Thompson Sampling

David S. Rosenberg

NYU: CDS

March 3, 2021

Contents

Bayesian updating for Gaussians

Review: Bayesian updating for Gaussians

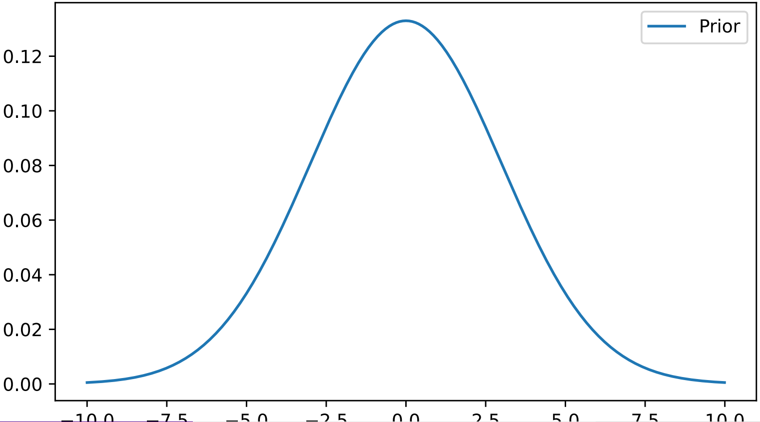
- Consider $R \sim \mathcal{N}\left(\mu, \sigma^2\right)$.
- Suppose we know σ^2, but don't know μ.
- We'll take a Bayesian approach.
- Put prior on μ: $p(\mu) = \mathcal{N}\left(\mu; 0, \sigma_0^2\right)$.
- Get data $R_1, \ldots, R_{t-1}$ i.i.d. $\mathcal{N}\left(\mu, \sigma^2\right)$.
- Posterior on μ: $p(\mu \mid R_1, \ldots, R_{t-1}) = \mathcal{N}\left(\mu; \mu_t, \sigma_t^2\right)$, where

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

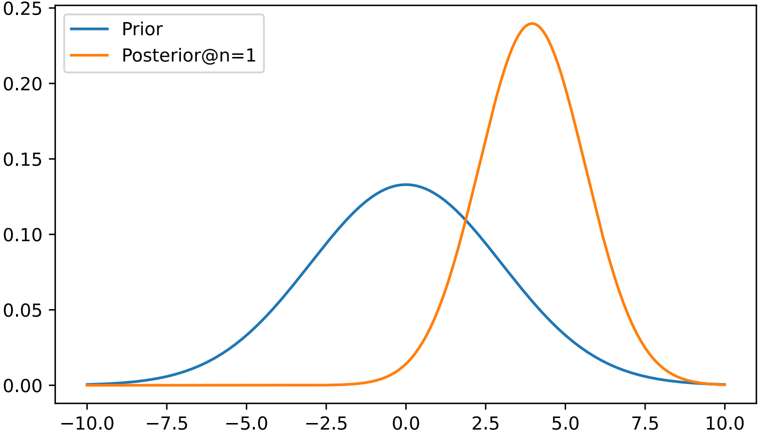
- Posterior mean μ_t is a weighted average of prior mean μ_0 and observed mean.

Gaussian prior distibution

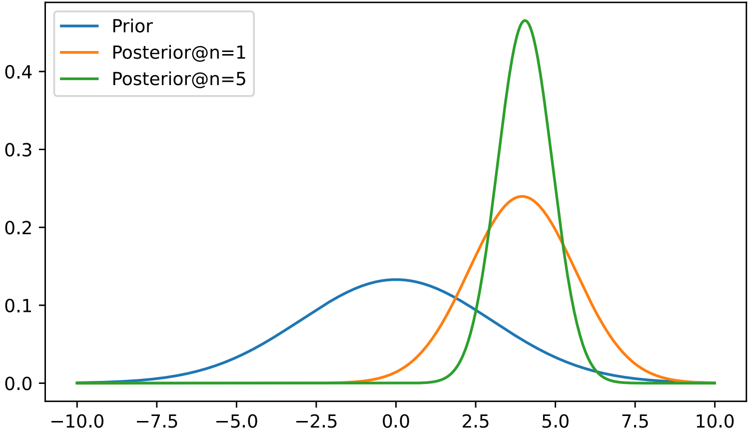
- Consider sampling from $R_1, R_2, \ldots \sim \mathcal{N}(5, \sigma = 2)$.
- Use prior $\mathcal{N}(0, \sigma = 3)$.



