DATA 37200: Learning, Decisions, and Limits
(Winter 2026)

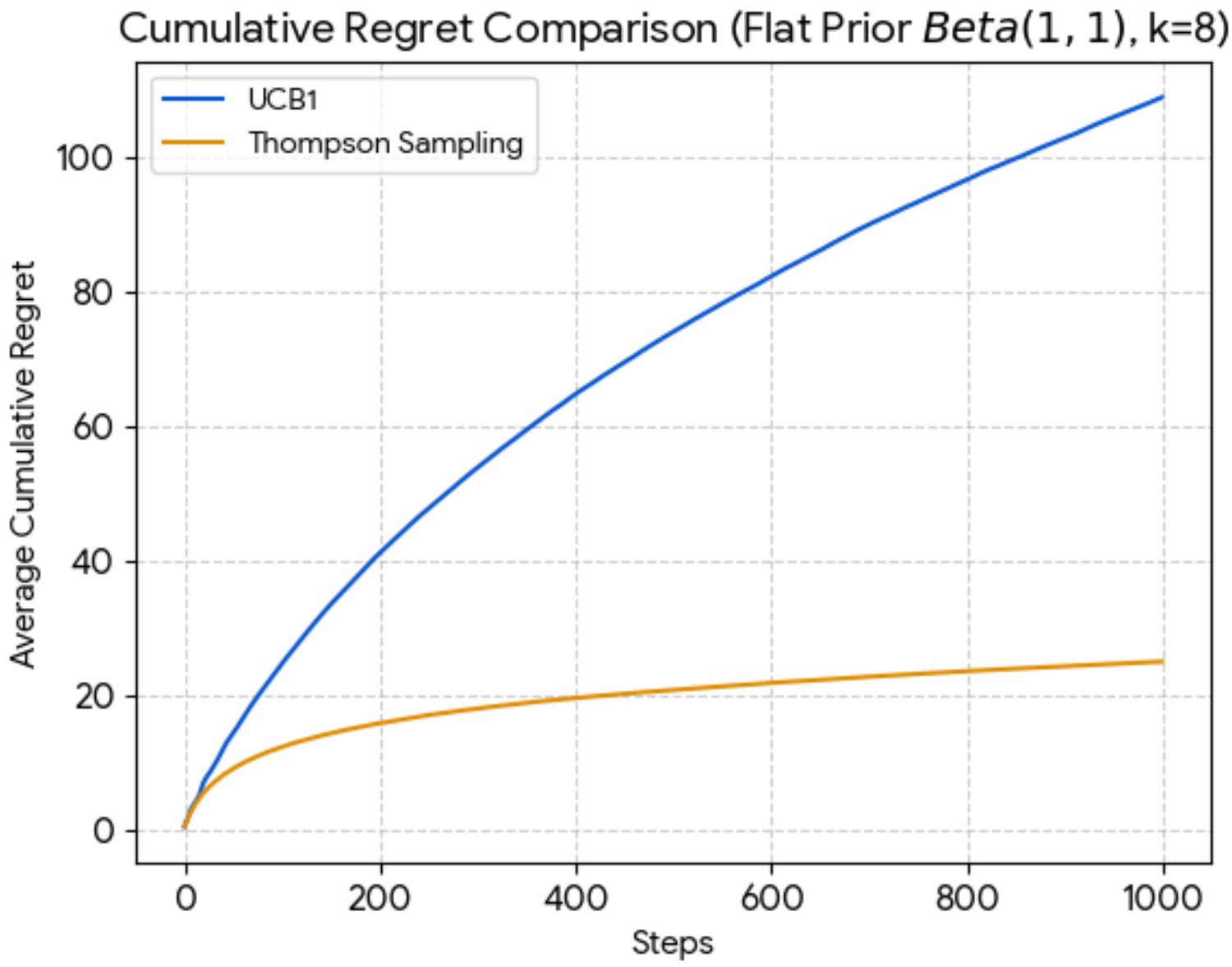
Lecture 5: Thompson sampling

Instructor: Frederic Koehler

Reference

Up to some minor details, we will follow Tor Lattimore and Csaba Szepesvari's Bandits book, Chapters 35-36.

A simulation



Beyond UCB1: The Bayesian Perspective

- **Recall UCB1:** Uses an Upper Confidence Bound (Optimism in the face of uncertainty).
- **Bayesian Setting:** We treat the unknown distribution parameters as random variables.
- **For simplicity:** today, we focus on a simple Bayesian setting where arm rewards are 0/1 valued. (But generalization is easy.)
- **Notation:**
  - k arms with distributions $\mathcal{D}_1, \ldots, \mathcal{D}_k$.
  - Bernoulli rewards: $r(t) \sim \text{Bernoulli}(\mu_j)$ where μ_j is unknown.
  - Arm chosen at time t: $i(t) \in \{1, \ldots, k\}$.
  - Reward at time t: $r(t) \in \{0, 1\}$.

