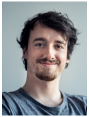
A Tale of Top Two Algorithms

Emilie Kaufmann

based on collaborations with

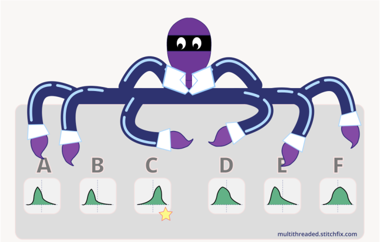
Marc Jourdan, Rémy Degenne, Dorian Baudry & Rianne de Heide

Workshop on Bandits and Statistical Tests, Potsdam, November 2023

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 $X_t \sim \nu_{A_t}$



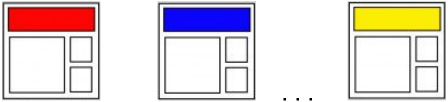
Sequential strategy / algorithm : A_{t+1} can depend on:

- 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: $\nu_a \in \mathcal{D}$

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

Two classical 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 $X_t = 1$ if a product is bought, 0 otherwise

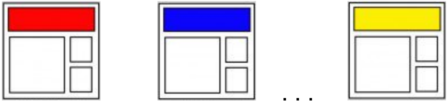
Objective 1: observation = reward → maximize rewards

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

a *reinforcement learning* problem

Two classical bandit problems

Example: A/B/n testing



p_1 p_2 $\ldots$ 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 $X_t = 1$ 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)

an *adaptive testing* problem

Other applications

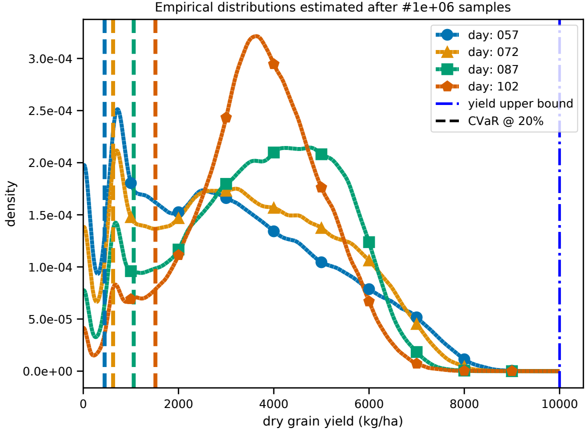
- clinical trials $\rightarrow$ observation: success/failure (Bernoulli)

- movie recommendation $\rightarrow$ observation: rating (multinomial)

- website optimization $\rightarrow$ observation: amount of money spent (Gaussian distribution?)

Other applications

- recommendation in agriculture → reward: yield (complex bounded distribution)



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

Outline

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

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

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

Performance measure

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

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

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

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

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

Best Achievable Regret

Lower bound [Burnetas and Katehakis, 1996]

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

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

where

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

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

Gaussian bandits

$$\mathcal{D} = \left\{ \mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R} \right\}$$

$$\mathcal{K}_{\text{inf}}^{\mathcal{D}}(\mathcal{N}(\mu, \sigma^2); \mu') = \frac{(\mu - \mu')^2}{2\sigma^2}$$

Best Achievable Regret

Lower bound [Burnetas and Katehakis, 1996]

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

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

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.

Bernoulli bandits

$$\mathcal{D} = \{\mathcal{B}(\mu), \mu \in [0,1]\}$$

$$\mathcal{K}_{\inf}^{\mathcal{D}}(\mathcal{B}(\mu);\mu') = \mu \log\frac{\mu}{\mu'} + (1-\mu)\log\frac{1-\mu}{1-\mu'}$$

Best Achievable Regret

Lower bound [Burnetas and Katehakis, 1996]

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

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

where

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

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

Single Parameter Exponential Family (SPEF)

$$\mathcal{D} = \{\nu_\mu, \mu \in \mathcal{I}\}$$

$$\mathcal{K}_{\text{inf}}^{\mathcal{D}}(\nu_\mu; \mu') = \mathrm{KL}\left(\nu_\mu, \nu_{\mu'}\right)$$

Best Achievable Regret

Lower bound [Burnetas and Katehakis, 1996]

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

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

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.

