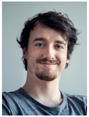
A tale of two (non-parametric) bandit problems

Emilie Kaufmann

Université de Lille

based on collaborations with
Dorian Baudry, Odalric-Ambrym Maillard,
Marc Jourdan, Rémy Degenne & Rianne de Heide

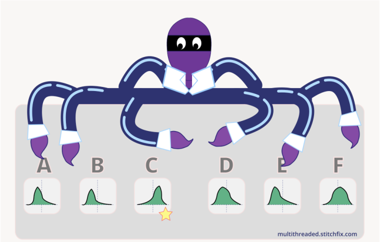




CWI, February 2023

The stochastic Multi Armed Bandit (MAB) model

- K *unknown* reward distributions $\nu_1, \dots, \nu_K$ called *arms*
- a each time t, select an arm A_t and observe a reward $X_t \sim \nu_{A_t}$



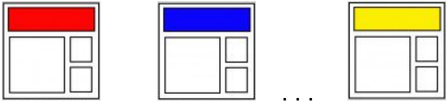
Sequential strategy / algorithm : A_{t+1} can depend on:

- previous observation $A_1, X_1, \dots, A_t, X_t$
- some external randomization $U_t \sim \mathcal{U}([0,1])$
- some knowledge about the type of reward distributions

[Thompson, 1933, Robbins, 1952, Lattimore and Szepesvari, 2019]

Bandit problems

Example: A/B/n testing



p_1 p_2 ... p_K

p_a: probability that a visitor seeing version a buys a product

For the t-th visitor:

- choose a version A_t to display
- observe the reward $X_t = 1$ if a product is bought, 0 otherwise

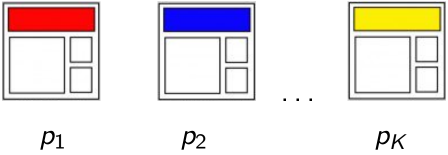
Objective 1: maximize rewards

- maximize $\mathbb{E}[\sum_{t=1}^{T} X_t]$ for some (possibly unknown) T
- maximize profit

a *reinforcement learning* problem

Bandit problems

Example: A/B/n testing



p_a: probability that a visitor seeing version a buys a product

For the t-th visitor:

- choose a version A_t to display
- observe the reward $X_t = 1$ if a product is bought, 0 otherwise

Objective 2: best arm identification

- identify quickly $a_\star = \arg\max_a \ p_a$
- find the best version (in order to keep displaying it)

a *pure exploration* problem

Other applications

- clinical trials → reward: success/failure (Bernoulli)

- movie recommendation → reward: rating (multinomial)



- recommendation in agriculture → reward: yield (complex, possibly multi-modal distribution)

Objective: design algorithms that leverage as little knowledge about the rewards distributions as possible

Outline

Performance measure

$$\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K) \qquad \mu_a = \mathbb{E}_{X\sim\nu_a}[X]$$

$$\mu_\star = \max_{a\in\{1,\ldots,K\}} \mu_a \qquad a_\star = \underset{a\in\{1,\ldots,K\}}{\arg\max}\ \mu_a.$$

Maximizing rewards $\leftrightarrow$ selecting $a_\star$ as much as possible
$\leftrightarrow$ minimizing the regret [Robbins, 52]

$$\mathcal{R}_{\boldsymbol{\nu}}(\mathcal{A}, T) = \underbrace{T\mu_\star}_{\substack{\text{sum of rewards of}\\ \text{an oracle strategy}\\ \text{always selecting } a_\star}} - \underbrace{\mathbb{E}_{\boldsymbol{\nu}}\left[\sum_{t=1}^{T} X_t\right]}_{\substack{\text{sum of rewards of}\\ \text{the strategy}\mathcal{A}}}$$

Regret decomposition

$$\mathcal{R}_{\boldsymbol{\nu}}(\mathcal{A}, T) = \mathbb{E}_{\boldsymbol{\nu}}\left[\sum_{t=1}^{T}(\mu_\star - \mu_a)\right]$$

$N_a(T)$: number of selections of arm a up to round T.

Performance measure

$$\boldsymbol{\nu} = (\nu_1, \dots, \nu_K) \qquad \mu_a = \mathbb{E}_{X \sim \nu_a}[X]$$

$$\mu_\star = \max_{a \in \{1,\dots,K\}} \mu_a \qquad a_\star = \operatorname*{arg\,max}_{a \in \{1,\dots,K\}} \mu_a.$$

Maximizing rewards $\leftrightarrow$ selecting $a_\star$ as much as possible

$\leftrightarrow$ minimizing the regret [Robbins, 52]

$$\mathcal{R}_{\boldsymbol{\nu}}(\mathcal{A}, T) = \underbrace{T\mu_\star}_{\substack{\text{sum of rewards of} \\ \text{an oracle strategy} \\ \text{always selecting } a_\star}} - \underbrace{\mathbb{E}_{\boldsymbol{\nu}}\left[\sum_{t=1}^{T} X_t\right]}_{\substack{\text{sum of rewards of} \\ \text{the strategy } \mathcal{A}}}$$

Regret decomposition

$$\mathcal{R}_{\boldsymbol{\nu}}(\mathcal{A}, T) = \sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\nu}}[N_a(T)](\mu_\star - \mu_a)$$

$N_a(T)$: number of selections of arm a up to round T.

