Thompson Sampling

David S. Rosenberg

NYU: CDS

October 6, 2021

Contents

1. Bayesian updating for Gaussians

2. Thompson sampling

3. Experimental results

Bayesian updating for Gaussians

Review: Bayesian updating for a 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 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 μ_0, σ_0^2 that **we choose**.
- 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 $R_1, R_2, \ldots,$ are conditionally independent given Q.
- Note that every parameter in our Bayesian model is known.

- “Every parameter in our Bayesian model is known.” – this is kind of the essence of a Bayesian model.
- In a Bayesian approach, all unknown parameters are replaced with unobserved random variables with known distributions (that we choose – this is the prior distribution).

Bayesian updating

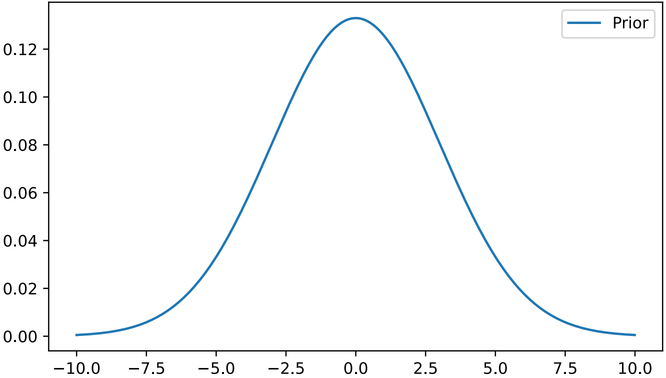
- Our prior distribution on Q is $\mathcal{N}\left(\mu_0, \sigma_0^2\right)$.
- After observing $\mathcal{D}_t = (R_1, \ldots, R_{t-1})$,
  - the posterior distribution on Q is $Q \mid \mathcal{D}_t \sim \mathcal{N}\left(\mu_t, \sigma_t^2\right)$, where

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

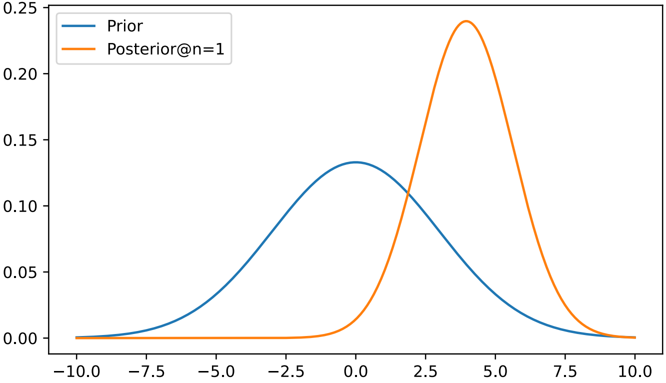
- Posterior mean μ_t is a weighted average of prior mean μ_0 and observed mean.

Gaussian prior distribution

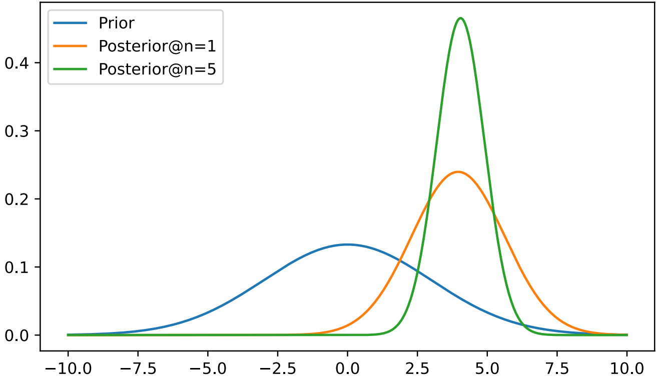
- Consider sampling from $R_1, R_2, \ldots \sim \mathcal{N}(5, \sigma = 2)$.
- Use prior $\mathcal{N}(0, \sigma = 3)$.