Bounded distributions

$$\mathcal{D}_B = \{\nu, \nu \text{ supported in } [0, B]\}$$

$$\mathcal{K}_{\inf}^{\mathcal{D}_B}(\nu; \mu') = \text{ non explicit, but computable}$$

A first (bad) algorithm

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

A first (bad) algorithm

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! $\nu_1 = \mathcal{B}(\mu_1), \nu_2 = \mathcal{B}(\mu_2)$, $\mu_1 > \mu_2$

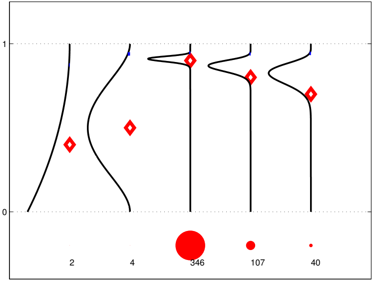
$$\mathbb{E}[N_2(T)] \geq (1 - \mu_1)\mu_2 \times (T - 1)$$

➜ **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 of 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*

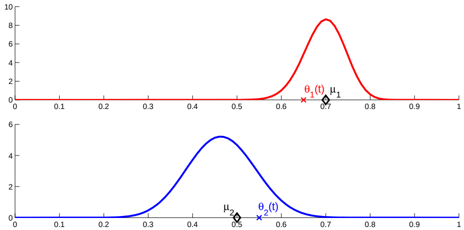
A Bayesian algorithm: Thompson Sampling

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

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

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 $\pi(0)$

$$\begin{cases} \forall a \in \{1..K\}, & \widetilde{\theta}_a(t) \sim \pi_a(t) \\ A_{t+1} = \underset{a=1\dots K}{\text{argmax}}\ \widetilde{\theta}_a(t). \end{cases}$$

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 $\nu_a = \mathcal{N}(\mu_a, \sigma^2)$

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

Theoretical guarantees

For different Single Parameter Exponential Families, TS with a conjugate prior satisfy the following:

Upper bound on sub-optimal selections

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

➜ matching the lower bound! TS is *asymptotically optimal*

[Kaufmann et al., 2012, Agrawal and Goyal, 2013, Korda et al., 2013]

Where does the KL come from?

1. the best arm (arm 1) has to be drawn a lot
2. probability of selecting a sub-optimal arm a:

$$\begin{aligned}
\mathbb{P}(A_t = a|\mathcal{F}_{t-1}) &= \mathbb{P}\left(\widetilde{\theta}_a(t) = \max_i \widetilde{\theta}_i(t)|\mathcal{F}_{t-1}\right) \\
&\simeq \mathbb{P}\left(\widetilde{\theta}_a(t) \geq \widetilde{\theta}_1(t)|\mathcal{F}_{t-1}\right) \\
&\simeq \mathbb{P}\left(\widetilde{\theta}_a(t) \geq \mu_1|\mathcal{F}_{t-1}\right)
\end{aligned}$$

Where does the KL come from?

1. the best arm (arm 1) has to be drawn a lot
2. probability of selecting a sub-optimal arm a:

$$\begin{aligned}\mathbb{P}(A_t = a|\mathcal{F}_{t-1}) &= \mathbb{P}\left(\widetilde{\theta}_a(t) = \max_i \widetilde{\theta}_i(t)|\mathcal{F}_{t-1}\right) \\ &\simeq \mathbb{P}\left(\widetilde{\theta}_a(t) \geq \widetilde{\theta}_1(t)|\mathcal{F}_{t-1}\right) \\ &\simeq \mathbb{P}\left(\widetilde{\theta}_a(t) \geq \mu_1|\mathcal{F}_{t-1}\right)\end{aligned}$$

For Gaussian bandits $\widetilde{\theta}_a(t)|\mathcal{F}_{t-1} = \mathcal{N}(\hat{\mu}_a(t), \sigma^2/N_a(t))$, thus

$$\begin{aligned}\mathbb{P}(A_t = a|\mathcal{F}_{t-1}) &\simeq \mathbb{P}\left(X \geq \frac{\sqrt{N_a(t)}(\mu_1 - \hat{\mu}_a(t))}{\sigma}\right) \\ &\leq \exp\left(-\frac{N_a(t)(\hat{\mu}_a(t) - \mu_1)^2}{2\sigma^2}\right)\end{aligned}$$

Where does the KL come from?