(Don't) Follow The Learder

Select each arm once, then exploit the current knowledge:

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

where

- $N_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$ is the number of selections of arm a
- $\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} X_s \mathbb{1}(A_s = a)$ is the empirical mean of the rewards collected from arm a

(Don't) Follow The Learder

Select each arm once, then exploit the current knowledge:

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

where

- $N_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$ is the number of selections of arm a
- $\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^{t} X_s \mathbb{1}(A_s = a)$ is the empirical mean of the rewards collected from arm a

Follow the leader can fail! $\nu_1 = \mathcal{B}(\mu_1), \nu_2 = \mathcal{B}(\mu_2)$, $\mu_1 > \mu_2$

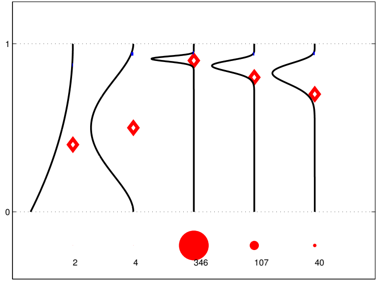
$$\mathbb{E}[N_2(T)] \geq (1 - \mu_1)\mu_2 \times (T - 1)$$

➜ **Exploitation** is not enough, we need to add some **exploration**

A Bayesian algorithm: Thompson Sampling

$\pi_a(0)$: prior distribution on μ_a

$\pi_a(t) = \mathcal{L}(\mu_a | Y_{a,1}, \dots, Y_{a,N_a(t)})$: posterior distribution on μ_a



Two equivalent interpretations:

- [Thompson, 1933]: "randomize the arms according to their posterior probability being optimal"
- modern view: "draw a possible bandit model from the posterior distribution and act optimally in this sampled model"

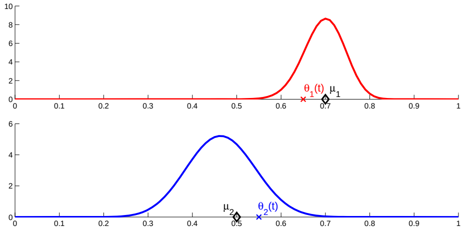
Russo et al. 2018, *A Tutorial on Thompson Sampling*

Input: a prior distribution $\pi(0)$

$$\begin{cases} \forall a \in \{1..K\}, \quad \theta_a(t) \sim \pi_a(t) \\ A_{t+1} = \underset{a=1...K}{\text{argmax}}\ \theta_a(t). \end{cases}$$

Thompson Sampling for Bernoulli distributions $\nu_a = \mathcal{B}(\mu_a)$

- $\pi_a(0) = \mathcal{U}([0,1])$
- $\pi_a(t) = \mathrm{Beta}\left(S_a(t)+1; N_a(t)-S_a(t)+1\right)$



A Bayesian algorithm: Thompson Sampling

Input: a prior distribution $\pi(0)$

$$\begin{cases} \forall a \in \{1..K\}, \quad \theta_a(t) \sim \pi_a(t) \\ A_{t+1} = \underset{a=1...K}{\operatorname{argmax}}\, \theta_a(t). \end{cases}$$

Thompson Sampling for Bernoulli distributions $\nu_a = \mathcal{B}(\mu_a)$

- $\pi_a(0) = \mathcal{U}([0,1])$
- $\pi_a(t) = \mathrm{Beta}\left(S_a(t) + 1; N_a(t) - S_a(t) + 1\right)$

Thompson Sampling for Gaussian distributions $\nu_a = \mathcal{N}(\mu_a, \sigma^2)$

- $\pi_a(0) \propto 1$
- $\pi_a(t) = \mathcal{N}\left(\hat{\mu}_a(t); \frac{\sigma^2}{N_a(t)}\right)$

An asymptotically optimal algorithm

Upper bound on sub-optimal selections

$$\forall a \neq a_\star, \quad \mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq \frac{\log(T)}{\mathrm{kl}(\mu_a, \mu_\star)} + o_\mu(\log(T)).$$

where $\mathrm{kl}(\mu_a, \mu_\star)$ is the KL divergence between ν_a and $\nu_{a_\star}$

- 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]

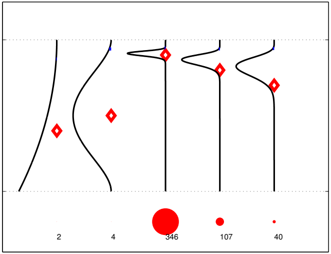
Lower bound [Lai and Robbins, 1985]

Let $\mathcal{D}$ be a family of rewards distribution that are continuously parameterized by their means. Any *good* bandit algorithm for $\mathcal{D}$ satisfies, on every instance with means $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$

$$\forall a \neq a_\star, \quad \liminf_{T \to \infty} \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]}{\log(T)} \geq \frac{1}{\mathrm{kl}(\mu_a; \mu_\star)}$$

Beyond parametric algorithms?

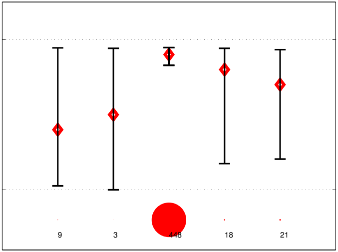
Thompson Sampling
(TS)



$$A_{t+1} = \underset{a \in [K]}{\text{argmax}}\ \theta_a(t)$$

where $\theta_a(t)$ is a sample from a posterior distribution on μ_a

Upper Confidence Bound
(UCB)



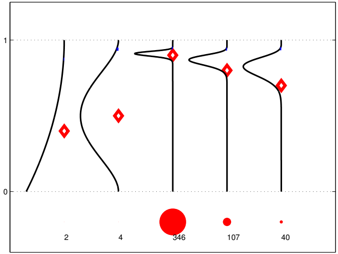
$$A_{t+1} = \underset{a \in [K]}{\text{argmax}}\ \mathrm{UCB}_a(t)$$

$\mathrm{UCB}_a(t)$ is an UCB on the unknown mean μ_a

➜ require some tuning depending on the distributions

Beyond parametric algorithms?

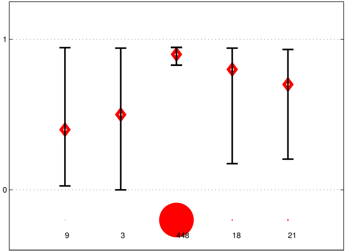
Thompson Sampling
(TS)



$$A_{t+1} = \underset{a \in [K]}{\text{argmax}}\ \theta_a(t)$$

where $\theta_a(t)$ is a sample from a posterior distribution on μ_a

Upper Confidence Bound
(UCB)



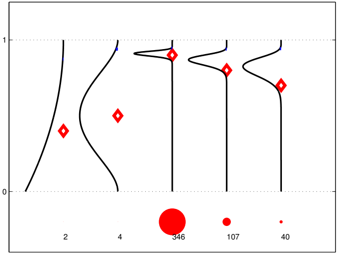
$$A_{t+1} = \underset{a \in [K]}{\text{argmax}}\ \mathrm{UCB}_a(t)$$

$\mathrm{UCB}_a(t)$ is an UCB on the unknown mean μ_a

➜ what is F_a is any distribution supported on $[0, B]$?

