Sequential Decision

Session 3: Thompson sampling and linear bandits

Igor Colin

October 2026

ESIEE Paris

Session 2

Quiz

1. At round $t = 100$, an arm has been pulled 4 times and its empirical mean is 0.5. Compute its UCB1 index, with $\log 100 \approx 4.6$.

2. Under UCB1, a suboptimal arm with gap Δ is pulled of the order of $\log(T)/\Delta^2$ times. Give the factor by which this number is multiplied when the gap is halved.

3. On the instance $(0.6, 0.4)$, UCB(0.5) has a smaller mean regret than UCB1. Name the risk taken with a small exploration parameter.

Where we are

Strategy	Needs to know	Regret
ε-greedy, fixed ε	nothing	linear
ETC, $m = m^\star$	T and Δ	$O(\log(T)/\Delta)$
UCB1	nothing	$O\big(\sum_a \log(T)/\Delta_a\big)$

Two limits of UCB1

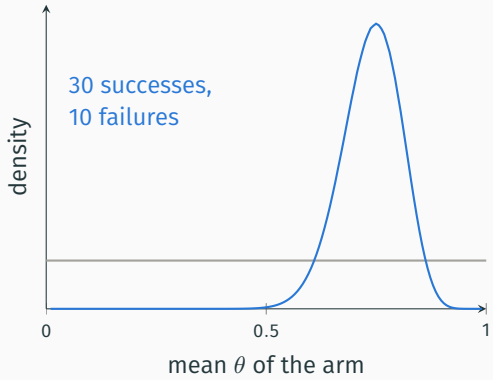
- Its constant is far from the Lai-Robbins lower bound
- Arms are unrelated: pulling one arm says nothing about the others

This session

- Thompson sampling: a randomized strategy that reaches the lower bound
- Linear bandits: arms described by features, so that information is shared between arms and between users

Thompson sampling

The Bayesian approach



The unknown mean θ of an arm is treated as a random variable.

- **Prior**: what we believe before any pull
- **Posterior**: the distribution of θ given the rewards observed so far (Bayes' rule)
- The posterior concentrates around the empirical mean, with a width of order $1/\sqrt{n}$: it plays the role of the confidence interval

Beta posteriors

Bernoulli rewards (tutorial, exercise 1)

If the prior on θ is $\mathrm{Beta}(a, b)$, then after S successes and F failures the posterior is

$$\mathrm{Beta}(a + S,\ b + F) \ .$$

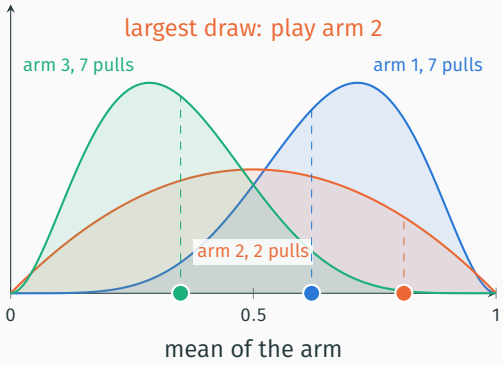
The Beta distribution

- Density proportional to $\theta^{a-1}(1-\theta)^{b-1}$ on $[0, 1]$
- $\mathrm{Beta}(1, 1)$ is the uniform distribution
- Mean $\dfrac{a}{a+b}$

Updating is counting

- Reward 1: add one to the first parameter
- Reward 0: add one to the second parameter
- With a uniform prior, the posterior mean is $\dfrac{1+S}{2+S+F}$

Thompson sampling



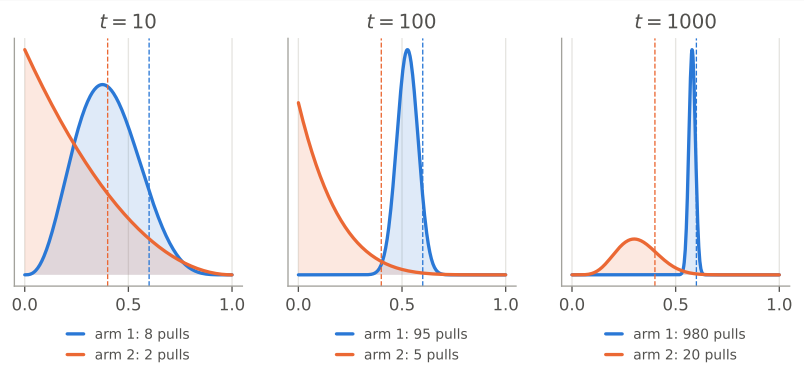
Strategy (Thompson, 1933)

At each round:

1. draw $\tilde{\theta}_a$ from the posterior of each arm a;
2. play $A_t \in \arg\max_a \tilde{\theta}_a$;
3. update the posterior of A_t.