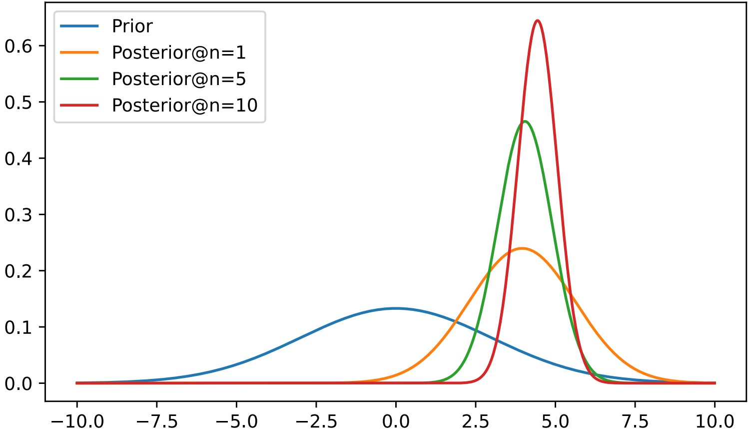
Posterior after $n = 1$ observations



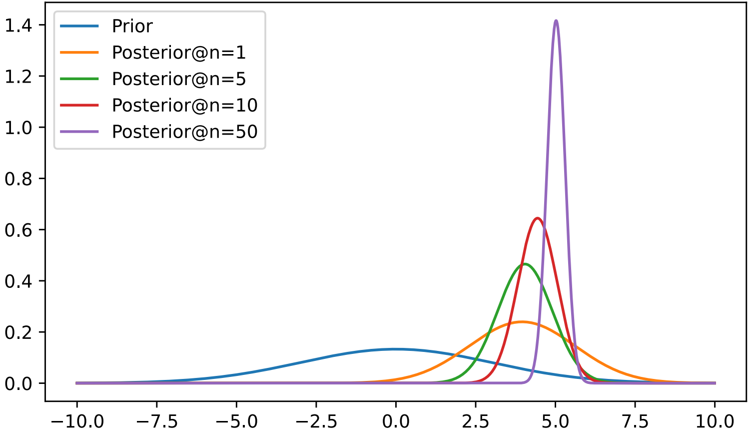
Posterior after $n = 5$ observations



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?

Gaussian priors

- For simplicity, we'll assume reward distribution is
$$\mathcal{N}(q_*(a), \sigma = 2),$$
for each action.
- The only thing we don't know is the expected reward $q_*(a)$.
- Let's put a $\mathcal{N}(0, \sigma = 5)$ prior on $q_*(a)$ for each action a.
- Let's write the posterior on $q_*(a)$ at **start** of round t as
$$\mathcal{N}(q_t(a), \sigma_t(a)),$$
where $q_t(a)$ and $\sigma_t(a)$ are updated based on
$$\mathcal{D}_t = ((A_1, R_1(A_1)), \ldots, (A_{t-1}, R_{t-1}(A_{t-1}))).$$

Action choice

.

- Ideally we'd choose action a with largest $q_*(a) = \mathbb{E}[R(a)]$.
- We only have a posterior on $q_*(a)$ for each a.
- We could choose a with maximum posterior mean $q_t(a)$.
- That would **pure exploitation.**

Thompson sampling action choice

Thompson sampling action choice

Sample action a with probability that a has the largest expected reward $q_*(a)$ (under our posterior).