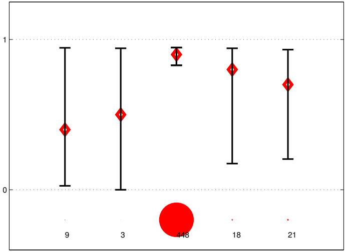
Beyond parametric algorithms?

Thompson Sampling (TS)



???

Upper Confidence Bound (UCB)



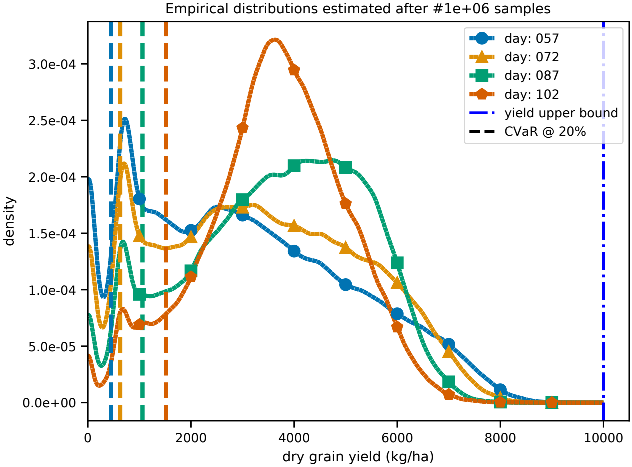
$$A_{t+1} = \underset{a \in [K]}{\text{argmax}} \ \mathrm{UCB}_a(t)$$

$$\mathrm{UCB}_a(t) = \hat{\mu}_a(t) + B\sqrt{\frac{\log(t)}{2N_a(t)}}$$

➜ what is F_a is any distribution supported on $[0, B]$?

Outline

Motivation: recommending planting dates to farmers



Distribution of the yield of a maize field for different planting dates obtained using the DSSAT crop-yield simulator

Optimality in Non Parametric families

Can we adapt *optimally* to complex bounded distributions?

Lower bound [Burnetas and Katehakis, 1996]

Under an algorithm achieving small regret for any bandit model $\nu \in \mathcal{D}^K$, it holds that

$$\forall a \neq a_\star(\boldsymbol{\nu}), \quad \liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\nu}}[N_a(T)]}{\log(T)} \geq \frac{1}{\mathcal{K}_{\inf}^{\mathcal{D}}(F_a;\mu_\star)}$$

where

$$\mathcal{K}_{\inf}^{\mathcal{D}}(\nu,\mu) = \inf\left\{ \mathrm{KL}(\nu,\nu') \middle| \, \nu' \in \mathcal{D} : \mathbb{E}_{X\sim\nu'}[X] \geq \mu \right\}$$

with $\mathrm{KL}(\nu,\nu')$ the Kullback-Leibler divergence.

$$\mathcal{D}_B = \left\{ \nu \in \mathcal{P}(\mathbb{R}), \nu \text{ is supported on } [0, B] \right\}$$

Non Parametric Thompson Sampling

$$A_{t+1} = \arg\max_{a \in [K]} \theta_a(t)$$

where

$$\theta_a(t) = \frac{1}{N_a(t)+1}\left(\sum_{i=1}^{N_a(t)} w_{a,t}(i) Y_{a,i} + w_{a,t}(N_a(t)+1) B\right)$$

with

- $(Y_{a,1}, \dots, Y_{a,N_a(t)}, B)$ is the augmented history of rewards gathered from arm a
- $w_{a,t} \sim \mathrm{Dir}(\underbrace{1, \dots, 1}_{N_a(t)+1})$ a random probability vector

[Riou and Honda, 2020]

Several interpretations:

- an extension of multinomial Thompson Sampling
- a variant of the Bayesian bootstrap
- posterior sampling using a Dirichlet Process prior

A risk-averse bandit problem

Specifics of our application:

➜ **bounded** distributions, with known upper bound B

➜ quality of an arm measured by its Conditional Value at Risk

$$\mathrm{CVaR}_\alpha(\nu_a) = \sup_{x \in \mathbb{R}} \left\{ x - \frac{1}{\alpha} \mathbb{E}_{X \sim \nu_a} \left[(x - X)^+ \right] \right\}$$

Interpretation of the CVaR:

- if ν is continuous, $\mathrm{CVaR}_\alpha(\nu) = \mathbb{E}_{X \sim \nu} \left[X | X \leq F^{-1}(\alpha) \right]$
- if ν is discrete, with values $x_1 \leq x_2 \leq \cdots \leq x_M$

$$\mathrm{CVaR}_\alpha(\nu) = \frac{1}{\alpha} \left[\sum_{i=1}^{n_\alpha - 1} p_i x_i + \left(\alpha - \sum_{i=1}^{n_\alpha - 1} p_i x_i \right) x_{n_\alpha} \right]$$

where $n_\alpha = \inf \{ n : \sum_{i=1}^{n} p_i x_i \geq \alpha \}$.