1. the best arm (arm 1) has to be drawn a lot
2. probability of selecting a sub-optimal arm a:

$$\begin{aligned}
\mathbb{P}(A_t = a|\mathcal{F}_{t-1}) &= \mathbb{P}\left(\widetilde{\theta}_a(t) = \max_i \widetilde{\theta}_i(t)|\mathcal{F}_{t-1}\right) \\
&\simeq \mathbb{P}\left(\widetilde{\theta}_a(t) \geq \widetilde{\theta}_1(t)|\mathcal{F}_{t-1}\right) \\
&\simeq \mathbb{P}\left(\widetilde{\theta}_a(t) \geq \mu_1|\mathcal{F}_{t-1}\right)
\end{aligned}$$

For Gaussian bandits $\widetilde{\theta}_a(t)|\mathcal{F}_{t-1} = \mathcal{N}(\hat{\mu}_a(t), \sigma^2/N_a(t))$, thus

$$\begin{aligned}
\mathbb{P}(A_t = a|\mathcal{F}_{t-1}) &\simeq \mathbb{P}\left(X \geq \frac{\sqrt{N_a(t)}(\mu_1 - \hat{\mu}_a(t))}{\sigma}\right) \\
&\leq \exp\left(-N_a(t)\mathrm{KL}(\hat{\mu}_a(t), \mu_1)\right)
\end{aligned}$$

Where does the KL come from?

1. the best arm (arm 1) has to be drawn a lot
2. probability of selecting a sub-optimal arm a:

$$\begin{aligned}\mathbb{P}(A_t = a|\mathcal{F}_{t-1}) &= \mathbb{P}\left(\widetilde{\theta}_a(t) = \max_i \widetilde{\theta}_i(t)|\mathcal{F}_{t-1}\right) \\ &\simeq \mathbb{P}\left(\widetilde{\theta}_a(t) \geq \widetilde{\theta}_1(t)|\mathcal{F}_{t-1}\right) \\ &\simeq \mathbb{P}\left(\widetilde{\theta}_a(t) \geq \mu_1|\mathcal{F}_{t-1}\right)\end{aligned}$$

For Gaussian bandits $\widetilde{\theta}_a(t)|\mathcal{F}_{t-1} = \mathcal{N}(\hat{\mu}_a(t), \sigma^2/N_a(t))$, thus

$$\begin{aligned}\mathbb{P}(A_t = a|\mathcal{F}_{t-1}) &\simeq \mathbb{P}\left(X \geq \frac{\sqrt{N_a(t)}(\mu_1 - \hat{\mu}_a(t))}{\sigma}\right) \\ &\leq \exp\left(-N_a(t)\mathrm{KL}(\hat{\mu}_a(t), \mu_1)\right)\end{aligned}$$

$$\mathbb{P}(A_t = a|\mathcal{F}_{t-1}) \leq \frac{1}{T} \quad \Rightarrow \quad N_a(t) \simeq \frac{\log(T)}{\mathrm{KL}(\hat{\mu}_a(t), \mu_1)}$$

Beyond Parametric Distributions

[Riou and Honda, 2020] : NPTS for Bounded distributions

$$\mathcal{D}_B = \{\nu : \nu \text{ is support in } [0, B]\}$$

Non Parametric Thompson Sampling

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

where

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

with

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

➜ TS is asymptotically optimal for bounded distributions!

Outline

Algorithm: made of three components:

➜ sampling rule: A_t (arm to explore)

➜ recommendation rule: $\hat{a}_t$ (current guess for the best arm)

➜ (optional) stopping rule τ (when do we stop exploring?)