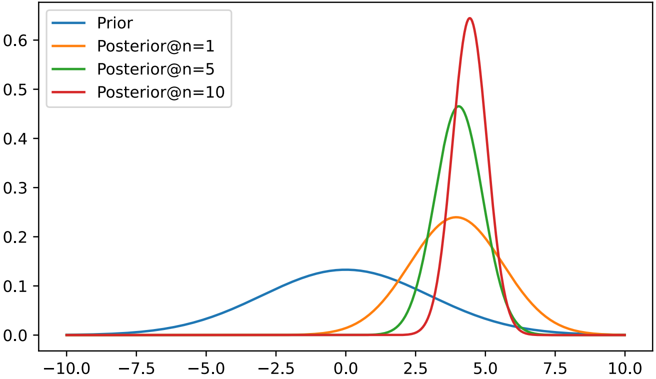
Posterior after $n = 1$ observations



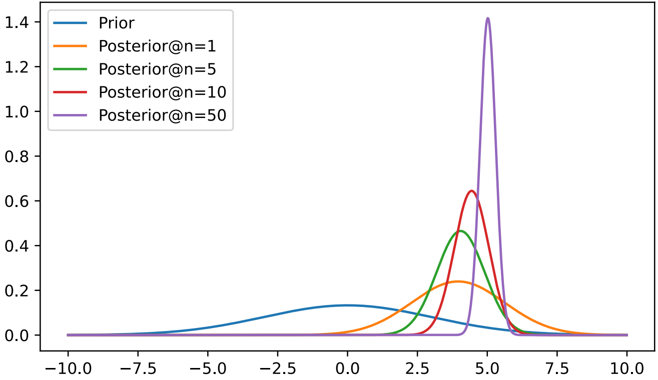
Posterior after $n = 5$ observations



Posterior after $n = 10$ observations

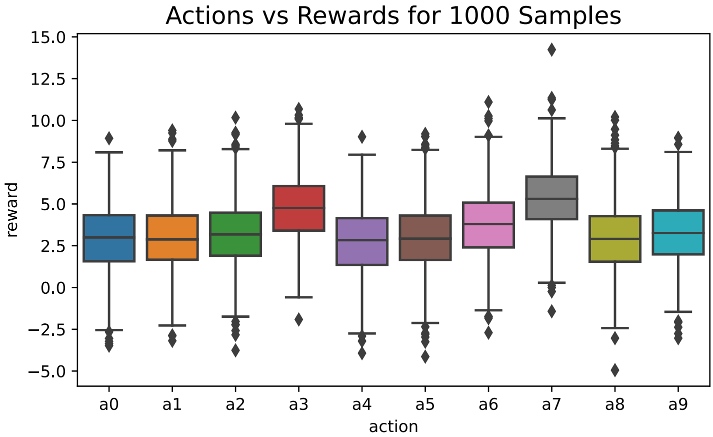


Posterior after $n = 50$ observations



Thompson sampling

Working example: 10-armed bandit



Plot and simulation code courtesy of Ryan Carroll.

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.
- TBD: How to choose the action we play?

Reward distribution

- The reward distribution for action a at round t is given by

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

  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 estimates 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

$$\begin{aligned} Q(a) &\sim \mathcal{N}(0, \sigma = 5) \\ R_t(a) \mid Q &\sim \mathcal{N}(Q(a), \sigma = 2), \end{aligned}$$

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 history

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

- 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.

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

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

- 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

$$a = \arg\max_a \mathbb{E}\left[R(a) \mid Q\right] = \arg\max_a Q(a).$$

- 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:

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

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

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

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

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 $a = \arg\max_a(Q(a))$, the more likely we are to select a.
- Thompson sampling is a **heuristic** approach to the explore/exploit tradeoff.
- How can we sample from this particular distribution?

The Thompson sampling trick

- Calculating $p_a = \mathbb{P}(a = \arg\max_a (Q(a)) \mid \mathcal{D}_t)$ 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 $A_t = \arg\max_a \mathbb{E}[R(a) \mid Q_t] = \arg\max_a Q_t(a)$.

- First we sample the $Q_t = (Q_t(1), \ldots, Q_t(k))$ from the posterior.
- Then we choose the action that has largest expected reward under Q_t.
- For our setting, this is exactly $\arg\max_a Q_t(a)$.
- What is the distribution of A_t chosen in this way? [want: $\mathbb{P}(A_t = a) = p_a$]

Thompson sampling trick does the right thing

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 $A_t = \arg\max_a Q_t(a)$.