➜ average of the lower part of the distribution

A risk-averse bandit problem

Specifics of our application:

➜ **bounded** distributions, with known upper bound B

➜ quality of an arm measured by its Conditional Value at Risk

$$\mathrm{CVaR}_\alpha(\nu_a) = \sup_{x \in \mathbb{R}} \left\{ x - \frac{1}{\alpha} \mathbb{E}_{X \sim \nu_a} \left[(x - X)^+ \right] \right\}$$

Interpretation of the CVaR:

Choosing α allows to customize the risk-aversion:

- $\alpha = 20\%$: farmer seeking to avoid very poor yield
- $\alpha = 80\%$: market-oriented farmer trying to optimize the yield of non-extraordinary years
- $\alpha = 100\%$: optimization of the average yield (no risk aversion)

Specifics of our application:

- ➜ **bounded** distributions, with known upper bound B
- ➜ quality of an arm measured by its Conditional Value at Risk

$$\mathrm{CVaR}_\alpha(\nu_a) = \sup_{x\in\mathbb{R}} \left\{ x - \frac{1}{\alpha}\mathbb{E}_{X\sim\nu_a}\left[(x-X)^+\right] \right\}$$

Interpretation of the CVaR:

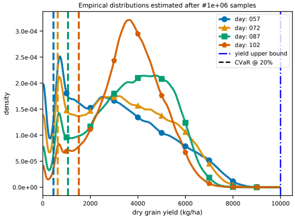


Table 3: Empirical yield distribution metrics in kg/ha estimated after 10^6 samples in `DSSAT` environment

day (action)	CVaR$_\alpha$			
	5%	20%	80%	100% (mean)
057	0	448	2238	3016
072	46	627	2570	3273
087	287	1059	3074	**3629**
102	**538**	**1515**	**3120**	3586

CVaR regret

Letting $c_a^\alpha = \mathrm{CVaR}_\alpha(\nu_a)$, the CVaR regret is defined as

$$\mathcal{R}_{\boldsymbol{\nu}}^\alpha(\mathcal{A}, T) = \mathbb{E}_{\boldsymbol{\nu}}\left[\sum_{t=1}^{T}(c_\star^\alpha - c_{A_t}^\alpha)\right] = \sum_{a=1}^{K}(c_\star^\alpha - c_a^\alpha)\,\mathbb{E}_{\boldsymbol{\nu}}[N_a(T)]$$

with $c_\star^\alpha = \max_a c_a^\alpha$.

Lower bound [Baudry et al., 2021]

Under an algorithm achieving small CVaR regret for any bandit model $\boldsymbol{\nu} \in \mathcal{D}^K$, it holds that

$$\forall a : c_a^\alpha < c_\star^\alpha, \quad \liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\nu}}[N_a(T)]}{\log(T)} \geq \frac{1}{\mathcal{K}_{\inf}^{\alpha,\mathcal{D}}(\nu_a; c_\star^\alpha)}$$

where $\mathcal{K}_{\inf}^{\alpha,\mathcal{D}}(\nu, c) = \inf\left\{\mathrm{KL}(\nu, \nu')\,|\nu' \in \mathcal{D} : \mathrm{CVaR}_\alpha(\nu') \geq c\right\}$.

Non Parametric Thompson Sampling for CVaR bandits

Assumption: $\nu_a \in \mathcal{D}_B = \{\text{distributions supported in } [0, B]\}$.

The **B-CVTS** algorithm selects

$$A_{t+1} \in \arg\max_{a \in [K]} C_a(t)$$

Index of arm a after t rounds

- $\overline{\mathcal{H}}_a(t) = (Y_{a,1}, \dots, Y_{a,N_a(t)}, B)$ be the augmented history of rewards gathered from this arm
- $w_{a,t} \sim \mathrm{Dir}(\underbrace{1, \dots, 1}_{N_a(t)+1})$ a random probability vector

➜ yields a random perturbation of the empirical distribution $\widetilde{F}_{a,t} = \sum_{i=1}^{N_a(t)} w_{a,t}(i)\delta_{Y_{a,i}} + w_{a,t}(N_a(t)+1)\delta_B$

$$C_a(t) = \mathrm{CVaR}_\alpha\left(\widetilde{F}_{a,t}\right)$$

$\alpha = 1 \rightarrow$ Non Parametric Thompson Sampling

[Riou and Honda, 2020]

B-CVTS is asymptotically optimal for bounded distributions.

Theorem [Baudry et al., 2021]

On an instance $\boldsymbol{\nu}$ such that $\boldsymbol{\nu} \in \mathcal{D}_B^K$, we have

$$\mathbb{E}_{\boldsymbol{\nu}}[N_a(T)] \leq \frac{\log T}{\mathcal{K}_{\inf}^{\alpha,\mathcal{D}_B}(\nu_a, c_1^\alpha)} + o(\log T) \,.$$

Key tool: new bounds on the *boundary crossing probability*

$$\mathbb{P}_{w\sim\mathcal{D}_n}\Big(\mathrm{C}_\alpha(\mathcal{Y}, w) > c\Big)$$

where

- $\mathcal{D}_n$ is a $\mathrm{Dir}(1, \dots, 1)$ distribution (with n ones)
- $\mathcal{Y} = \{y_1, \dots, y_n\}$ is a fixed support
- $\mathrm{C}_\alpha(\mathcal{Y}, w)$ is the α CVaR of a discrete distribution with support $\mathcal{Y}$ and weights w

Theory

B-CVTS is asymptotically optimal for bounded distributions.

Theorem [Baudry et al., 2021]

On an instance $\boldsymbol{\nu}$ such that $\boldsymbol{\nu} \in \mathcal{D}_B^K$, we have

$$\mathbb{E}_{\boldsymbol{\nu}}[N_a(T)] \leq \frac{\log T}{\mathcal{K}_{\text{inf}}^{\alpha, \mathcal{D}_B}(\nu_a, c_1^\alpha)} + o(\log T) \ .$$

Key tool: new bounds on the *boundary crossing probability*

$$\mathbb{P}_{w \sim \mathcal{D}_n}\left(\mathrm{C}_\alpha(\mathcal{Y}, w) > c\right) \simeq \exp\left(-n \mathcal{K}_{\text{inf}}^{\alpha, \mathcal{D}_B}(\mathcal{U}(\mathcal{Y}), c)\right)$$

where

- $\mathcal{D}_n$ is a $\mathrm{Dir}(1, \dots, 1)$ distribution (with n ones)
- $\mathcal{Y} = \{y_1, \dots, y_n\}$ is a fixed support
- $\mathrm{C}_\alpha(\mathcal{Y}, w)$ is the α CVaR of a discrete distribution with support $\mathcal{Y}$ and weights w