- Most draws favor the arm with the best posterior
- A wide posterior sometimes gives the largest draw: this is the exploration

Posteriors along a run



- Each arm is played with the posterior probability that it is the best one
- Arm 2 is abandoned once the two posteriors no longer overlap, after only 20 pulls out of 1000: its mean is still poorly known, but known to be smaller

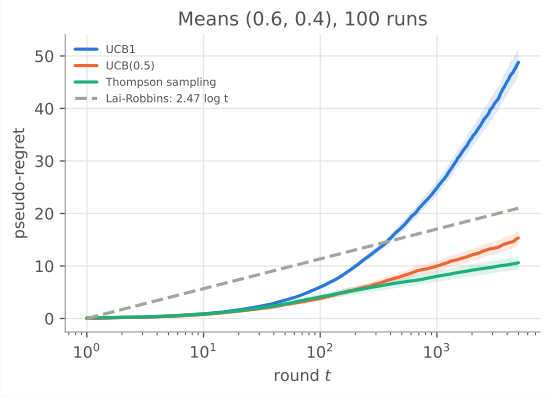
Regret of Thompson sampling

Theorem (Kaufmann et al., 2012)

For Bernoulli arms and uniform priors,

$$\lim_{T\to\infty} \frac{R_T}{\log T} = \sum_{a\,:\,\Delta_a>0} \frac{\Delta_a}{\mathrm{kl}(\mu_a, \mu^\star)} \quad .$$

- The Lai-Robbins constant: it cannot be improved
- Asymptotic: curves can be below the dashed line at a finite horizon
- At $T = 5000$: regret ≈ 11, against 49 for UCB1



UCB and Thompson sampling

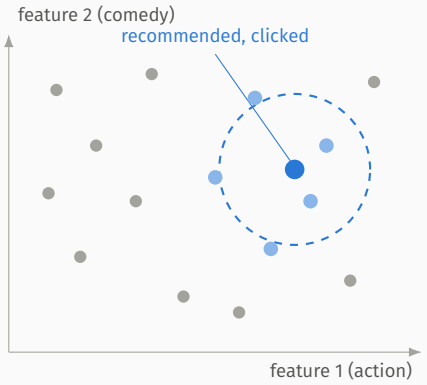
	UCB	Thompson sampling
Uncertainty	confidence interval	posterior distribution
Choice	argmax of the upper bound	argmax of a random draw
Decision given the history	deterministic	randomized
Tuning	exploration parameter α	prior
Delayed or batched feedback	repeats the same arm	spreads its choices

Sampling and argmax

Taking the argmax of an estimate (greedy) locks; adding a bonus (UCB) or drawing from a distribution (Thompson sampling) explores. The same choice comes back in language models: greedy decoding takes the most likely token, sampling draws from the distribution.

Linear bandits

Arms with features



A catalog of 10 000 movies, each described by a few features.

- With unrelated arms, every movie must be recommended many times: the regret of UCB1 grows with K
- One click is observed for one movie
- If similar movies have similar click rates, this click informs about its neighbors
- Learning d coefficients instead of K means

The linear model

Linear bandit

K arms with known feature vectors $x_1, \dots, x_K \in \mathbb{R}^d$ and an unknown parameter $\theta^\star \in \mathbb{R}^d$. At each round t, the agent chooses A_t and observes X_t with

$$\mathbb{E}[X_t \mid \text{history}] = \langle \theta^\star, x_{A_t} \rangle \ .$$

K-armed bandit

- $d = K$, $x_a = e_a$
- $\theta^\star = (\mu_1, \dots, \mu_K)$

Catalog

- x_a: features of item a, with $d \ll K$
- $\theta^\star$: tastes of the audience

Regret: as before, with $\mu_a = \langle \theta^\star, x_a \rangle$,

$$R_T = T \max_a \langle \theta^\star, x_a \rangle - \mathbb{E}\Big[\sum_{t=1}^{T} \langle \theta^\star, x_{A_t} \rangle \Big] \ .$$