Frequentist vs Bayesian statistics

- *Frequentist statistics:* the true means $\mu_1, \ldots, \mu_k$ are fixed and unknown. Randomness in our experiment is only due to randomness of our rewards (and policy, if randomized).
  - To measure uncertainty, build "confidence intervals": e.g. a random interval $[L_i, U_i]$ such that with 95% probability over the randomness of the interval, $\mu_i \in [L_i, U_i]$.
- *Bayesian statistics:* additionally models $\mu_1, \ldots, \mu_k$ as random variables sampled from a "prior" distribution.
  - Using Bayes rule, we can update the prior to a posterior distribution after seeing the data. For observation X and unknown parameter θ,

$$P(\theta|X) \propto P(X|\theta)P(\theta)$$

    (posterior $\propto$ likelihood $\times$ prior).
  - Posterior $=$ our beliefs, including our uncertainty.
  - Sample from prior to understand possible θ given the data.

The Prior: Beta Distribution

As a Bayesian, how to pick a prior distribution? In real life: often determined by computational considerations.

For the Bernoulli distribution, the **Beta Distribution** is the "conjugate prior" (so it will be very easy to calculate with).

Probability Density Function (PDF)

For $x \in [0, 1]$, and parameters $\alpha, \beta > 0$:

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

- $B(\alpha, \beta)$ is an (explicit) normalizing constant so $\int f dx = 1$.
- $\alpha - 1$: "Pseudo-counts" of successes (1s).
- $\beta - 1$: "Pseudo-counts" of failures (0s).
- **Flat Prior:** $Beta(1, 1)$ is $Uniform(0, 1)$.
- **Jeffreys Prior:** $Beta(0.5, 0.5)$ (non-informative for Bernoulli).

Key properties of Beta

Mean of $Z \sim \mathit{Beta}(\alpha, \beta)$:

$$\mathbb{E}[Z] = \frac{\alpha}{\alpha + \beta}.$$

Conjugacy with Bernoulli: suppose $Z \sim \mathit{Beta}(\alpha, \beta)$ and $X \sim \mathit{Ber}(Z)$. Then

$$Z \mid X \sim \mathit{Beta}(\alpha + X, \beta + (1 - X)).$$

(Similarly extends to $\mathit{Bin}(n, Z)$.) Why? Using Bayes rule

$$p(Z = z \mid X) \propto p(X \mid Z = z)p(Z = z) \propto z^x(1-z)^{1-x}z^{\alpha-1}(1-z)^{\beta-1}$$

which is exactly the density of $\mathit{Beta}(\alpha + X, \beta + (1 - X))$.

Posterior Updates

In the Bayesian setting, we maintain a belief (posterior) for each arm j, and update it based on our experiences.

The Update Rule: If we choose arm $i(t)$ and observe reward $r(t)$:

1. If $r(t) = 1$: $\alpha_{i(t)} \leftarrow \alpha_{i(t)} + 1$
2. If $r(t) = 0$: $\beta_{i(t)} \leftarrow \beta_{i(t)} + 1$

The posterior distribution for μ_j at time t is:

$$\pi_{j,t} = \mathit{Beta}(\alpha_{j,0} + S_{j,t}, \beta_{j,0} + F_{j,t})$$

where $S_{j,t}$ and $F_{j,t}$ are the number of successes and failures for arm j up to time t.

Thompson Sampling (TS)

Thompson Sampling (or Posterior Sampling) implements **Probability Matching**.

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

1. **Sample:** For each arm $j \in \{1, \ldots, k\}$, sample:

$$\hat{\mu}_j(t) \sim \pi_{j,t}$$

2. **Act:** Choose arm $i(t) = \text{argmax}_j \hat{\mu}_j(t)$.
3. **Observe:** Get reward $r(t) \sim \text{Bernoulli}(\mu_{i(t)})$.
4. **Update:** Update $\pi_{i(t),t+1}$ using $r(t)$. (I.e., increment either $\alpha_{i(t)}$ or $\beta_{i(t)}$ appropriately.)

Bayesian Regret Definition

Unlike frequentist regret (which is specific to fixed parameters θ), Bayesian Regret (BR) averages over the prior ν.

$$BR(T) = \mathbb{E}_{\theta \sim \nu}\left[\mathbb{E}\left[\sum_{t=1}^{T}(\mu^* - \mu_{i(t)})\right]\right]$$

- $\mu^* = \max_j \mu_j$.
- The inner expectation is over the randomness of rewards and the algorithm.
- The outer expectation is over the prior distributions $\pi_{1,0} \dots \pi_{k,0}$.

Regret Guarantees (Lattimore & Szepesvári)

For k-armed bandits with rewards in $[0, 1]$, Thompson Sampling satisfies a Bayesian regret bound:

Theorem (Bayesian Regret of TS)

For any prior ν, the Bayesian regret of Thompson Sampling after T rounds is bounded by:

$$BR(T) \leq \sqrt{\frac{1}{2} kT \log T}$$

Why is this important?