Practice

Competitors: two styles of UCB algorithms

- U-UCB [Cassel et al., 2018] uses the empirical cdf $\hat{F}_{a,t}$

$$\mathrm{UCB}_a^{(1)}(t) = \mathrm{CVaR}_\alpha(\hat{F}_{a,t}) + \frac{B}{\alpha}\sqrt{\frac{c\log(t)}{2N_a(t)}}$$

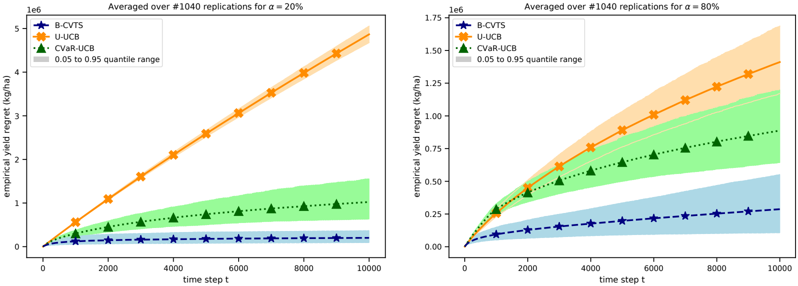
- CVaR-UCB: [Tamkin et al., 2020] buids an optimistic cdf $\overline{F}_{a,t}$

$$\mathrm{UCB}_a^{(2)}(t) = \mathrm{CVaR}_\alpha(\overline{F}_{a,t})$$

Table 4: Empirical yield regrets at horizon 10^4 in t/ha in DSSAT environment, for 1040 replications. Standard deviations in parenthesis.

α	U-UCB	CVaR-UCB	B-CVTS
5%	3128 (3)	760 (14)	**192 (11)**
20%	4867 (11)	1024 (17)	**202 (10)**
80%	1411 (13)	888 (13)	**287 (12)**

Practice



Regret as a function of time averaged over $N = 1040$ simulations for $\alpha = 20\%$ (left) and $\alpha = 80\%$ (right)

Outline

Best Arm Identification

Algorithm: made of three components:

- ➜ sampling rule: A_t (arm to explore)
- ➜ recommendation rule: B_t (current guess for the best arm)
- ➜ stopping rule τ (when do we stop exploring?)

- **Objectives studied in the literature:**

Fixed-budget setting	Fixed-confidence setting
input: budget T	**input:** risk parameter δ
$\tau = T$ minimize $\mathbb{P}(B_T \neq a_\star)$	minimize $\mathbb{E}[\tau]$ $\mathbb{P}(B_\tau \neq a_\star) \leq \delta$
[Bubeck et al., 2011] [Audibert et al., 2010]	[Even-Dar et al., 2006]

Finding the Best Arm with Thompson Sampling

B_T: guess for the best arm after T samples.

Thompson Sampling selects a lot the best arm...

- idea (1): $B_T = \arg\max_a N_a(T)$
- idea (2) : $\mathbb{P}(B_T = a) = \frac{N_a(T)}{T}$

Thompson Sampling + (2):

$$\begin{aligned} \mathbb{E}[\mu_\star - \mu_{B_T}] &= \mathbb{E}\left[\sum_{a=1}^{K}(\mu_\star - \mu_a)\frac{N_a(T)}{T}\right] \\ &= \frac{\mathcal{R}(\mathrm{TS}, T)}{T} = O\left(\frac{K\log(T)}{\Delta T}\right) \end{aligned}$$

☺ the estimation error decays with T

Finding the Best Arm with Thompson Sampling

B_T: guess for the best arm after T samples.

Thompson Sampling selects a lot the best arm...

- idea (1): $B_T = \arg\max_a N_a(T)$
- idea (2) : $\mathbb{P}(B_T = a) = \frac{N_a(T)}{T}$

Thompson Sampling + (2):

$$\begin{aligned}\mathbb{E}[\mu_\star - \mu_{B_T}] &= \mathbb{E}\left[\sum_{a=1}^{K}(\mu_\star - \mu_a)\frac{N_a(T)}{T}\right] \\ &= \frac{\mathcal{R}(\mathrm{TS}, T)}{T} = O\left(\frac{K\log(T)}{\Delta T}\right)\end{aligned}$$

☺ the estimation error decays with T

Uniform Sampling + Empirical Best Arm:

$$\mathbb{E}[\mu_\star - \mu_{B_T}] = O\left(K\exp\left(-\frac{T}{K}\Delta^2\right)\right)$$

☹ but not as fast as with uniform sampling...

Finding the Best Arm with Thompson Sampling

B_T: guess for the best arm after T samples.

Thompson Sampling selects a lot the best arm...

- idea (1): $B_T = \arg\max_a N_a(T)$
- idea (2) : $\mathbb{P}(B_T = a) = \frac{N_a(T)}{T}$

Thompson Sampling + (2):

$$\Delta \mathbb{P}\left(B_T \neq a_\star\right) \simeq \mathbb{E}\left[\sum_{a=1}^{K} (\mu_\star - \mu_a) \frac{N_a(T)}{T}\right]$$
$$= \frac{\mathcal{R}(\mathrm{TS}, T)}{T} = O\left(\frac{K \log(T)}{\Delta T}\right)$$

☺ the estimation error decays with T

Uniform Sampling + Empirical Best Arm:

$$\Delta \mathbb{P}\left(B_T \neq a_\star\right) \simeq O\left(K \exp\left(-\frac{T}{K}\Delta^2\right)\right)$$

☹ but not as fast as with uniform sampling...

Top Two Thompson Sampling

$\Pi_t = (\pi_1(t), \ldots, \pi_K(t))$ posterior distribution on $(\mu_1, \ldots, \mu_K)$

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] performs a Bayesian analysis of TTTS:

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

where the rate is proved to be optimal.