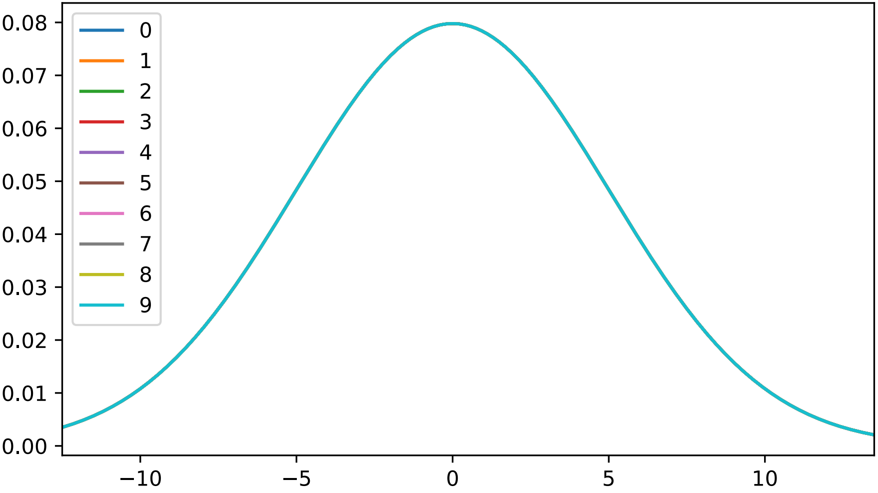
- The more certain we are that a is the best, the more likely we are to select a.
- Thompson sampling amounts to a **heuristic** strategy.
- It's an approach to the explore/exploit tradeoff.
- How to sample from this particular distribution?

Thompson sampling recipe

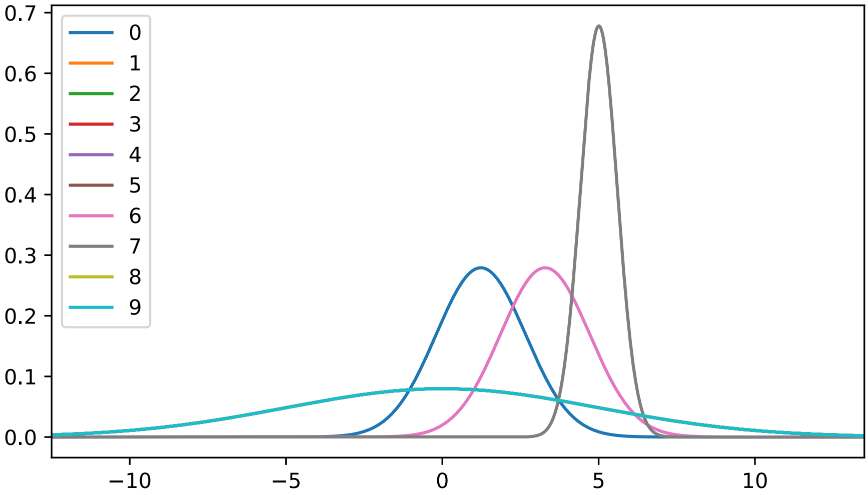
- For each action a
  - sample synthetic reward R_a from the posterior over $q_*(a)$.
- Choose action A corresponding to $\arg\max_a R_a$.
- Turns out, A has the desired distribution.
- That is, $A = a$ with probability that a has the largest expected reward, under our posterior.