Ridge regression

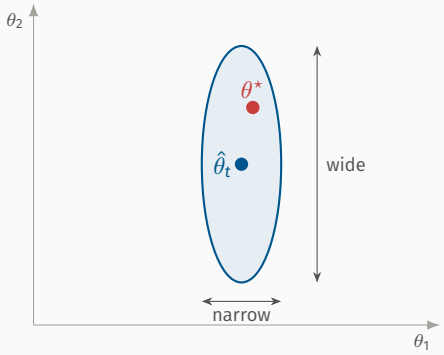
After t rounds, with $x_s = x_{A_s}$ the feature vector played at round s:

Estimator (tutorial, exercise 2)

$$\hat{\theta}_t = \mathbf{V}_t^{-1} b_t \ , \qquad \mathbf{V}_t = \lambda \mathbf{I} + \sum_{s=1}^{t} x_s x_s^\top \ , \qquad b_t = \sum_{s=1}^{t} X_s \, x_s \ .$$

- $\hat{\theta}_t$ minimizes $\sum_s (X_s - \langle \theta, x_s \rangle)^2 + \lambda \|\theta\|^2$: least squares with a penalty, which makes $\mathbf{V}_t$ invertible
- $\mathbf{V}_t$ counts how much each direction has been explored: for $x_a = e_a$ it is the diagonal matrix of the $\lambda + N_a(t)$
- Incremental update in $O(d^2)$ with the Sherman-Morrison formula: no linear system to solve at each round

The confidence ellipsoid

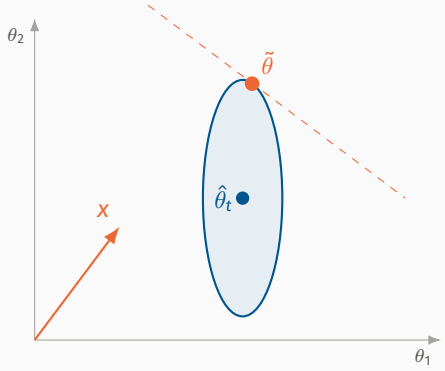


With high probability, $\theta^\star$ belongs to

$$\mathcal{C}_t = \left\{ \theta : (\theta - \hat{\theta}_t)^\top \mathbf{V}_t (\theta - \hat{\theta}_t) \leq \beta^2 \right\} .$$

- An ellipsoid centered at $\hat{\theta}_t$: the confidence interval of dimension d
- Here the pulls were mostly along the first coordinate: θ_1 is well known, θ_2 is not
- Each pull shrinks the ellipsoid in the direction of the feature vector that was played

LinUCB



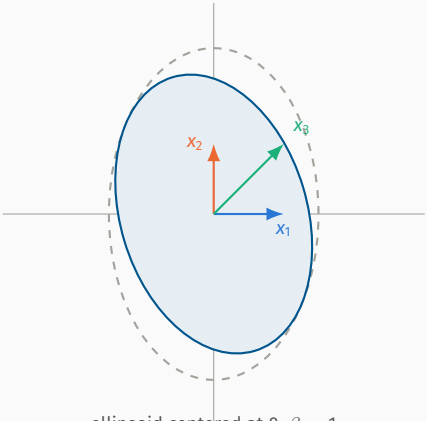
Optimism: evaluate each arm with the most favorable parameter of the ellipsoid.

Index of the arm with features x

$$U_t(x) = \max_{\theta \in \mathcal{C}_t} \langle \theta, x \rangle = \langle \hat{\theta}_t, x \rangle + \beta \sqrt{x^\top \mathbf{V}_t^{-1} x}$$

- Play $A_t \in \arg\max_a U_t(x_a)$
- The bonus is the half-width of the ellipsoid in the direction x: large for directions that were rarely played
- For $x_a = e_a$: bonus $\beta / \sqrt{\lambda + N_a}$, as in UCB

Sharing information between arms



ellipsoid centered at 0, $\beta = 1$

Tutorial, exercise 3: $x_1 = (1, 0)$ pulled 4 times, $x_2 = (0, 1)$ once, $x_3 = (1, 1)$ never.

- Bonuses: 0.45, 0.71 and 0.84
- After one pull of x_3: the ellipsoid shrinks along x_3, and tilts
- The bonus of x_3 drops to 0.64, and the one of x_1 to 0.42, although x_1 was not pulled
- With unrelated arms, only the bonus of the pulled arm changes

Regret of LinUCB

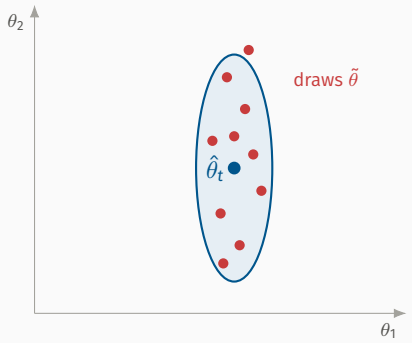
Theorem (Abbasi-Yadkori, Pál and Szepesvári, 2011)