(for exponential families, and some restricted family of priors)

The optimal exponent

➜ connected with the optimal sample complexity of *fixed-confidence* best arm identification

Lower bound [Garivier and Kaufmann, 2016]

For any strategy such that $\mathbb{P}_{\boldsymbol{\nu}}\left(B_\tau \neq a_\star(\nu)\right) \leq \delta$ for all $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K) \in \mathcal{D}^K$,

$$\forall \boldsymbol{\nu} \in \mathcal{D}^K, \quad \mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \geq T^\star(\boldsymbol{\nu}) \ln\left(\frac{1}{3\delta}\right),$$

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

General expression:

$$T^\star_\beta(\boldsymbol{\nu})^{-1} = \sup_{\substack{\boldsymbol{w} \in \triangle_K \\ w_{a_\star} = \beta}} \min_{a \neq a^\star} \inf_{x \in \mathcal{I}} \left[w_{a_\star} \mathcal{K}^-_{\mathrm{inf}}(\nu_{a_\star}, x) + w_a \mathcal{K}^+_{\mathrm{inf}}(\nu_a, x) \right] .$$

The optimal exponent

➜ connected with the optimal sample complexity of *fixed-confidence* best arm identification

Lower bound [Garivier and Kaufmann, 2016]

For any strategy such that $\mathbb{P}_{\boldsymbol{\nu}}\left(B_\tau \neq a_\star(\nu)\right) \leq \delta$ for all $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K) \in \mathcal{D}^K$,

$$\forall \boldsymbol{\nu} \in \mathcal{D}^K, \quad \mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta] \geq T^\star(\boldsymbol{\nu}) \ln\left(\frac{1}{3\delta}\right),$$

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

Parametric example: Gaussian bandits

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

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 (GLR) stopping rule

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

with threshold $\beta(t,\delta) \simeq \log(1/\delta) + K \log\log(t)$.

Theorem [Shang et al., 2020]

TTTS(β) is δ-correct and

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

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 (GLR) stopping rule

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

with threshold $\beta(t,\delta) \simeq \log(1/\delta) + K \log\log(t)$.

Theorem [Shang et al., 2020]

TTTS(1/2) is δ-correct and

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

Outline

The Top Two structure

Top Two algorithm

Given a parameter $\beta \in (0, 1)$, in round t:

- define a leader $B_t \in [K]$
- define a challenger $C_t \neq B_t$
- select arm $A_t \in \{B_t, C_t\}$ at random:

$$\mathbb{P}(A_t = B_t) = \beta \qquad \mathbb{P}(A_t = C_t) = 1 - \beta$$

In Top Two Thompson Sampling,

- TS leader: $B_t = a_\star(\boldsymbol{\theta})$ with $\boldsymbol{\theta} \sim \Pi_{t-1}$
- Re-Sampling (RS) challenger: $C_t = a_\star(\boldsymbol{\theta}')$ where

$$\boldsymbol{\theta}' \sim \Pi_{t-1} | \left(a_\star(\boldsymbol{\theta}') \neq B_t\right)$$

The Top Two structure

Top Two algorithm

Given a parameter $\beta \in (0, 1)$, in round t:

- define a leader $B_t \in [K]$
- define a challenger $C_t \neq B_t$
- select arm $A_t \in \{B_t, C_t\}$ at random:

$$\mathbb{P}(A_t = B_t) = \beta \qquad \mathbb{P}(A_t = C_t) = 1 - \beta$$

In Top Two Thompson Sampling,

- TS leader: $B_t = a_\star(\boldsymbol{\theta})$ with $\boldsymbol{\theta} \sim \Pi_{t-1}$
- Re-Sampling (RS) challenger: $C_t = a_\star(\boldsymbol{\theta}')$ where

$$\boldsymbol{\theta}' \sim \Pi_{t-1} | \left(a_\star(\boldsymbol{\theta}') \neq B_t\right)$$

Liminations:

➜ re-sampling can be numerically costly

➜ beyond parametric distributions?

Approximating Re-Sampling

Under the RS challenger,

$$\mathbb{P}\left(C_t = a \middle| B_t = b\right) = \frac{p_{t,a}}{\sum_{i \neq b} p_{t,i}}$$

where $p_{t,a} = \Pi_t\left(\theta_a = \max_j \theta_j\right) \simeq \Pi_t\left(\theta_a > \theta_b\right)$.

For Gaussian bandits when $\hat{\mu}_b(t) > \hat{\mu}_a(t)$,

$$\Pi_t\left(\theta_a > \theta_b\right) \simeq \exp\left(-t \frac{(\hat{\mu}_b(t) - \hat{\mu}_a(t))^2}{2\sigma^2\left(\frac{1}{N_b(t)} + \frac{1}{N_a(t)}\right)}\right)$$

Idea: compute the mode instead of sampling!

$$C_t = \underset{a \neq B_t}{\arg\min} \frac{(\hat{\mu}_{B_t}(t) - \hat{\mu}_a(t))^2}{2\sigma^2\left(\frac{1}{N_{B_t}(t)} + \frac{1}{N_a(t)}\right)} \mathbb{1}(\hat{\mu}_{B_t}(t) \geq \hat{\mu}_a(t))$$

[Shang et al., 2020]

Links with the stopping rule

Recall that TTTS was analyzed with

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

➜ another interpretation: challenger that minimizes the Transportation Cost (TC) featured in the stopping rule

Recall that TTTS was analyzed with

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

➜ another interpretation: challenger that minimizes the Transportation Cost (TC) featured in the stopping rule

This idea extends to the non-parametric setting

$$\begin{aligned} W_t(i,j) &= \inf_x \left[N_i(t) \mathcal{K}_{\inf}^{\mathcal{D},-}(F_i(t), x) + N_j(t) \mathcal{K}_{\inf}^{\mathcal{D},+}(F_j(t), x) \right] \\ C_t &= \arg\min_{a \neq B_t} W_t(B_t, a) \end{aligned}$$

Links with the stopping rule

Recall that TTTS was analyzed with

$$\tau_\delta = \inf \left\{ t \in \mathbb{N} : \min_{a \neq \hat{a}_t^\star} W_t\left(\hat{a}_t, a\right) > \beta(t, \delta) \right\}$$

➜ another interpretation: challenger that minimizes the Transportation Cost (TC) featured in the stopping rule

