Posterior Sampling Beyond Regret Minimization

Emilie Kaufmann



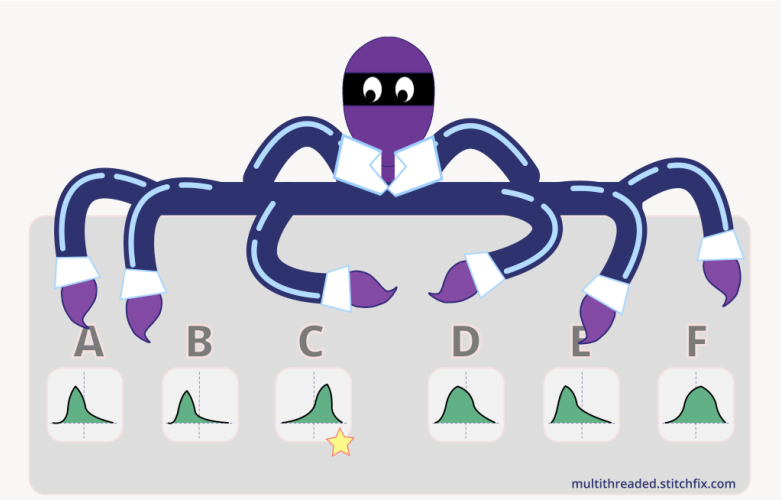




Statlearn, Grenoble, April 2026

The stochastic Multi Armed Bandit (MAB) model

- K *unknown* distributions $\nu_1, \ldots, \nu_K$ called *arms*
- a time t, select an arm A_t and collect an observation $X_t \sim \nu_{A_t}$



Sequential strategy / algorithm : A_{t+1} can depend on:

- previous observation $A_1, X_1, \ldots, A_t, X_t$
- some external randomization $U_t \sim \mathcal{U}([0,1])$
- some knowledge about the possible distributions: $\nu_a \in \mathcal{D}$

Outline

1. Posterior Sampling
2. Beyond Regret
3. Posterior Sampling for Pure Exploration
   - Top Two Thompson Sampling
   - Posterior Sampling Beyond Best Arm Identification
   - Murphy Sampling

Thompson Sampling

[Thompson, 1933] suggests the very first bandit algorithm.

A Bayesian algorithm:

- $K = 2$ Bernoulli arms, $\mathcal{B}(p_1), \mathcal{B}(p_2)$
- uniform prior distribution $(p_1, p_2) \sim \mathcal{U}([0,1]) \otimes \mathcal{U}([0,1])$

At time $t+1$, choose

- arm 1 with probability P_t
- arm 2 with probability $1 - P_t$

where P_t is the posterior probability that arm 1 is optimal:

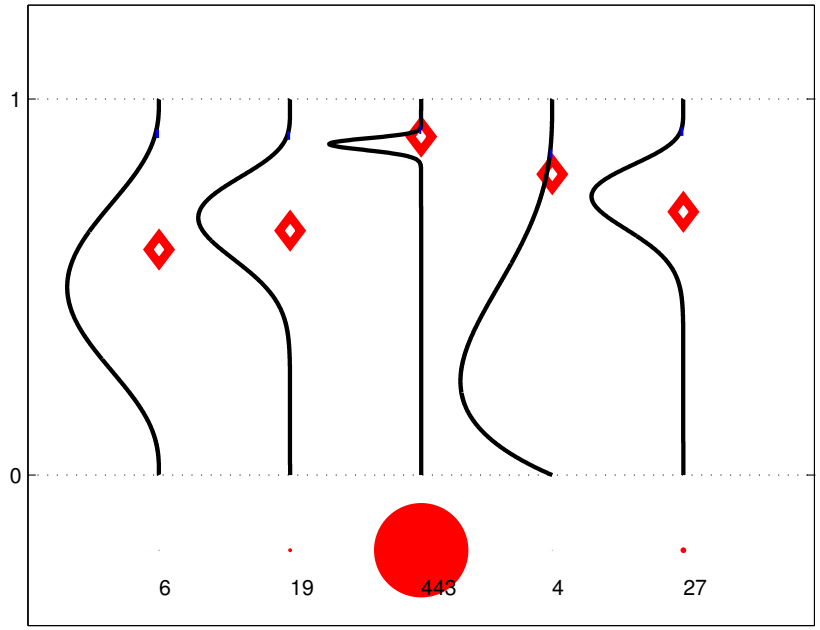
$$\begin{aligned} P_t &= \mathbb{P}\left(p_1 \geq p_2 | \mathcal{H}_t\right) \\ &= \mathbb{P}_{\substack{\widetilde{p}_1 \sim \mathrm{Beta}(S_1(t)+1, F_1(t)+1) \\ \widetilde{p}_2 \sim \mathrm{Beta}(S_2(t)+1, F_2(t)+1)}} \left(\widetilde{p}_1 \geq \widetilde{p}_2\right) \end{aligned}$$

$\mathcal{H}_t = (A_1, X_1, \ldots, A_t, X_t)$ observation collected in t time steps

$S_i(t)/F_i(t)$: number of successes / failures observed on arm i

Posterior Sampling: Modern View

Input: prior distribution over the arms' means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$
posterior distribution $\Pi_t = \mathcal{L}(\boldsymbol{\mu}|\mathcal{H}_t)$



In each round, TS is equivalent to sampling a bandit model from the posterior and selecting the best arm in the sampled model

$$\mathbb{P}(A_{t+1} = a|\mathcal{H}_t) = \Pi_t \left(a = \arg\max_i \mu_i \right)$$

[Russo et al., 2018]

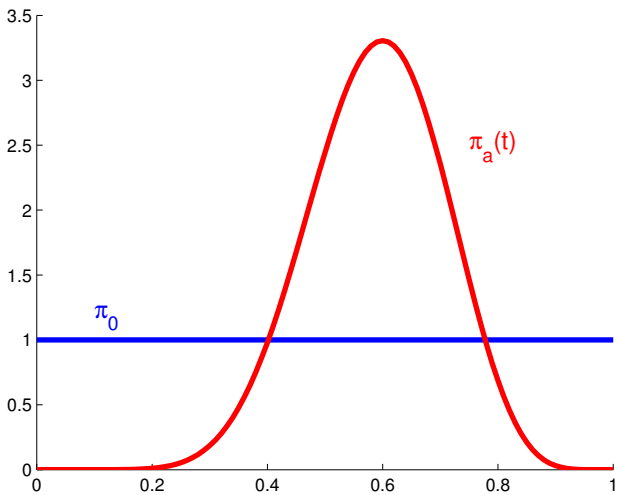
Posterior Sampling for Bernoulli bandits

Example: Bernoulli bandit with means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$

- prior distribution: $\mu_a \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0,1])$

➜ posterior distribution:

$$\begin{aligned} \pi_a(t) &= \mathcal{L}(\mu_a | X_1, \ldots, X_t) \\ &= \text{Beta}\Big(\underbrace{S_a(t)}_{\#ones} + 1, \underbrace{N_a(t) - S_a(t)}_{\#zeros} + 1 \Big) \end{aligned}$$



- $N_a(t) = \sum_{s=1}^{t} \mathbb{1}_{(A_s=a)}$ number of observations from arm a
- $S_a(t) = \sum_{s=1}^{t} X_s \mathbb{1}_{(A_s=a)}$ sum of the rewards from arm a

Posterior Sampling for Bernoulli bandits

Example: Bernoulli bandit with means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$