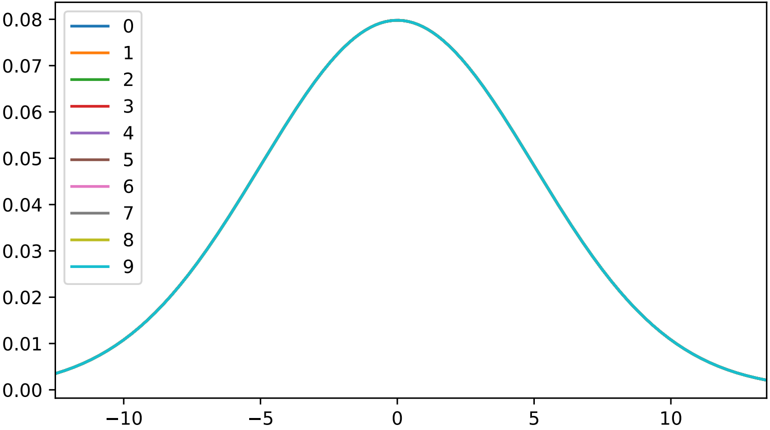
- Note that

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

- So A_t has exactly the desired distribution.

Experimental results

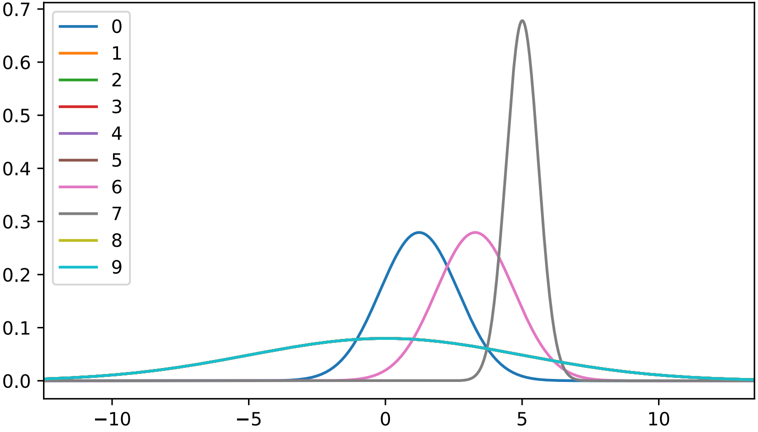
Prior distributions on reward means



Plot and simulation code courtesy of Ryan Carroll.

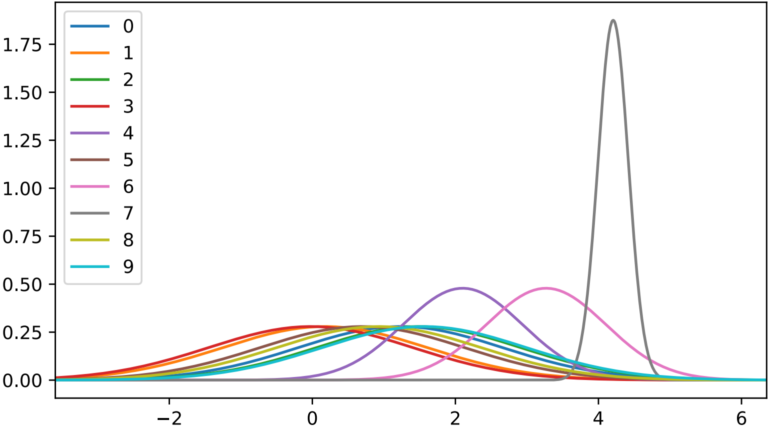
- Recall that the reward distribution for action a is given by $\mathcal{N}(Q(a), \sigma = 2)$.
- What we're showing here is the prior distribution on $Q(a)$, which we've taken to be $\mathcal{N}(0, \sigma = 5)$ for each a.

Posterior reward means $n = 5$



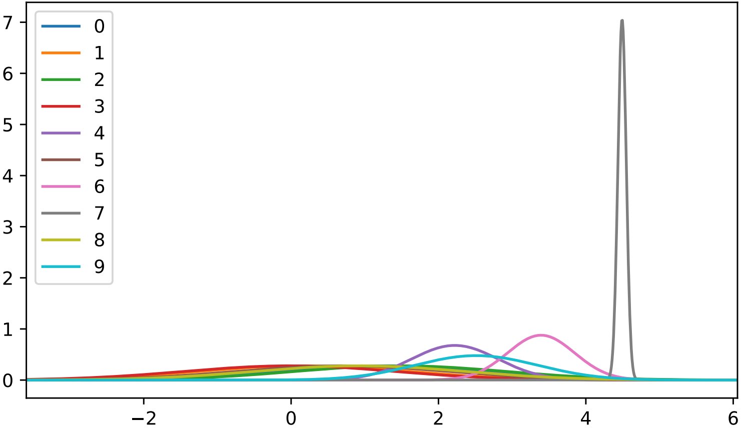
Plot and simulation code courtesy of Ryan Carroll.