This idea extends to the non-parametric setting

$$\begin{aligned} W_t(i,j) &= \inf_x \left[N_i(t)\mathcal{K}_{\inf}^{\mathcal{D},-}(F_i(t), x) + N_j(t)\mathcal{K}_{\inf}^{\mathcal{D},+}(F_j(t), x) \right] \\ C_t &= \arg\min_{a \neq B_t} W_t(B_t, a) \end{aligned}$$

... provided that we know how to calibrate the stopping rule

Top Two Algorithms

- Choices of the leader:

TS - Sample $\theta \sim \Pi_{t-1}$ then set $B_t^{\text{TS}} \in \arg\max_{a \in [K]} \theta_a$

EB - $B_t^{\text{EB}} \in \arg\max_{a \in [K]} \hat{\mu}_a(t-1)$

- Choices of the challenger:

RS - repeat $\theta \sim \Pi_{t-1}$ until $C_t^{\text{RS}} \in \arg\max_{a \in [K]} \theta_a \neq B_t$

TC - $C_t^{\text{TC}} \in \arg\min_{a \neq B_t} W_{t-1}(B_t, a)$

TCI - $C_t^{\text{TCI}} \in \arg\min_{a \neq B_t} W_{t-1}(B_t, a) + \log N_a(t)$

Top Two Algorithms

- Choices of the leader:

TS - Sample $\theta \sim \Pi_{t-1}$ then set $B_t^{\text{TS}} \in \arg\max_{a \in [K]} \theta_a$

EB - $B_t^{\text{EB}} \in \arg\max_{a \in [K]} \hat{\mu}_a(t-1)$

- Choices of the challenger:

RS - repeat $\theta \sim \Pi_{t-1}$ until $C_t^{\text{RS}} \in \arg\max_{a \in [K]} \theta_a \neq B_t$

TC - $C_t^{\text{TC}} \in \arg\min_{a \neq B_t} W_{t-1}(B_t, a)$

TCI - $C_t^{\text{TCI}} \in \arg\min_{a \neq B_t} W_{t-1}(B_t, a) + \log N_a(t)$

Π_t: a sampler (e.g. posterior distribution)

➜ parameteric setting: posterior distribution

➜ bounded distribution: Dirichlet Sampling

Asymptotic Guarantees

Theorem

Given a calibrated GLR stopping rule, instantiating the Top Two sampling rule with any pair of leader/challenger satisfying some properties yields a δ-correct algorithm satisfying for all $\boldsymbol{\nu} \in \mathcal{D}^K$ with distincts means

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\nu}}[\tau_\delta]}{\log(1/\delta)} \leq T^\star_\beta(\boldsymbol{\nu}) \, .$$

Distributions	TS	EB	RS	TC	TCI
Gaussian KV	✓	✓	✓	✓	✓
Bernoulli	✓	✓	✓	✓	✓
sub-Exp SPEF	?	✓	?	✓	✓
Gaussian UV	?	✓	?	✓	✓
Bounded	✓	✓	✓	✓	✓

[Jourdan et al., 2022, Jourdan et al., 2023]

Experiments: Bounded distributions

arm = planting date / observation = yield

Moderate regime, $\delta = 0.01$. Top Two algorithms with $\beta = 1/2$.

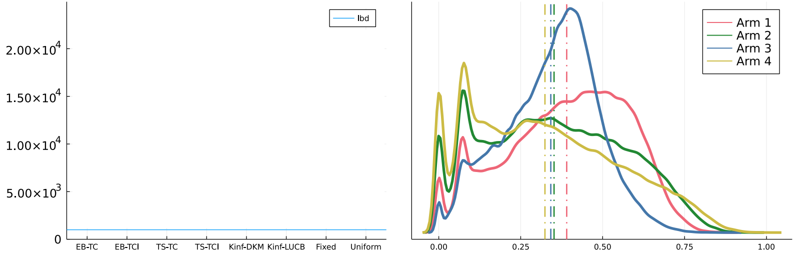


Figure: Empirical stopping time (a) on scaled DSSAT instances with their density and mean (b). Lower bound is $T^\star(\boldsymbol{\nu})\ln(1/\delta)$.

Experiments: Bounded distributions

arm = planting date / observation = yield

Moderate regime, $\delta = 0.01$. Top Two algorithms with $\beta = 1/2$.

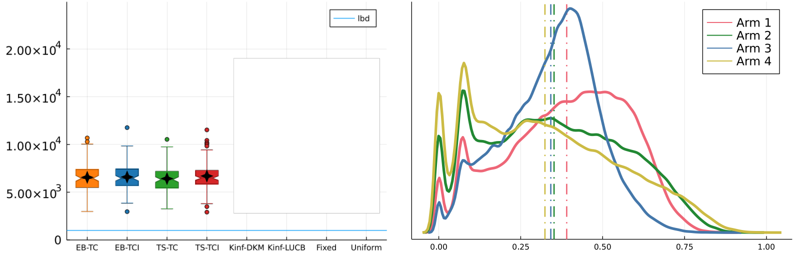


Figure: Empirical stopping time (a) on scaled DSSAT instances with their density and mean (b). Lower bound is $T^\star(\boldsymbol{\nu}) \ln(1/\delta)$.

Experiments: Bounded distributions

arm = planting date / observation = yield

Moderate regime, $\delta = 0.01$. Top Two algorithms with $\beta = 1/2$.

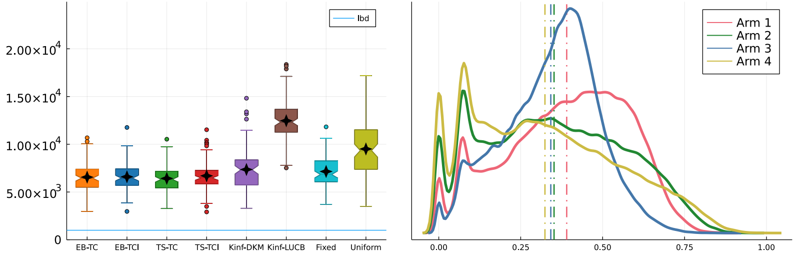


Figure: Empirical stopping time (a) on scaled DSSAT instances with their density and mean (b). Lower bound is $T^{\star}(\boldsymbol{\nu}) \ln(1/\delta)$.