- prior distribution: $\mu_a \overset{\text{i.i.d.}}{\sim} \mathcal{U}([0,1])$

➜ posterior distribution:

$$\begin{aligned} \pi_a(t) &= \mathcal{L}\left(\mu_a | X_1, \ldots, X_t\right) \\ &= \text{Beta}\Big(\underbrace{S_a(t)}_{\#ones} + 1, \underbrace{N_a(t) - S_a(t)}_{\#zeros} + 1 \Big) \end{aligned}$$

Thompson Sampling

In round $t+1$:

$$\forall a \in [K], \quad \widetilde{\theta}_a(t) \sim \pi_a(t)$$

$$A_{t+1} = \underset{a \in [K]}{\arg\max}\ \widetilde{\theta}_a(t)$$

Posterior Sampling for Gaussian bandits

Example: Gaussian bandit with means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, var σ^2

- prior distribution: $\mu_a \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \kappa^2)$

➜ posterior distribution:

$$\begin{aligned} \pi_a(t) &= \mathcal{L}\left(\mu_a | X_1, \ldots, X_t\right) \\ &= \mathcal{N}\left(\frac{S_a(t)}{N_a(t) + \frac{\sigma^2}{\kappa^2}}, \frac{\sigma^2}{N_a(t) + \frac{\sigma^2}{\kappa^2}}\right) \end{aligned}$$

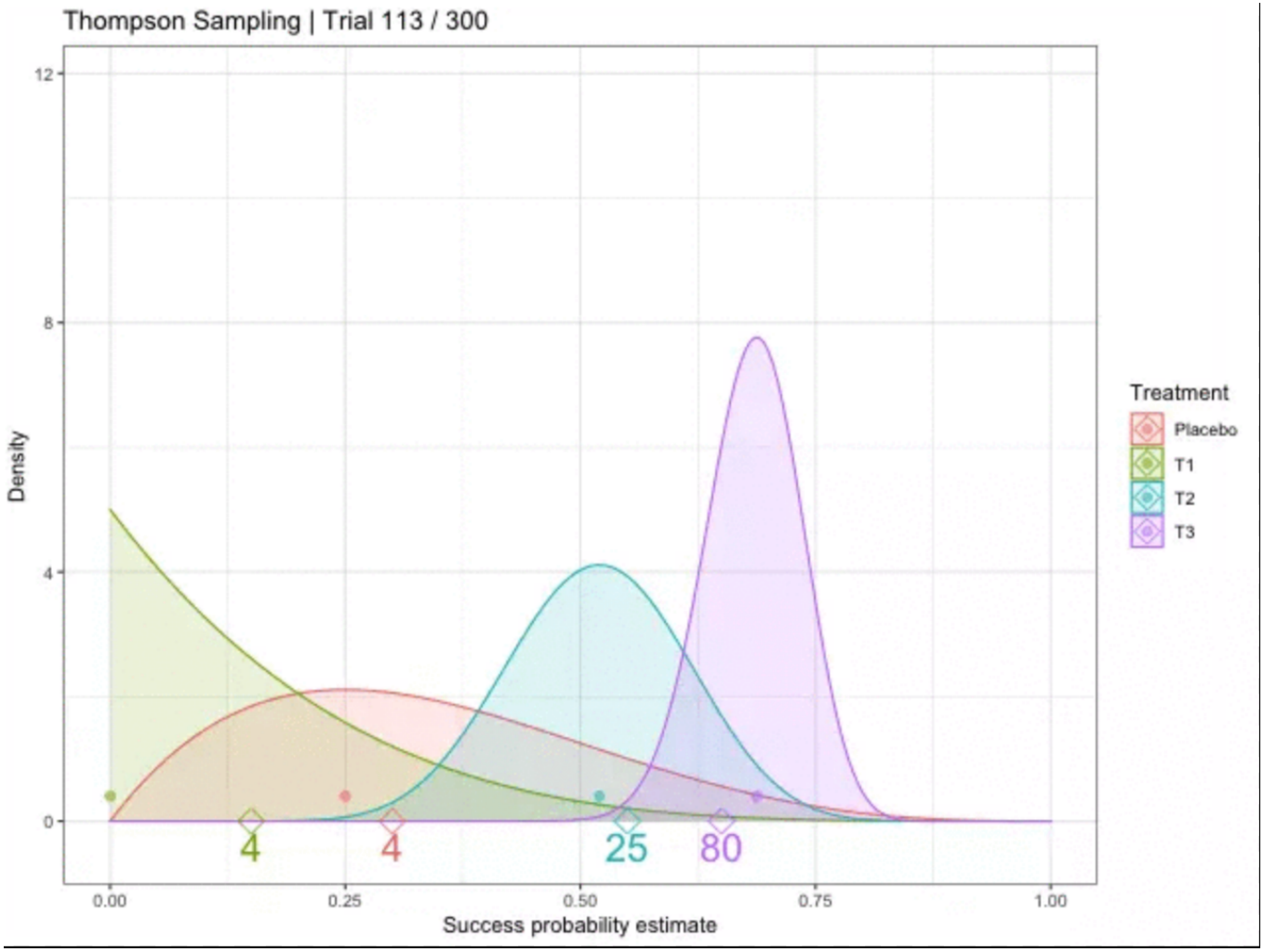
Thompson Sampling

In round $t+1$:

$$\forall a \in [K], \quad \widetilde{\theta}_a(t) \sim \pi_a(t)$$

$$A_{t+1} = \underset{a \in [K]}{\arg\max}\ \widetilde{\theta}_a(t)$$

Posterior Sampling in action



source: Wikipedia

Posterior Sampling: Theory

$$\mu_\star = \max_a \mu_a, \qquad a_\star = \arg\max_a \ \mu_a$$

Upper bound on sub-optimal selections

$$\forall a : \mu_a < \mu_\star, \quad \mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq \frac{\log(T)}{\mathcal{K}_{\inf}(\nu_a, \mu_\star)} + o_\mu(\log(T)).$$

where $\mathcal{K}_{\inf}(\nu_a, \mu_\star) = \min\{\mathrm{KL}(\nu_a, \nu) : \mathbb{E}_{X\sim\nu}[X] \geq \mu_\star\}$.

- proved for Bernoulli bandits, with a uniform prior

  [Kaufmann et al., 2012, Agrawal and Goyal, 2013]
- for 1-dimensional exponential families, with a conjuguate prior

  [Agrawal and Goyal, 2017, Korda et al., 2013]
- distributions with finite support with a Dirichlet prior

  [Riou and Honda, 2020]

Posterior Sampling: Theory

$$\mu_\star = \max_a \mu_a, \qquad a_\star = \arg\max_a \ \mu_a$$

Upper bound on sub-optimal selections

$$\forall a : \mu_a < \mu_\star, \quad \mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq \frac{\log(T)}{\mathcal{K}_{\inf}(\boldsymbol{\nu_a}, \boldsymbol{\mu_\star})} + o_\mu(\log(T)).$$

where $\mathcal{K}_{\inf}(\nu_a, \mu_\star) = \min\{\mathrm{KL}(\nu_a, \nu) : \mathbb{E}_{X\sim\nu}[X] \geq \mu_\star\}$.