Experimental results

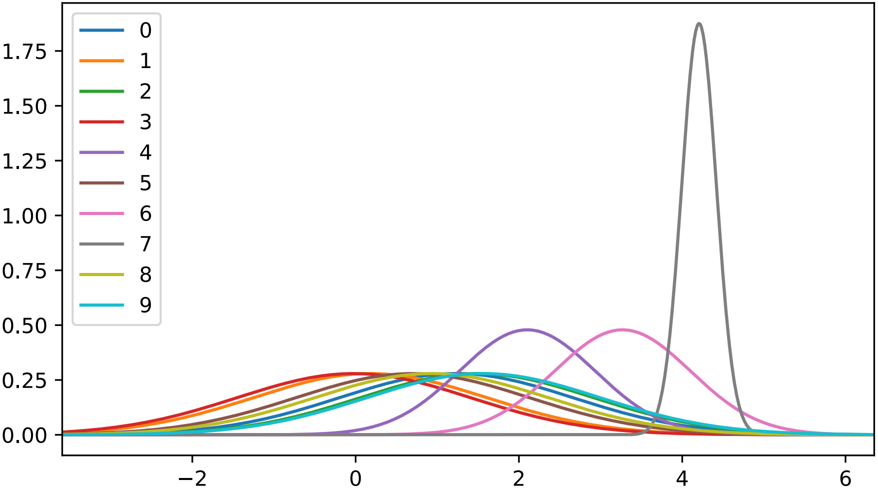
Prior distributions



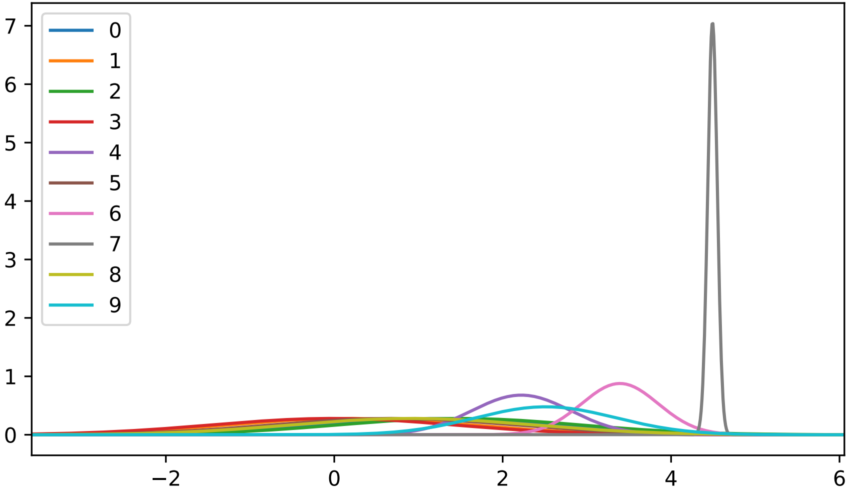
Posterior distributions $n = 5$



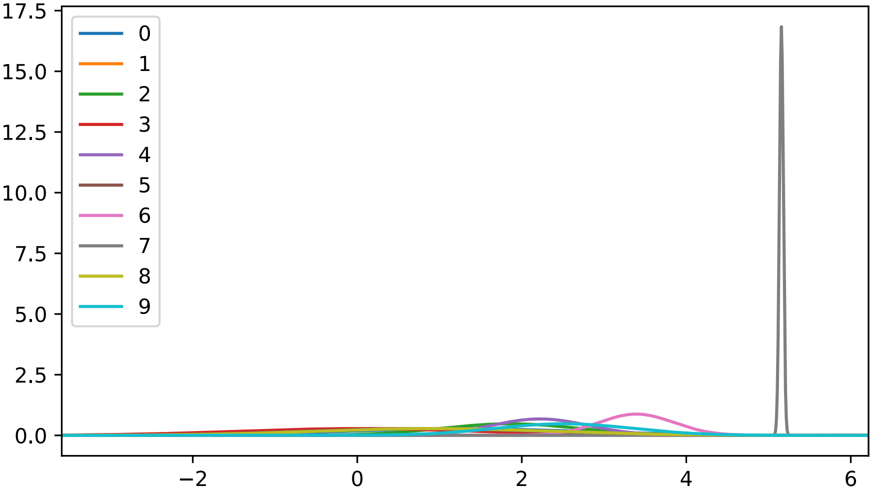
Posterior distribution $n = 20$



