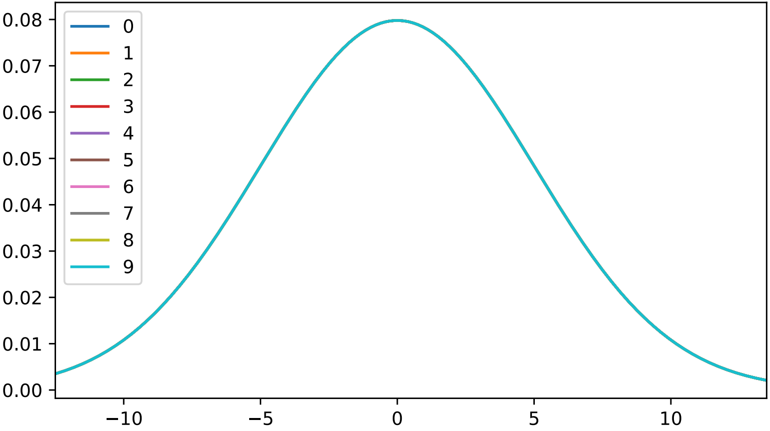
- Note that

$$\begin{aligned} \mathbb{P}(A_t = a) &= \mathbb{P}\left(a = \arg\max_a Q_t(a)\right) \\ &= \mathbb{P}\left(a = \arg\max_a Q(a) \mid \mathcal{D}_t\right) \\ &= p_a. \end{aligned}$$

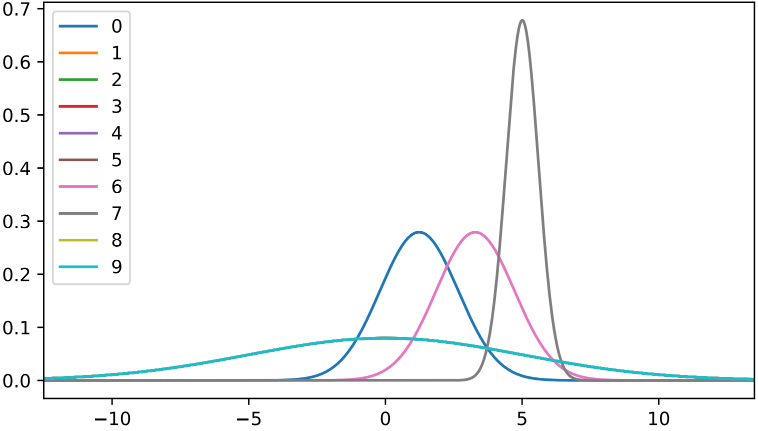
- So A_t has exactly the desired distribution.