- proved for Bernoulli bandits, with a uniform prior

  [Kaufmann et al., 2012, Agrawal and Goyal, 2013]
- for 1-dimensional exponential families, with a conjuguate prior

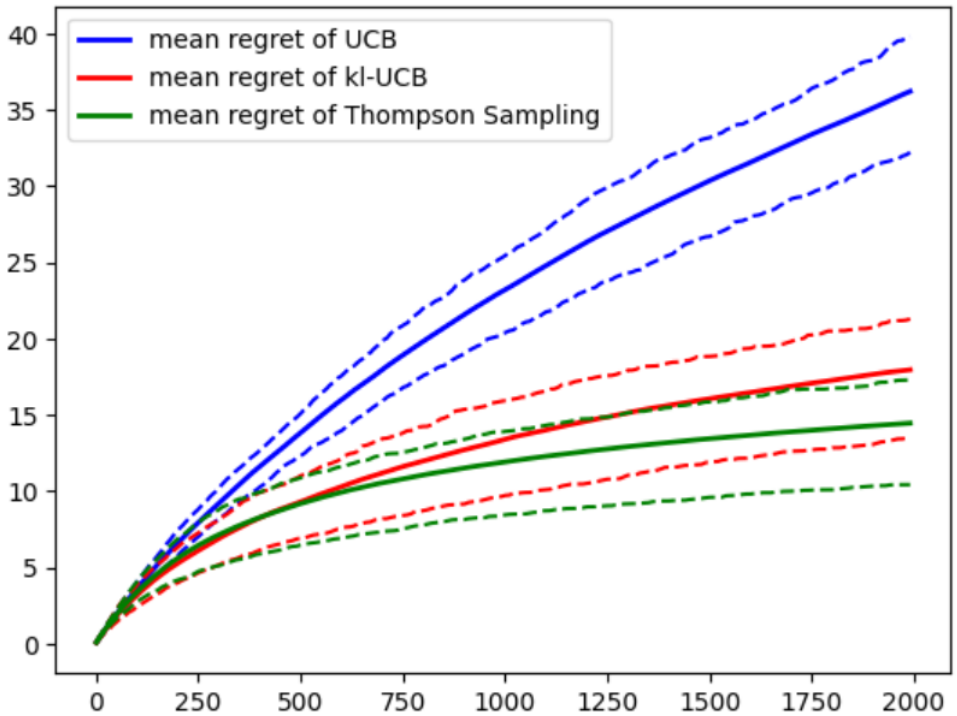
  [Agrawal and Goyal, 2017, Korda et al., 2013]
- distributions with finite support with a Dirichlet prior

  [Riou and Honda, 2020]

➜ its *regret* $\mathbb{E}\left[\sum_{t=1}^{T}(\mu_\star - \mu_{A_t})\right]$ is logarithmic (and **optimal**)

Posterior Sampling: Practice

$$R_{\boldsymbol{\mu}}(T) = \mathbb{E}\left[\sum_{t=1}^{T}(\mu_\star - \mu_{A_t})\right] = \sum_{a=1}^{K}(\mu_\star - \mu_a)\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]$$



regret estimated over $N = 500$ repetitions ($+$ 25%, 75% empirical quantiles) on a Bernoulli bandit $\boldsymbol{\mu} = [0.1\ 0.05\ 0.02\ 0.01]$

Outline

1. Posterior Sampling
2. Beyond Regret
3. Posterior Sampling for Pure Exploration
   - Top Two Thompson Sampling
   - Posterior Sampling Beyond Best Arm Identification
   - Murphy Sampling

Example: Adaptive Clinical Trial







$\mathcal{B}(\mu_1)$ $\mathcal{B}(\mu_2)$ $\mathcal{B}(\mu_3)$ $\mathcal{B}(\mu_4)$ $\mathcal{B}(\mu_5)$

For the t-th patient in a clinical trial:

- choose a treatment A_t
- observe a response $X_t \in \{0, 1\}$: $\mathbb{P}(X_t = 1) = \mu_{A_t}$

Minimize regret $\leftrightarrow$ cure as many patients as possible

$$\neq$$

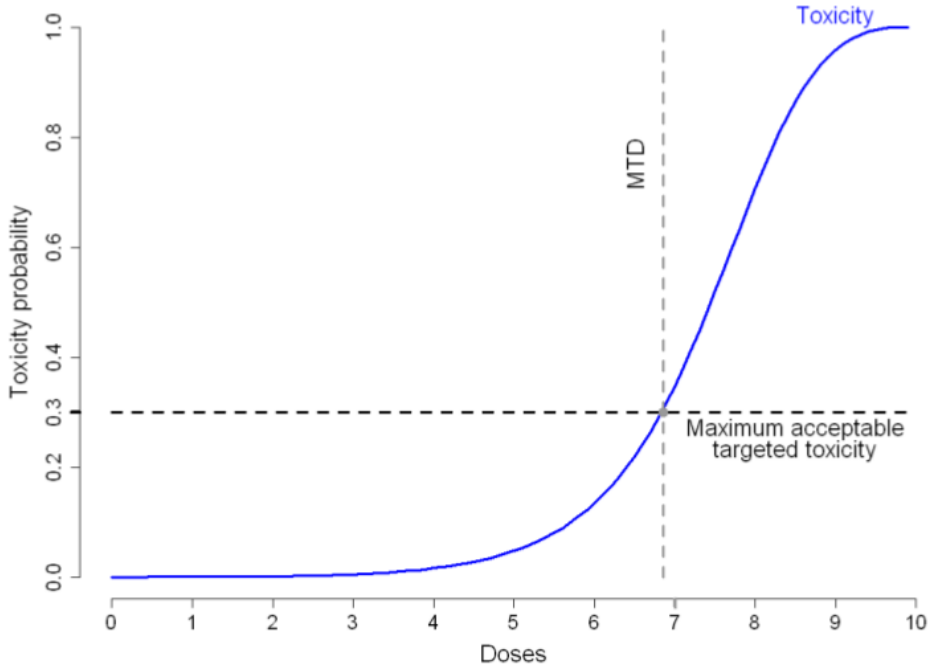
Best Arm Identification: identify as quickly as possible the best treatment (without trying to cure patients during the trial)

$$a_\star(\boldsymbol{\mu}) = \arg\max_a \ \mu_a$$

Another Pure Exploration Task

Phase I (Dose Finding) clinical trial:

- arms = doses of the same medication
- observation $X_t \in \{0, 1\}$ indicating *toxicity*



Goal: identify as quickly as possible the maximum tolerated dose:

$$i_\star(\boldsymbol{\mu}) = \arg\max_{a \in [K]} |\gamma - \mu_a|$$

[Garivier et al., 2019, Aziz et al., 2021]

Can Thompson Sampling find the best arm?

At time t, TS is selecting

$$A_t = \arg\max_{a \in [K]} \widetilde{\theta}_a(t-1)$$

Is arm A_t a reasonnable guess for the best arm?

Can Thompson Sampling find the best arm?

At time t, TS is selecting

$$A_t = \arg\max_{a \in [K]} \widetilde{\theta}_a(t-1)$$

Is arm A_t a reasonnable guess for the best arm?

Less explorative **recommendation rules**:

- empirical best arm: $\hat{\imath}_t = \arg\max_{a \in [K]} \hat{\mu}_a(t)$
- most played arm: $\hat{\imath}_t = \arg\max_{a \in [K]} N_a(t)$
- a smoother (randomized) version

$$\mathbb{P}(\hat{\imath}_t = b | \mathcal{H}_t) = \frac{N_b(t)}{t}$$

Can Thompson Sampling find the Best Arm?