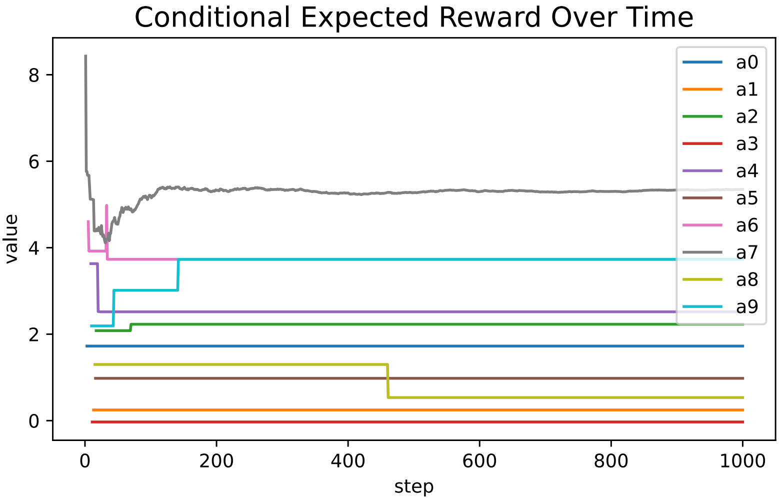
Posterior distribution $n = 50$



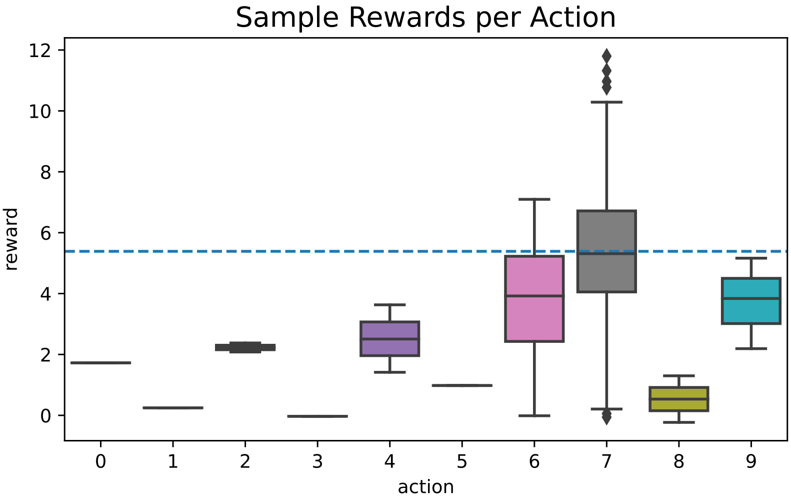
Posterior distribution $n = 100$



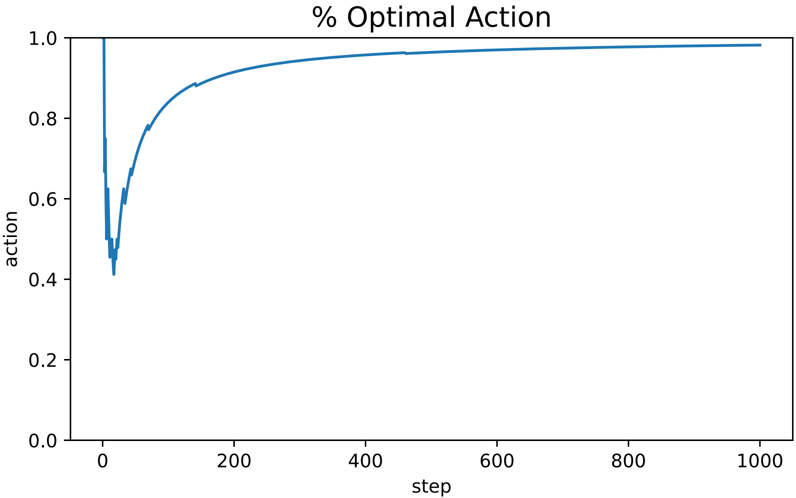
Posterior expected reward



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.