For bounded features and parameter, and $\beta = \beta_t$ of order $\sqrt{d \log t}$, LinUCB satisfies, up to logarithmic factors,

$$R_T = O\big(d\sqrt{T}\big) \ .$$

- The bound does not depend on the number of arms K, only on the dimension d
- The pulls depend on the past rewards: $\hat{\theta}_t$ is not a regression on a fixed design, and the confidence ellipsoid requires a concentration inequality for martingales
- The rate is $\sqrt{T}$ and not $\log T$: no gap is assumed between the arms
- In practice β is a constant to tune, as α for UCB (lab)

Linear Thompson sampling



Strategy

At each round, draw $\tilde{\theta} \sim \mathcal{N}\big(\hat{\theta}_t,\, v^2 \mathbf{V}_t^{-1}\big)$ and play $A_t \in \arg\max_a \langle \tilde{\theta}, x_a \rangle$.

- The Gaussian has the shape of the ellipsoid: it is the posterior of $\theta^\star$ for a Gaussian prior and Gaussian noise
- One draw of $\tilde{\theta}$ instead of one bonus per arm: convenient when the set of arms is large or continuous
- Regret of order $d^{3/2}\sqrt{T}$ up to logarithmic factors (Agrawal and Goyal, 2013)

Limits of Thompson sampling

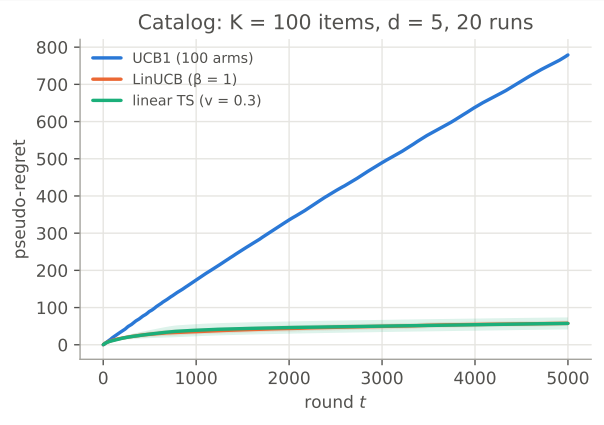
For K Bernoulli arms, Thompson sampling is as cheap and simple as UCB. Beyond:

- **Theory.** The guarantees of linear Thompson sampling require inflating the posterior variance by a factor of order $\sqrt{d}$, hence $d^{3/2}\sqrt{T}$ instead of $d\sqrt{T}$; without inflation, the regret can be linear on some instances (Hamidi and Bayati, 2020)
- **Computation.** A draw from $\mathcal{N}(\hat{\theta}_t, v^2\mathbf{V}_t^{-1})$ costs a Cholesky factorization, $O(d^3)$. For a Gaussian process (session 4), one draw is a whole function: a joint Gaussian over the n candidates, $O(n^3)$, where GP-UCB only needs the variance at each point
- **Models.** Without conjugacy (logistic model, neural network) the posterior has no closed form: Laplace approximation, MCMC, bootstrap or ensembles
- **Other objectives.** Identifying the best arm (session 4): Thompson sampling focuses on the best arm and samples the second one too little. Off-policy evaluation (session 5): the probability of each action has no closed form

Catalog of 100 items

$K = 100$ items with $d = 5$ features, $T = 5000$.

- UCB1 pulls each item about 50 times: its regret (≈ 780) is close to the one of uniform exploration
- LinUCB and linear Thompson sampling estimate 5 coefficients: regret ≈ 58



Summary

Strategy	Model	Regret
UCB1	K unrelated arms	$O\left(\sum_a \log(T)/\Delta_a\right)$
Thompson sampling	K Bernoulli arms	Lai-Robbins constant
LinUCB	linear, dimension d	$O(d\sqrt{T})$ up to logarithms
Linear Thompson sampling	linear, dimension d	$O(d^{3/2}\sqrt{T})$ up to logarithms

Tutorial 3

- Beta posteriors
- Ridge regression and its recursive update
- LinUCB on an example

Lab 3

- Thompson sampling against UCB
- Ridge, LinUCB, linear Thompson sampling
- Catalog of 100 items

Reading

- Lattimore and Szepesvári, *Bandit Algorithms*, chapters 19, 20 and 36
- Li et al., WWW 2010