For Thompson Sampling $+ \hat{\imath}_t \sim \left(\frac{N_1(t)}{t}, \dots, \frac{N_K(t)}{t} \right)$

$$\begin{aligned} \mathbb{P}_\nu(\hat{\imath}_t \neq a_\star) &= \sum_{b \neq a_\star} \mathbb{P}_\nu(\hat{\imath}_t = b) = \sum_{b \neq a_\star} \mathbb{E}_\nu \left[\mathbb{P}_\nu(\hat{\imath}_t = b | \mathcal{H}_t) \right] \\ &= \sum_{b \neq a_\star} \frac{\mathbb{E}_\nu[N_b(t)]}{t} \leq C_\nu \frac{\log(t)}{t} \end{aligned}$$

Can Thompson Sampling find the Best Arm?

For Thompson Sampling + $\hat{\imath}_t \sim \left(\frac{N_1(t)}{t}, \ldots, \frac{N_K(t)}{t}\right)$

$$\begin{aligned}\mathbb{P}_\nu(\hat{\imath}_t \neq a_\star) &= \sum_{b \neq a_\star} \mathbb{P}_\nu(\hat{\imath}_t = b) = \sum_{b \neq a_\star} \mathbb{E}_\nu\left[\mathbb{P}_\nu(\hat{\imath}_t = b | \mathcal{H}_t)\right] \\ &= \sum_{b \neq a_\star} \frac{\mathbb{E}_\nu[N_b(t)]}{t} \leq C_\nu \frac{\log(t)}{t}\end{aligned}$$

How good is this decay rate?

- worse than uniform sampling + empirical best arm (exponential decay)
- in order to guarantee $\mathbb{P}_\nu(B_t \neq a_\star) \leq \delta$, t has to be chosen as a function of the unknown instance ν

Outline

1. Posterior Sampling
2. Beyond Regret
3. Posterior Sampling for Pure Exploration
   - Top Two Thompson Sampling
   - Posterior Sampling Beyond Best Arm Identification
   - Murphy Sampling

Outline

1. Posterior Sampling
2. Beyond Regret
3. Posterior Sampling for Pure Exploration
   - Top Two Thompson Sampling
   - Posterior Sampling Beyond Best Arm Identification
   - Murphy Sampling

Top Two Thompson Sampling

$\Pi_t = (\pi_1(t), \dots, \pi_K(t))$ posterior distribution on $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$

Top-Two Thompson Sampling (TTTS) [Russo, 2016]

Input: parameter $\beta \in (0, 1)$. In round $t + 1$:

- draw a posterior sample $\boldsymbol{\theta} \sim \Pi_t$, $a_\star(\boldsymbol{\theta}) = \arg\max_a \theta_a$
- with probability β, select $A_{t+1} = a_\star(\boldsymbol{\theta})$
- with probability $1 - \beta$, re-sample the posterior $\boldsymbol{\theta}' \sim \Pi_t$ until $a_\star(\boldsymbol{\theta}') \neq a_\star(\boldsymbol{\theta})$, select $A_{t+1} = a_\star(\boldsymbol{\theta}')$

[Russo, 2016] Bayesian analysis of TTTS (for exp. families):

$$\Pi_t\left(\{\boldsymbol{\theta} : a_\star(\boldsymbol{\theta}) \neq a_\star\}\right) \lesssim C \exp\left(-t / T^\star_\beta(\boldsymbol{\mu})\right) \quad \text{a.s.}$$

where the rate is proved to be optimal.

Relation to Fixed Confidence BAI

For Gaussian bandits (known variance σ^2):

$$T^\star_\beta(\boldsymbol{\mu})^{-1} = \sup_{\substack{\boldsymbol{w}\in\triangle_K \\ w_{a^\star}=\beta}} \min_{a\neq a^\star} \frac{(\mu_{a_\star}-\mu_a)^2}{2\sigma^2\left(\frac{1}{w_{a_\star}}+\frac{1}{w_a}\right)}$$

Fixed-Confidence Best Arm Identification algorithm:

- sampling rule A_t
- recommendation rule $\hat{\imath}_t$ (guess for the best arm)
- stopping rule τ (when we think our guess is correct)

Lower bound [Garivier and Kaufmann, 2016]

For any δ-correct BAI algorithm, i.e. $\mathbb{P}_{\boldsymbol{\mu}}(\hat{\imath}_\tau \neq a_\star(\mu)) \leq \delta$ for all $\boldsymbol{\mu}$,

$$\forall \boldsymbol{\mu}, \quad \mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq T^\star(\boldsymbol{\mu}) \ln\left(\frac{1}{3\delta}\right),$$

where $T^\star(\boldsymbol{\mu}) = \min_{\beta\in(0,1)} T^\star_\beta(\boldsymbol{\mu})$.

Sample complexity of TTTS

For Gaussian bandits, one can analyze TTTS with the posterior

$$\pi_a(t) = \mathcal{N}\left(\hat{\mu}_a(t), \frac{\sigma^2}{N_a(t)}\right)$$

coupled with the (Generalized Likelihood Ratio) stopping rule

$$\tau_\delta = \inf\left\{ t \in \mathbb{N} : \min_{a \neq \hat{a}_t^\star} \frac{(\hat{\mu}_{\hat{a}_t^\star}(t) - \hat{\mu}_a(t))^2}{2\sigma^2\left(\frac{1}{N_{\hat{a}_t^\star}(t)} + \frac{1}{N_a(t)}\right)} \gtrsim \log\left(\frac{\log(t)}{\delta}\right) \right\}$$

with $\hat{a}_t^\star = \arg\max_a \hat{\mu}_a(t)$ and $\hat{\imath}_\tau = \hat{a}_\tau^\star$.

Theorem [Shang et al., 2020]

TTTS(β) is δ-correct and

$$\forall \boldsymbol{\mu}, \quad \lim_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leq T^\star_\beta(\boldsymbol{\mu})$$

Sample complexity of TTTS

For Gaussian bandits, one can analyze TTTS with the posterior

$$\pi_a(t) = \mathcal{N}\left(\hat{\mu}_a(t), \frac{\sigma^2}{N_a(t)}\right)$$

coupled with the (Generalized Likelihood Ratio) stopping rule

$$\tau_\delta = \inf\left\{ t \in \mathbb{N} : \min_{a \neq \hat{a}_t^\star} \frac{(\hat{\mu}_{\hat{a}_t^\star}(t) - \hat{\mu}_a(t))^2}{2\sigma^2\left(\frac{1}{N_{\hat{a}_t^\star}(t)} + \frac{1}{N_a(t)}\right)} \gtrsim \log\left(\frac{\log(t)}{\delta}\right) \right\}$$

with $\hat{a}_t^\star = \arg\max_a \hat{\mu}_a(t)$ and $\hat{\imath}_\tau = \hat{a}_\tau^\star$.

Theorem [Shang et al., 2020]

TTTS(1/2) is δ-correct and

$$\forall \boldsymbol{\mu}, \quad \lim_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leq 2T^\star(\boldsymbol{\mu})$$

Sample complexity of TTTS

For Gaussian bandits, one can analyze TTTS with the posterior

$$\pi_a(t) = \mathcal{N}\left(\hat{\mu}_a(t), \frac{\sigma^2}{N_a(t)}\right)$$

coupled with the (Generalized Likelihood Ratio) stopping rule

$$\tau_\delta = \inf\left\{ t \in \mathbb{N} : \min_{a \neq \hat{a}_t^\star} \frac{(\hat{\mu}_{\hat{a}_t^\star}(t) - \hat{\mu}_a(t))^2}{2\sigma^2\left(\frac{1}{N_{\hat{a}_t^\star}(t)} + \frac{1}{N_a(t)}\right)} \gtrsim \log\left(\frac{\log(t)}{\delta}\right) \right\}$$

with $\hat{a}_t^\star = \arg\max_a \hat{\mu}_a(t)$ and $\hat{\imath}_\tau = \hat{a}_\tau^\star$.