- **Settings studied in the literature:**

Fixed-confidence	Fixed-budget	Anytime
input: error bound δ	**input:** budget T	
min. $\mathbb{E}[\tau]$ $\mathbb{P}(\hat{a}_\tau \neq a_\star) \leq \delta$	$\tau = T$ min. $\mathbb{P}(\hat{a}_T \neq a_\star)$	$\forall t \in \mathbb{N}$, min. $\mathbb{P}(\hat{a}_t \neq a_\star)$
[Even-Dar et al., 2006]	[Audibert et al., 2010]	[Bubeck et al., 2011]

Best Arm Identification

Algorithm: made of three components:

- ➜ sampling rule: A_t (arm to explore)
- ➜ recommendation rule: $\hat{a}_t$ (current guess for the best arm)
- ➜ (optional) stopping rule τ (when do we stop exploring?)

- **Settings studied in the literature:**

Fixed-confidence	Fixed-budget	Anytime
input: error bound δ	**input:** budget T	
min. $\mathbb{E}[\tau]$ $\mathbb{P}(\hat{a}_\tau \neq a_\star) \leq \delta$	$\tau = T$ min. $\mathbb{P}(\hat{a}_T \neq a_\star)$	$\forall t \in \mathbb{N}$, min. $\mathbb{E}[\mu_\star - \mu_{\hat{a}_t}]$
[Even-Dar et al., 2006]	[Audibert et al., 2010]	[Bubeck et al., 2011]

Can Thompson Sampling find the best arm?

$\hat{a}_T$: guess for the best arm after T samples.

Thompson Sampling selects a lot the best arm...

- idea (1): $\hat{a}_T = \arg\max_a N_a(T)$
- idea (2) : $\mathbb{P}(\hat{a}_T = a | \mathcal{F}_T) = \frac{N_a(T)}{T}$

Thompson Sampling + (2):

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

☺ the estimation error decays with T

Can Thompson Sampling find the best arm?

$\hat{a}_T$: guess for the best arm after T samples.

Thompson Sampling selects a lot the best arm...

- idea (1): $\hat{a}_T = \arg\max_a N_a(T)$
- idea (2) : $\mathbb{P}(\hat{a}_T = a | \mathcal{F}_T) = \frac{N_a(T)}{T}$

Thompson Sampling + (2):

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

☺ the estimation error decays with T

Uniform Sampling + Empirical Best Arm:

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

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

Can Thompson Sampling find the best arm?

$\hat{a}_T$: guess for the best arm after T samples.

Thompson Sampling selects a lot the best arm...

- idea (1): $\hat{a}_T = \arg\max_a N_a(T)$
- idea (2) : $\mathbb{P}(\hat{a}_T = a | \mathcal{F}_T) = \frac{N_a(T)}{T}$

Thompson Sampling + (2):

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

☺ the estimation error decays with T

Uniform Sampling + Empirical Best Arm:

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

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

Top Two Thompson Sampling

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

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.

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 $\boldsymbol{\nu} = (\nu_1, \dots, \nu_K) \in \mathcal{D}^K$,

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

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

General expression:

$$T^\star_\beta(\boldsymbol{\nu})^{-1} = \sup_{\substack{\boldsymbol{w} \in \triangle_K \\ w_{a_\star} = \beta}} \min_{a \neq a^\star} \underbrace{\inf_{\lambda_a \geq \lambda_{a_\star}} \left[w_{a_\star} \mathcal{K}^-_{\inf}(\nu_{a_\star}, \lambda_{a_\star}) + w_a \mathcal{K}^+_{\inf}(\nu_a, \lambda_a) \right]}_{\text{"transportation cost"}}.$$

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

Back to the parametric case: Gaussian bandits

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

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 (Generalized Likelihood Ratio) stopping rule

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

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

$$T_\beta^\star(\boldsymbol{\mu})^{-1} = \min_{a \neq a^\star} \frac{(\mu_{a_\star} - \mu_a)^2}{2\sigma^2\left(\frac{1}{w_{a_\star}^\star} + \frac{1}{w_a^\star}\right)}$$

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 (Generalized Likelihood Ratio) stopping rule

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

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

Theorem [Shang et al., 2020]

TTTS(β) is δ-correct and

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

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

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

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

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

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

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

In Top Two Thompson Sampling,

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

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

Liminations:

➜ re-sampling can be numerically costly

➜ do we need a posterior distribution?