- It matches the lower bound of $O(\sqrt{kT})$ up to a logarithmic factor.
- Despite having a similar guarantee, we saw experimentally that TS can outperform UCB1. Nice to know that this does not require throwing away mathematical guarantees.
- Surprising trick to proof: analysis argues that Thompson sampling is "optimistic" similar to UCB1.

The Core of the Proof: Probability Matching

Let $\mathcal{F}_{t-1}$ be the history (filtration) up to time $t-1$. Let i^* be the index of the true optimal arm, $i^* = \text{argmax}_j \mu_j$.

The Probability Matching Property

Under Thompson Sampling, the conditional distribution of the chosen arm $i(t)$ is the same as the conditional distribution of the optimal arm i^*:

$$\mathbb{P}(i(t) = j \mid \mathcal{F}_{t-1}) = \mathbb{P}(i^* = j \mid \mathcal{F}_{t-1})$$

(Why? symmetry/Nishimori identity)
By the same symmetry principle, we have

$$\mathbb{E}[\mu_{i^*} \mid \mathcal{F}_{t-1}] = \mathbb{E}[\hat{\mu}_{i(t)} \mid \mathcal{F}_{t-1}]$$

where $\hat{\mu}_{i(t)}$ is the **sample** drawn by the algorithm.

Aside: probability matching in psychology

"In this situation S is asked to predict on each of a series of trials whether some designated event, e.g., the flash of a light, will occur; this event, the analogue of the US (Unconditioned Stimulus) in a conditioning experiment, is presented in accordance with a predetermined schedule, usually random with some fixed probability. Several recent investigators (3, 5) have noted that S tends to match his response rate to the rate of occurrence of the predicted event so that if the probability of the latter is, say, .75, the mean response curve for a group of Ss tends over a series of trials toward an apparently stable final level at which the event is predicted on approximately 75% of the trials. This behavior has seemed puzzling to most investigators since it does not maximize the proportion of successful predictions..."

(Estes-Straughan, "Analysis of a verbal conditioning situation in terms of statistical learning theory", Journal of Experimental Psychology '54)

Step 1: Regret Decomposition

We can decompose the instantaneous Bayesian regret $r(t)$ as follows:

$$\mathbb{E}[r(t) \mid \mathcal{F}_{t-1}] = \mathbb{E}[\mu_{i^*} - \mu_{i(t)} \mid \mathcal{F}_{t-1}]$$

Using the Probability Matching property:

$$\mathbb{E}[r(t) \mid \mathcal{F}_{t-1}] = \mathbb{E}[\underbrace{\hat{\mu}_{i(t)} - \mu_{i(t)}}_{\text{Sampled vs. True Mean}} \mid \mathcal{F}_{t-1}]$$

Insight: The regret is determined by how much our *sample* ($\hat{\mu}_{i(t)}$) deviates from the *true mean* ($\mu_{i(t)}$) of the arm we actually pulled.

Step 2: Introducing a Confidence Bound

To bound $\mathbb{E}[\hat{\mu}_{i(t)} - \mu_{i(t)} \mid \mathcal{F}_{t-1}]$, we introduce an Upper Confidence Bound (UCB) $U_j(t)$ for each arm j.

We split the term into two parts:

$$\hat{\mu}_{i(t)} - \mu_{i(t)} = (\hat{\mu}_{i(t)} - U_{i(t)}(t)) + (U_{i(t)}(t) - \mu_{i(t)})$$

1. **Part A:** $\mathbb{E}[U_{i(t)}(t) - \mu_{i(t)}]$. How much the UCB estimate is overoptimistic. This was bounded inside of the UCB1 proof (if UCB were very overoptimistic, it would not have had a good regret bound).
2. **Part B:** $\mathbb{E}[\hat{\mu}_{i(t)} - U_{i(t)}(t)]$. This measures how much the Thompson sample exceeds our "safe" upper bound. (Think: positive only in rare events where we are "surprised".)

Step 3: Concentration and Summation

By choosing $U_j(t) = \bar{\theta}_j(t-1) + \sqrt{\frac{2 \log T}{N_j(t-1)}}$:

- The term $\sum_t \mathbb{E}[U_{i(t)}(t) - \mu_{i(t)}]$ scales as $O(\sqrt{kT \log T})$ using the UCB1 analysis.
- Similarly, we can consider (the positive contribution to)

$$\sum_t \mathbb{E}[\hat{\mu}_{i(t)} - U_{i(t)}(t)]$$

and this is also bounded by a similar factor (corresponds to the "bad events" in UCB1 analysis).

Final Result

We have the in-expectation bound

$$BR(T) = \sum_{t=1}^{T} \mathbb{E}[r(t)] \lesssim \sqrt{kT \log T}.$$

High probability bound by Markov.