Rationale

$$T_\beta^\star(\boldsymbol{\mu})^{-1} = \sup_{\substack{\boldsymbol{w} \in \triangle_K \\ w_{a^\star} = \beta}} \min_{a \neq a^\star} \frac{(\mu_{a_\star} - \mu_a)^2}{2\sigma^2\left(\frac{1}{w_{a_\star}} + \frac{1}{w_a}\right)} \quad \text{and} \quad \frac{N_a(t)}{t} \overset{a.s.}{\to} w_a^\star(\boldsymbol{\mu})$$

Sample complexity of TTTS

For Gaussian bandits, one can analyze TTTS with the posterior

$$\pi_a(t) = \mathcal{N}\left(\hat{\mu}_a(t), \frac{\sigma^2}{N_a(t)}\right)$$

coupled with the (Generalized Likelihood Ratio) stopping rule

$$\tau_\delta = \inf\left\{ t \in \mathbb{N} : \min_{a \neq \hat{a}_t^\star} \frac{(\hat{\mu}_{\hat{a}_t^\star}(t) - \hat{\mu}_a(t))^2}{2\sigma^2\left(\frac{1}{N_{\hat{a}_t^\star}(t)} + \frac{1}{N_a(t)}\right)} \gtrsim \log\left(\frac{\log(t)}{\delta}\right) \right\}$$

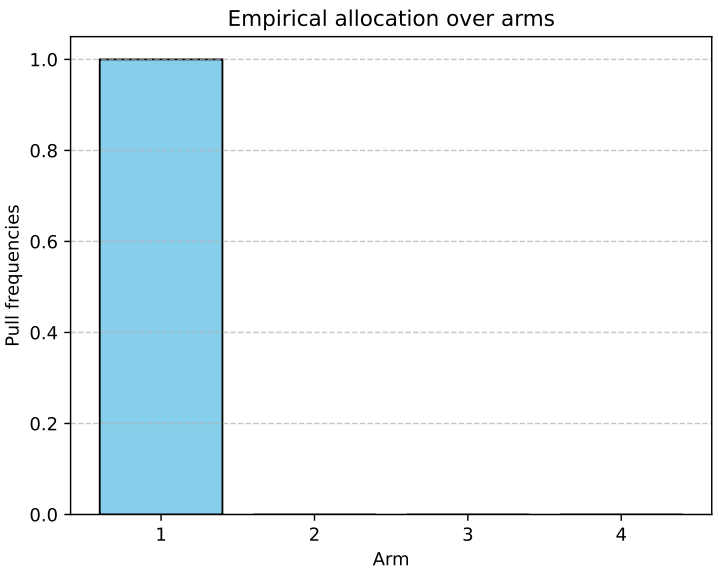
with $\hat{a}_t^\star = \arg\max_a \hat{\mu}_a(t)$ and $\hat{\imath}_\tau = \hat{a}_\tau^\star$.

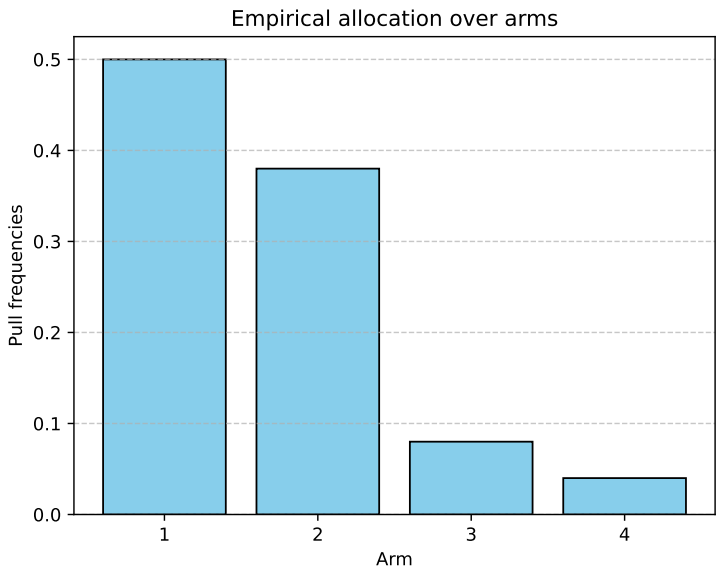
Rationale

$$T_\beta^\star(\boldsymbol{\mu})^{-1} = \min_{a \neq a^\star} \frac{(\mu_{a_\star} - \mu_a)^2}{2\sigma^2\left(\frac{1}{w_{a_\star}^\star(\boldsymbol{\mu})} + \frac{1}{w_a^\star(\boldsymbol{\mu})}\right)} \quad \text{and} \quad \frac{N_a(t)}{t} \overset{a.s.}{\to} w_a^\star(\boldsymbol{\mu})$$

TS versus TTTS

Fraction of samples allocated to each arm as $t \to \infty$:

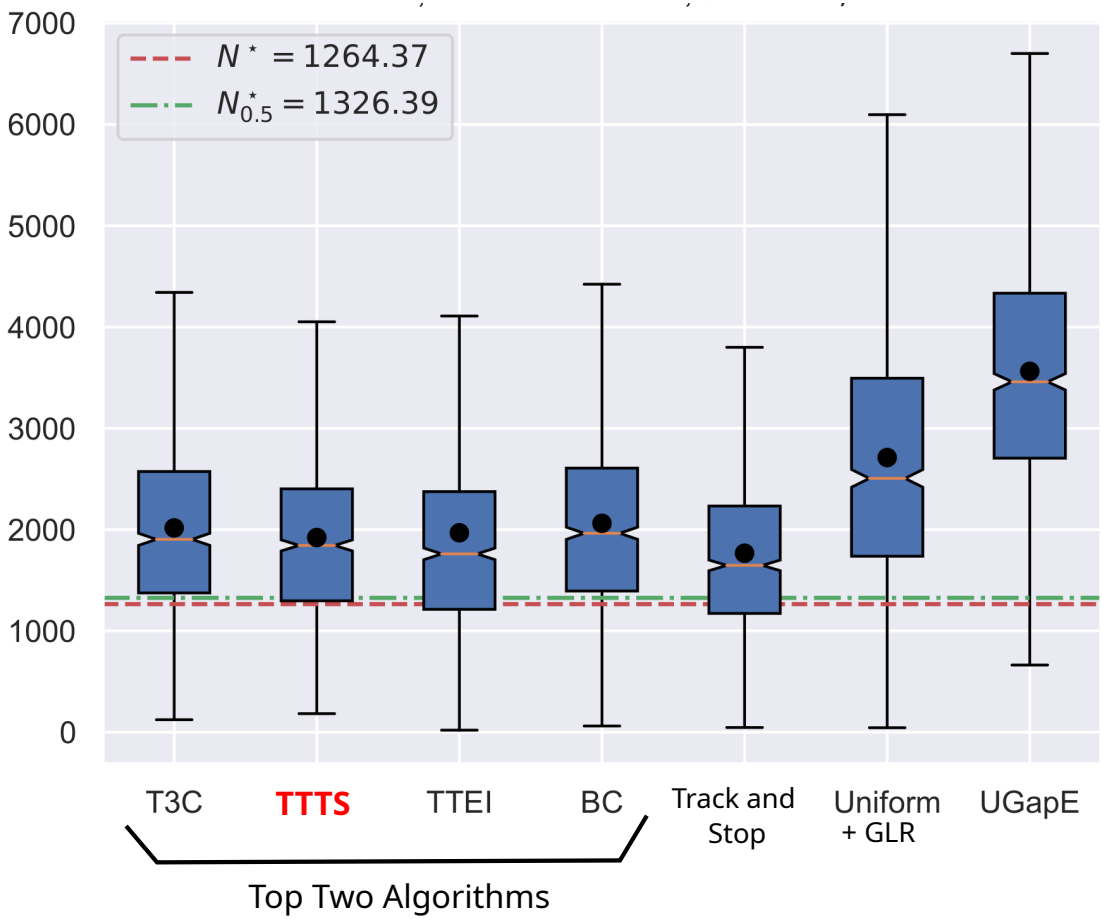




Thompson Sampling v.s. Top Two TS ($\beta = 1/2$)