Posterior reward means $n = 20$



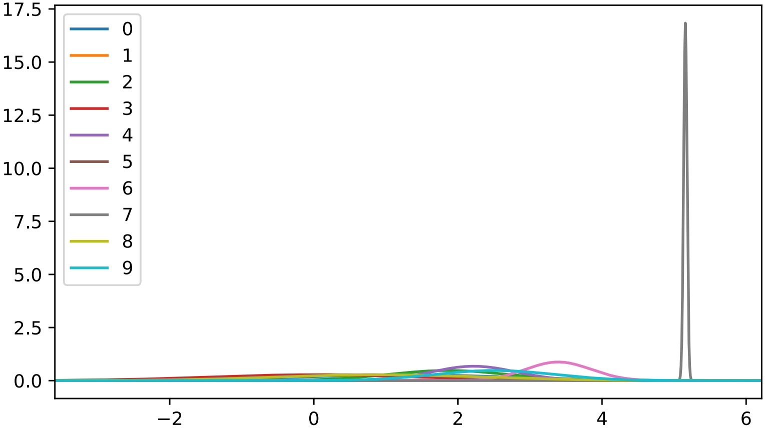
Plot and simulation code courtesy of Ryan Carroll.

Posterior reward means $n = 50$



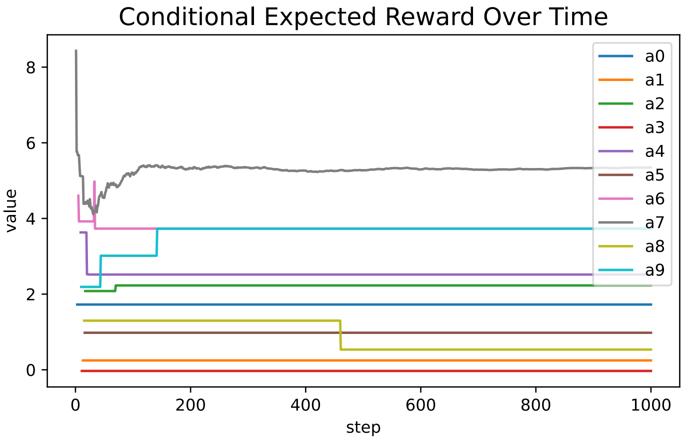
Plot and simulation code courtesy of Ryan Carroll.

Posterior reward means $n = 100$



Plot and simulation code courtesy of Ryan Carroll.

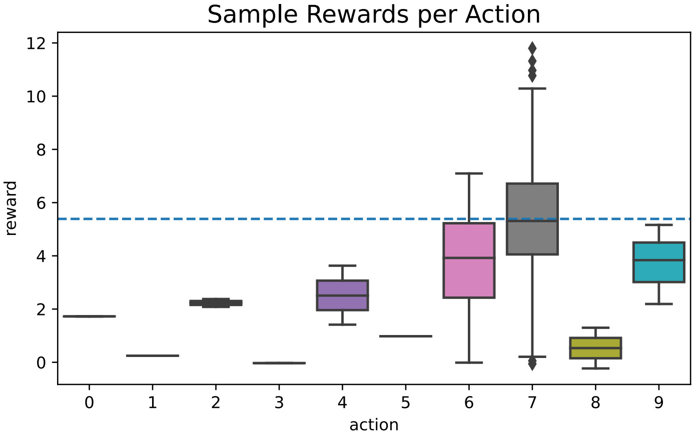
Posterior expected reward



Plot and simulation code courtesy of Ryan Carroll.

- Here we're plotting $\mathbb{E}[Q(a) \mid \mathcal{D}_t]$ for each action $a \in \{0, \ldots, 9\}$ over 1000 rounds.
- You can see that the expected rewards for most actions don't change much, since their performance doesn't look good, and thus they're not played often.

Received rewards by action



Plot and simulation code courtesy of Ryan Carroll.

- Here we're seeing a representation of the distribution of observed rewards for each 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

- Following the same protocol as in [SB18, Sec 2.3] and in the previous module, we repeat 1000-round bandit the 2000 drawn bandit distributions. Here we display statistics on the average rewards received for different prior distributions.

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.

[SB18] Richard S. Sutton and Andrew G. Barto, *Reinforcement learning: An introduction*, A Bradford Book, Cambridge, MA, USA, 2018.

[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.