Moderate regime, $\delta = 0.1$. Top Two algorithms with $\beta = 1/2$.

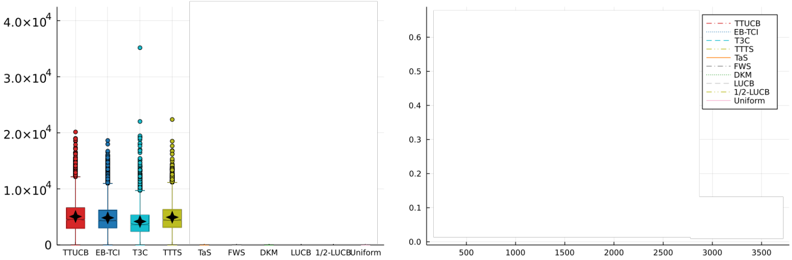


Figure: (Left) Empirical stopping time τ_δ. (Right) Empirical errors $\mathbb{P}(\hat{a}_t^\star \neq a_\star)$ at time $t < \tau_\delta$ on random instances with $K = 10$, $\mu_1 = 0.6$, $\mu_a \sim \mathcal{U}([0.2, 0.5])$.

Experiments: Gaussian distributions

Moderate regime, $\delta = 0.1$. Top Two algorithms with $\beta = 1/2$.

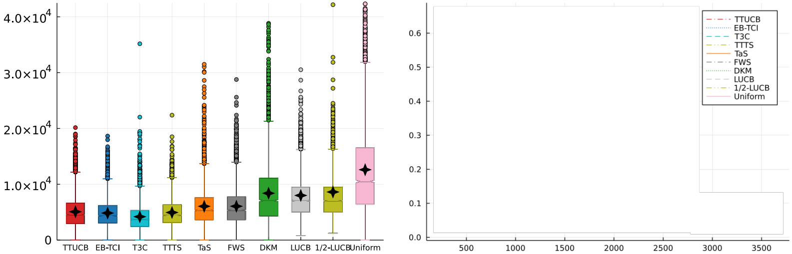


Figure: (Left) Empirical stopping time τ_δ. (Right) Empirical errors $\mathbb{P}(\hat{a}_t^\star \neq a_\star)$ at time $t < \tau_\delta$ on random instances with $K = 10$, $\mu_1 = 0.6$, $\mu_a \sim \mathcal{U}([0.2, 0.5])$.

Moderate regime, $\delta = 0.1$. Top Two algorithms with $\beta = 1/2$.

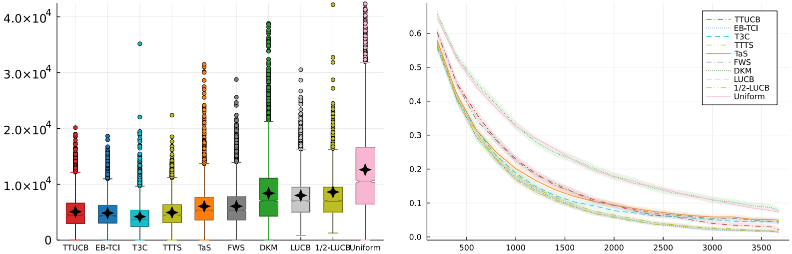


Figure: (Left) Empirical stopping time τ_δ. (Right) Empirical errors $\mathbb{P}(\hat{a}_t^\star \neq a_\star)$ at time $t < \tau_\delta$ on random instances with $K = 10$, $\mu_1 = 0.6$, $\mu_a \sim \mathcal{U}([0.2, 0.5])$.

Conclusion

Thompson Sampling for maximizing rewards:

- is asymptotically optimal for simple parametric distributions
- can be extended to some non-parametric settings
- is flexible enough to tackle alternative performance criterion

Top Two Thompson Sampling for best arm identification:

- may be viewed as a fix of TS for BAI
- is a inspiration for others (non-Bayesian) Top Two algorithms
- ... which are near optimal in theory and very good in practice

Perspective: finite-time performance?

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.

Audibert, J.-Y., Bubeck, S., and Munos, R. (2010).
Best Arm Identification in Multi-armed Bandits.
In *Proceedings of the 23rd Conference on Learning Theory*.

Baudry, D., Gautron, R., Kaufmann, E., and Maillard, O. (2021).
Optimal Thompson Sampling strategies for support-aware CVaR bandits.
In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Bubeck, S., Munos, R., and Stoltz, G. (2011).
Pure Exploration in Finitely Armed and Continuous Armed Bandits.
Theoretical Computer Science 412, 1832-1852, 412:1832–1852.

Burnetas, A. and Katehakis, M. (1996).
Optimal adaptive policies for sequential allocation problems.
Advances in Applied Mathematics, 17(2):122–142.

Cassel, A., Mannor, S., and Zeevi, A. (2018).
A general approach to multi-armed bandits under risk criteria.
In *Proceedings of the 31st Annual Conference On Learning Theory*.

Even-Dar, E., Mannor, S., and Mansour, Y. (2006).

Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems.
Journal of Machine Learning Research, 7:1079–1105.

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

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)*.

Jourdan, M., Degenne, R., and Kaufmann, E. (2023).
Dealing with unknown variances in best-arm identification.
In *Algorithmic Learning Theory (ALT)*.

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*.

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

Lai, T. and Robbins, H. (1985).
Asymptotically efficient adaptive allocation rules.
Advances in Applied Mathematics, 6(1):4–22.

Lattimore, T. and Szepesvari, C. (2019).
Bandit Algorithms.

Cambridge University Press.

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

Robbins, H. (1952).
Some aspects of the sequential design of experiments.
Bulletin of the American Mathematical Society, 58(5):527–535.

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

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)*.

Tamkin, A., Keramati, R., Dann, C., and Brunskill, E. (2020).
Distributionally-aware exploration for cvar bandits.
In *NeurIPS 2019 Workshop on Safety and Robustness in Decision Making; RLDM 2019*.

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