TTTS in Practice

Empirical sample complexity for $\mu = [1\ 0.8\ 0.75\ 0.7]$ and $\sigma^2 = 1$ over $N = 1000$ runs, $\delta = 0.01$



TTTS: the first of many Top Two Algorithms

(analyzed beyond Gaussian distributions)

[Jourdan et al., 2022]

Outline

1. Posterior Sampling
2. Beyond Regret
3. Posterior Sampling for Pure Exploration
   - Top Two Thompson Sampling
   - Posterior Sampling Beyond Best Arm Identification
   - Murphy Sampling

Pure Exploration

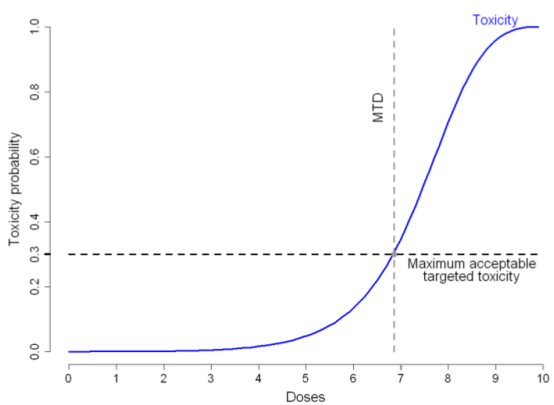
(For simplicity) Gaussian bandits, known variance σ^2
Possible vectors of arms means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K) \in \mathbb{R}^K$

Identification task

Given a correct answer function

$$i_\star : \mathcal{M} \longrightarrow \mathcal{I}$$
$$\boldsymbol{\mu} \mapsto i_\star(\boldsymbol{\mu})$$

find a correct answer with high probability.



$$i_\star(\boldsymbol{\mu}) = \arg\max_a |\gamma - \mu_a|$$

Pure Exploration with Fixed Confidence

An algorithm is made of:

- a sampling rule $A_t \in [K]$: what is the next arm to explore?

➜ get a new observation $X_t \sim \nu_{A_t}$

- a recommendation rule $\hat{\imath}_t$: a guess for the correct answer
- a stopping rule τ: when to stop the data collection?

Definition

An algorithm is δ-correct if, for all $\boldsymbol{\mu} \in \mathcal{M}$, $\mathbb{P}_{\boldsymbol{\mu}}(\hat{\imath}_\tau \neq i_\star(\boldsymbol{\mu})) \leq \delta$.

Goal: a δ-correct algorithm with small sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$

Lower bound and Stopping Rule

Lower bound (e.g., [Degenne et al., 2019])

For any δ-correct algorithm, i.e. $\mathbb{P}_{\boldsymbol{\mu}}\left(\hat{\imath}_\tau \neq i_\star(\mu)\right) \leq \delta$ for all $\boldsymbol{\mu}$,

$$\forall \boldsymbol{\mu}, \quad \mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq T^\star(\boldsymbol{\mu}) \ln\left(\frac{1}{3\delta}\right),$$

where

$$T^\star(\boldsymbol{\mu})^{-1} = \sup_{w \in \Sigma_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(i_\star(\boldsymbol{\mu}))} \sum_{a=1}^{K} w_a \frac{(\mu_a - \lambda_a)^2}{2\sigma^2}$$

with $\mathrm{Alt}(i) = \{\boldsymbol{\lambda} : i_\star(\boldsymbol{\lambda}) \neq i\}$.

Generalized Likelihood Ratio stopping rule: let

$$\tau_\delta = \inf\left\{ t : \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\imath}_t)} \sum_{a=1}^{K} N_a(t) \frac{(\hat{\mu}_a(t) - \lambda_a)^2}{2\sigma^2} \geq c(t,\delta) \right\}$$

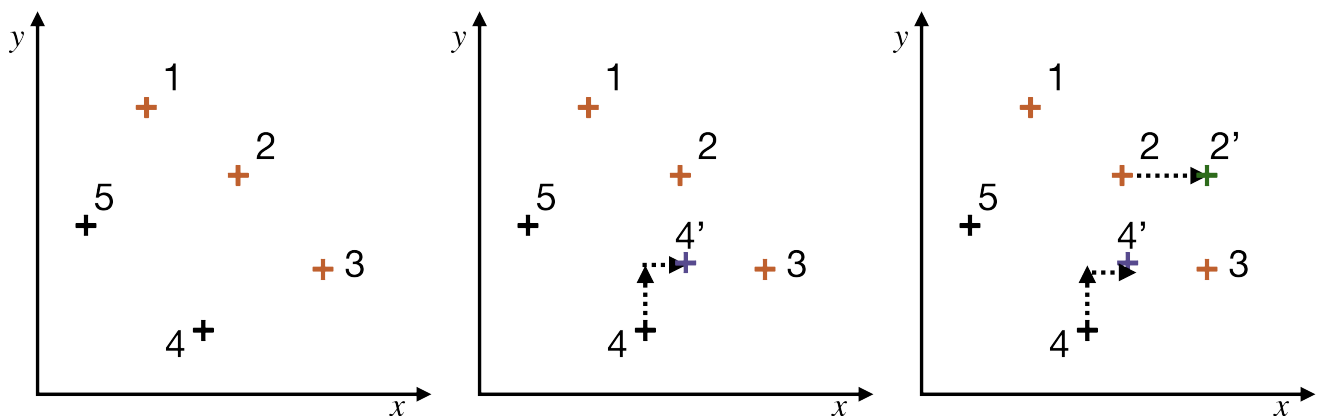
with $\hat{\imath}_t = i_\star(\hat{\mu}_a(t))$ the empirical correct answer.

The Generalized Likelihood Ratio Stopping Rule

The GLR stopping rule:

- is δ-correct for *any* pure exploration task for a threshold $c(t, \delta) \simeq \log(1/\delta) + K \log\log(t)$ [Kaufmann and Koolen, 2021]
- ... but can be intractable for some complex tasks $i_\star$

Example: Pareto Set Identification in multi-objective bandits [Kone et al., 2025]



Posterior Sampling to the Rescue

PS Stopping rule

For all $m \leq M(t,\delta)$, sample $\widetilde{\boldsymbol{\theta}}^m = (\widetilde{\boldsymbol{\theta}}_1^m, \dots, \widetilde{\boldsymbol{\theta}}_K^m)$ with

$$\widetilde{\boldsymbol{\theta}}_a^m \sim \mathcal{N}\left(\hat{\boldsymbol{\mu}}_a(t), \frac{c(t,\delta)\sigma^2}{N_a(t)}\right)$$

If for all m, $i_\star(\widetilde{\boldsymbol{\theta}}^m) = i_\star(\hat{\boldsymbol{\mu}}(t))$, **stop** and recommend $i_\star(\hat{\boldsymbol{\mu}}(t))$

- inspired by the TS-Explore strategy for Combinatorial bandits [Wang and Zhu, 2022]
- analyzed in [Kone et al., 2025] for Pareto Set Identification for

$$c(t,\delta) \simeq \frac{\log(\log(t)/\delta)}{\log(1/\delta)} \quad \text{and} \quad M(t,\delta) \simeq \frac{\log(t/\delta)}{\delta}$$

A Posterior Sampling Sampling Rule?