Experimental results

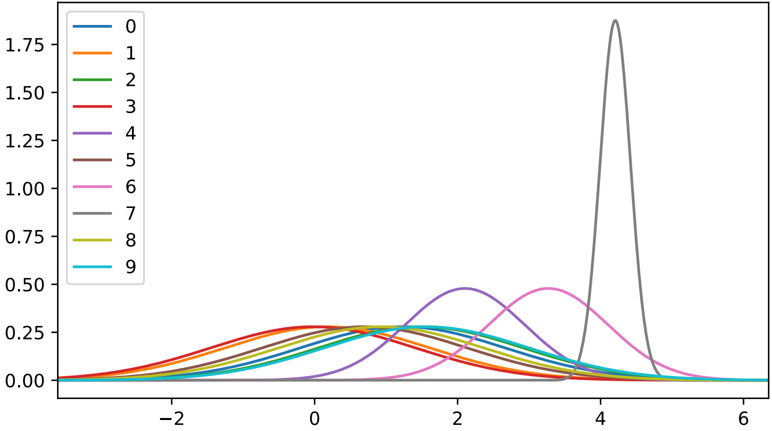
Prior distributions



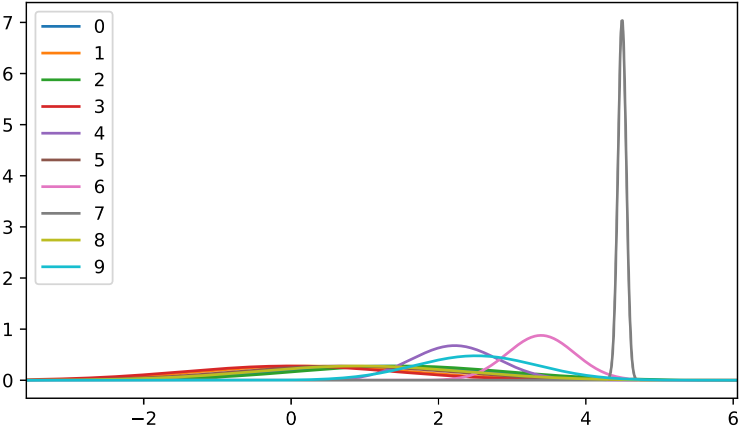
Posterior distributions $n = 5$



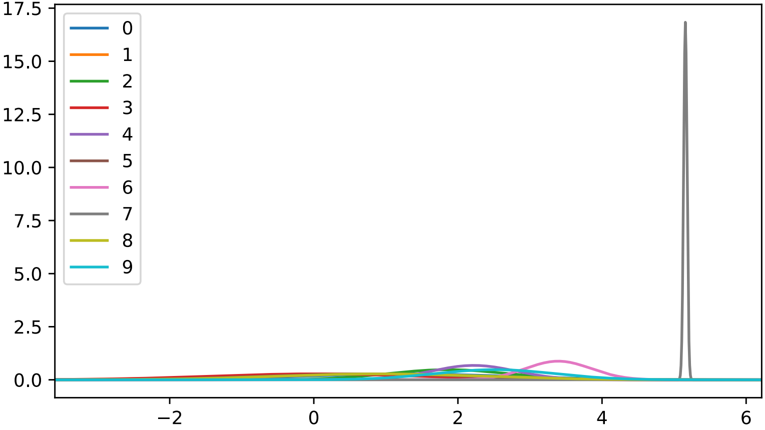
Posterior distribution $n = 20$



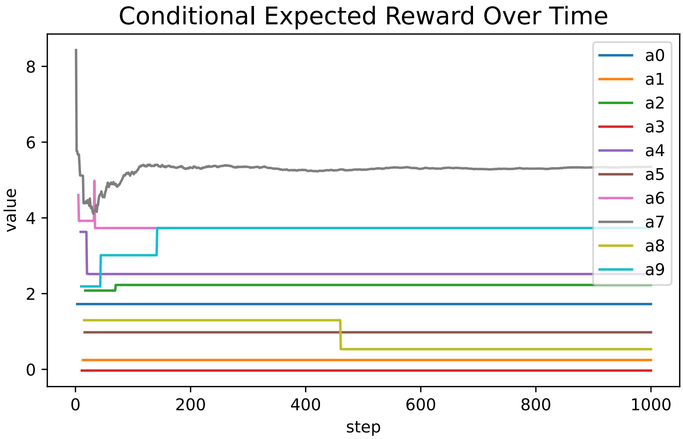
Posterior distribution $n = 50$



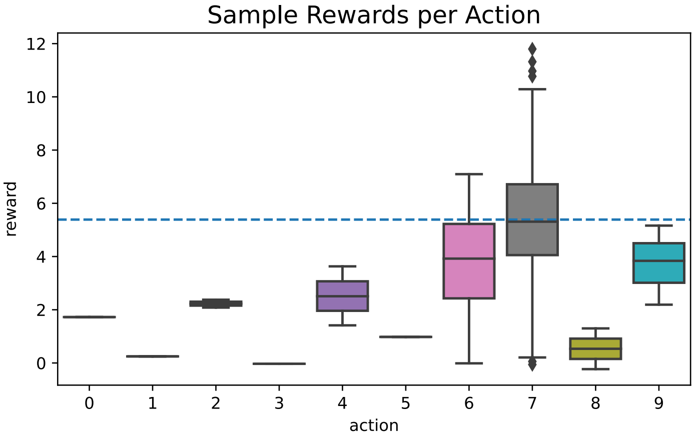
Posterior distribution $n = 100$



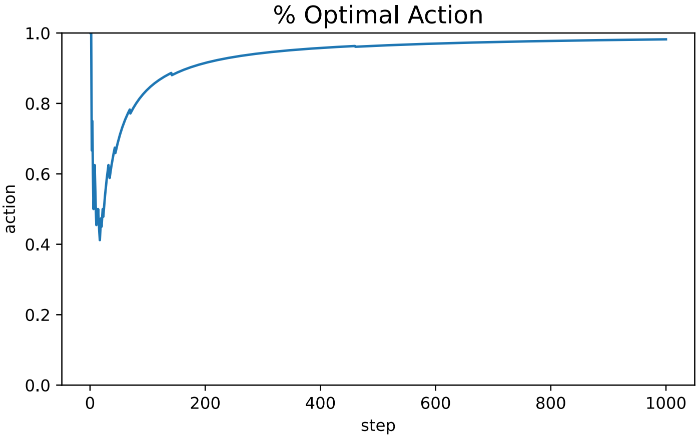
Posterior expected reward



Received rewards by action



Percent optimal action



Tuning parameter?

- What are the "hyperparameters" for Thompson sampling?
- Everything related to the prior distribution.
- In our setting, we can vary the prior variance and see the effect.

strategy	mean	SD	SE
Thompson sampling $\sigma_0 = 2$	5.129	0.306	0.022
Thompson sampling $\sigma_0 = 5$	5.229	0.214	0.015
Thompson sampling $\sigma_0 = 10$	5.279	0.169	0.012

References

Resources

- A Tutorial on Thompson Sampling by Russo et al is a nice [long] tutorial on Thompson sampling [RRK$^+$18].
- You could take a look at Thompson's original work [Tho33] for fun.

References I

[RRK$^+$18] Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen, *A tutorial on thompson sampling*, Foundations and TrendsÂ® in Machine Learning **11** (2018), no. 1, 1–96.

[Tho33] William R. Thompson, *On the likelihood that one unknown probability exceeds another in view of the evidence of two samples*, Biometrika **25** (1933), no. 3/4, 285.