(Repeated) Posterior Sampling can be used to select an alternative answer that is promising: given

$$\underline{m} = \inf\{m : i_\star(\widetilde{\theta}_t^m) \neq i_\star(\hat{\mu}(t))\}$$

we need to discriminate between

$$i_\star(\hat{\mu}(t)) \quad \text{and} \quad i_\star\left(\widetilde{\theta}_t^{\underline{m}}\right)$$

➜ What arm to sample to do that is however non-trivial...

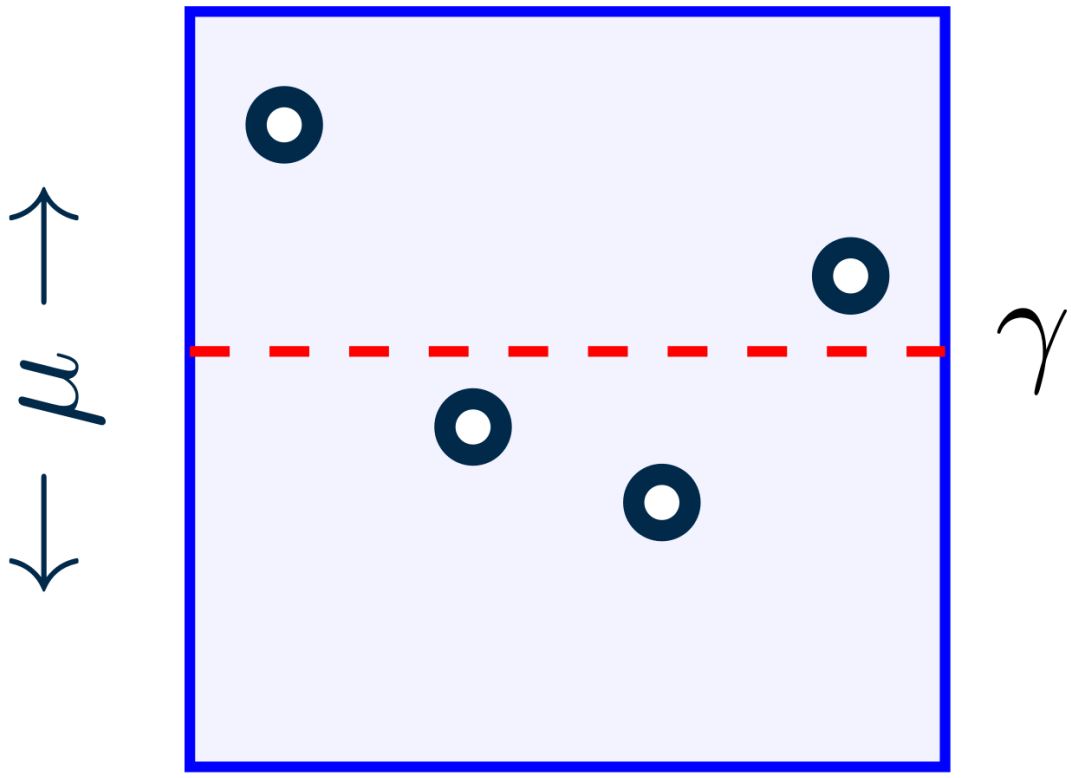
[Kone et al., 2025, Qin and You, 2025]

Outline

1. Posterior Sampling
2. Beyond Regret
3. Posterior Sampling for Pure Exploration
   - Top Two Thompson Sampling
   - Posterior Sampling Beyond Best Arm Identification
   - Murphy Sampling

Comparing the Smallest Mean to a Threshold

Is there an arm below γ?



$$i_\star(\boldsymbol{\mu}) = \mathbb{1}(\min_a \mu_a < \gamma) \in \{0, 1\}$$

[Kaufmann et al., 2018]

Lower Bound and Oracle Allocation

Lower bound: for *any* δ-correct algorithm,

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \;\geq\; T_\star(\boldsymbol{\mu}) \ln\left(\frac{1}{3\delta}\right).$$

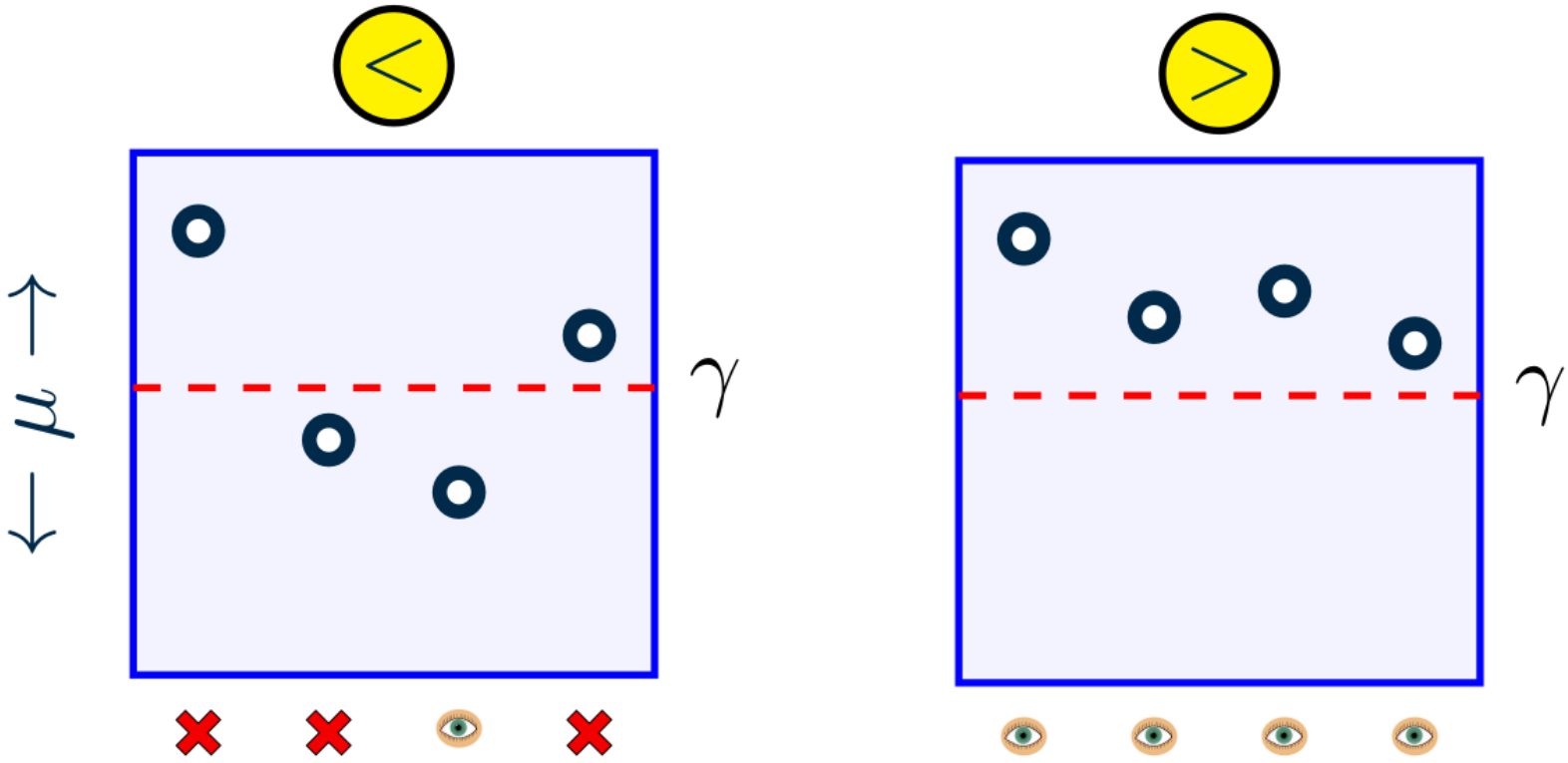
For our problem the *characteristic time* and *oracle weights* are

$$T_\star(\boldsymbol{\mu}) \;=\; \begin{cases} \dfrac{2\sigma^2}{(\mu_{\min}-\gamma)^2} & \mu_{\min} < \gamma, \\ \displaystyle\sum_a \dfrac{2\sigma^2}{(\mu_a-\gamma)^2} & \mu_{\min} > \gamma \end{cases}$$

$$(w_\star(\boldsymbol{\mu}))_a = \begin{cases} \mathbf{1}_{(a=a_\star)} & \mu_{\min} < \gamma, \\ \dfrac{\frac{1}{(\mu_a-\gamma)^2}}{\sum_j \frac{1}{(\mu_j-\gamma)^2}} & \mu_{\min} > \gamma \end{cases}$$

$(w_\star(\boldsymbol{\mu}))_a$: fraction of selections of arm a under a strategy that would match the lower bound

Murphy Sampling



Murphy Sampling [Kaufmann et al., 2018]

Sample $\boldsymbol{\theta} \sim \Pi_{t-1}(\cdot \,|\, \min_a \theta_a < \gamma)$, then play $A_t = \arg\min_a \ \theta_a$.

always condition on the "worst alternative"

Properties of Murphy Sampling

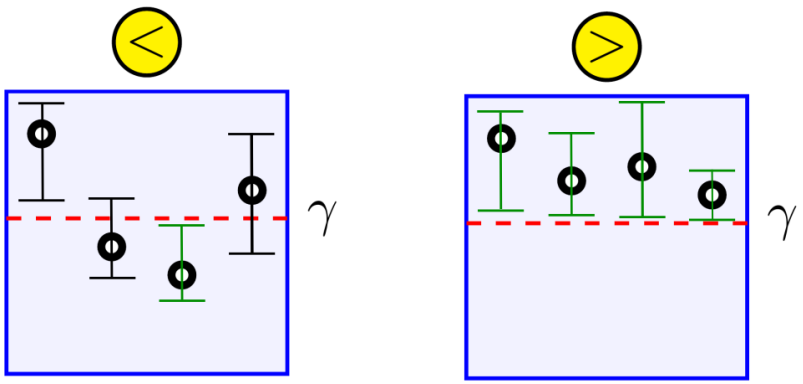
Theorem

For all μ, Murphy Sampling satisfies, for all a,

$$\frac{N_a(t)}{t} \to (w_\star(\mu))_a, \ \ a.s.$$

Murphy sampling together with $\tau = \tau_< \wedge \tau_>$ where

$$\begin{aligned} \tau_< &= \inf\{t \in \mathbb{N} : \exists a : \mathrm{UCB}_a(t) < \gamma\} \\ \tau_> &= \inf\{t \in \mathbb{N} : \forall a, \mathrm{LCB}_a(t) > \gamma\} \end{aligned}$$



satisfies $\limsup_{\delta \to 0} \frac{\tau_\delta}{\ln \frac{1}{\delta}} \leq T_\star(\mu), a.s.$.

Conclusion

Adding some appropriate conditionning to posterior sampling can make this algorithmic principle useful for pure exploration tasks.

We saw:

- a universal stopping rule
  (but its calibration is specific to each task)
- two task-specific (simple) sampling rules
  - ➜ Top-Two Thompson Sampling (BAI)
  - ➜ Murphy Sampling (Is the Minimum below γ?)

Perspective:

- For what kind of pure exploration task can we have a simple posterior-sampling based sampling rule?
- When to use the PS stopping rule over the frequentist GLR?


Agrawal, S. and Goyal, N. (2013).
Further Optimal Regret Bounds for Thompson Sampling.
In *Proceedings of the 16th Conference on Artificial Intelligence and Statistics*.


Agrawal, S. and Goyal, N. (2017).
Near-optimal regret bounds for thompson sampling.
J. ACM, 64(5):30:1–30:24.


Aziz, M., Kaufmann, E., and Riviere, M. (2021).
On multi-armed bandit designs for dose-finding clinical trials.
Journal of Machine Learning Research, 22(14):1–38.


Degenne, R., Nedelec, T., Calauzènes, C., and Perchet, V. (2019).
Bridging the gap between regret minimization and best arm identification, with application to A/B tests.
In *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*.


Garivier, A. and Kaufmann, E. (2016).
Optimal best arm identification with fixed confidence.
In *Proceedings of the 29th Conference On Learning Theory*.


Garivier, A., Ménard, P., and Rossi, L. (2019).
Thresholding bandit for dose-ranging: The impact of monotonicity.
In *International Conference on Machine Learning, Artificial Intelligence and Applications*.


Jourdan, M., Degenne, R., Baudry, D., de Heide, R., and Kaufmann, E. (2022).
Top two algorithms revisited.
In *Advances in Neural Information Processing Systems (NeurIPS)*.


Kaufmann, E. and Koolen, W. (2021).
Mixture martingales revisited with applications to sequential tests and confidence intervals.
Journal of Machine Learning Research, 22(246).


Kaufmann, E., Koolen, W., and Garivier, A. (2018).
Sequential test for the lowest mean: From Thompson to Murphy Sampling.
In *Advances in Neural Information Processing Systems (NeurIPS)*.


Kaufmann, E., Korda, N., and Munos, R. (2012).
Thompson Sampling : an Asymptotically Optimal Finite-Time Analysis.
In *Proceedings of the 23rd conference on Algorithmic Learning Theory (ALT)*.


Kone, C., Jourdan, M., and Kaufmann, E. (2025).
Pareto set identification with posterior sampling.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.


Korda, N., Kaufmann, E., and Munos, R. (2013).
Thompson Sampling for 1-dimensional Exponential family bandits.
In *Advances in Neural Information Processing Systems*.


Qin, C. and You, W. (2025).
Dual-directed algorithm design for efficient pure exploration.
arXiv:2310.19319.


Riou, C. and Honda, J. (2020).
Bandit algorithms based on thompson sampling for bounded reward distributions.
In *Algorithmic Learning Theory (ALT)*.


Russo, D. (2016).
Simple Bayesian algorithms for best arm identification.
In *Proceedings of the 29th Conference on Learning Theory (COLT)*.


Russo, D., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. (2018).
A tutorial on thompson sampling.
Foundations and Trends in Machine Learning, 11(1):1–96.


Shang, X., de Heide, R., Kaufmann, E., Ménard, P., and Valko, M. (2020).
Fixed-confidence guarantees for bayesian best-arm identification.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.


Thompson, W. (1933).
On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.
Biometrika, 25:285–294.


Wang, S. and Zhu, J. (2022).
Thompson sampling for (combinatorial) pure exploration.
In *International Conference on Machine Learning*. PMLR.