Approximating Re-Sampling

Under the RS challenger,

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

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: select the mode from this distribution instead of sampling!

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

[Shang et al., 2020]

Another (non Bayesian) interpretation

Recall that TTTS was analyzed with

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

➜ another interpretation: C_t^{TC} minimizes the Empirical Transportation Cost (TC) featured in the stopping rule

Another (non Bayesian) interpretation

Recall that TTTS was analyzed with

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

- ➜ another interpretation: C_t^{TC} minimizes the Empirical Transportation Cost (TC) featured in the stopping rule
- ➜ could we use $B_T^{\mathrm{EB}} = \hat{a}_t^\star$, i.e. Empirical Best leader?

Asymptotically... yes!

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., 2023a]

But exploration is nice for finite-time performance

TS-TC

$$B_t \sim \arg\max_{a \in [K]} \widetilde{\theta}_a(t) \quad \widetilde{\theta}(t) \sim \Pi_t$$

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

EB-TCI

$$B_t = \arg\max_{a \in [K]} \hat{\mu}_a(t)$$

$$C_t = \arg\min_{a \neq B_t} \left[\frac{(\hat{\mu}_{B_t}(t) - \hat{\mu}_a(t))_+^2}{2\sigma^2 \left(\frac{1}{N_{B_t}(t)} + \frac{1}{N_a(t)} \right)} + \log N_a(t) \right]$$

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

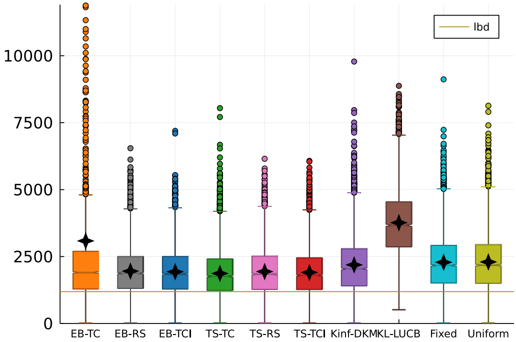
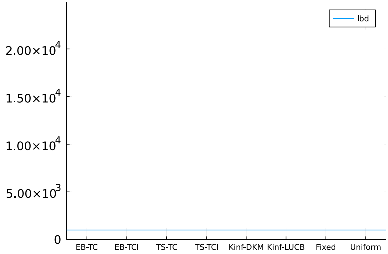


Figure: Empirical sample complexity averaged over 5000 random (Bernoulli) instances with $K = 8$ and $\Delta_{\min} \geq 0.01$.

Numerical experiments

arm = planting date / observation = yield

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



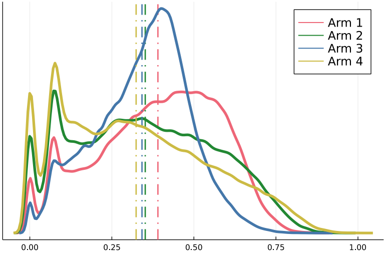


Figure: Empirical stopping time (a) on scaled DSSAT instances with their density and mean (b). Lower bound is $T^\star(\boldsymbol{\nu}) \ln(1/\delta)$.

Numerical experiments

arm = planting date / observation = yield

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

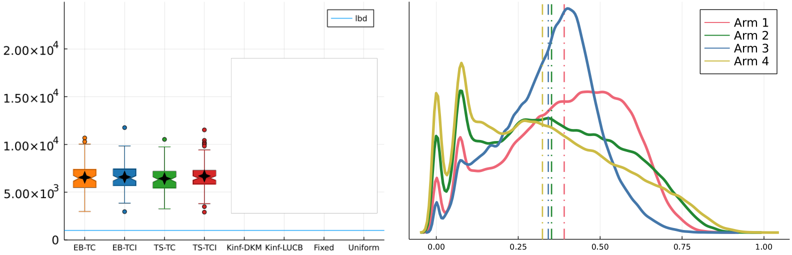


Figure: Empirical stopping time (a) on scaled DSSAT instances with their density and mean (b). Lower bound is $T^\star(\boldsymbol{\nu})\ln(1/\delta)$.

Experiments: Bounded distributions

arm = planting date / observation = yield

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

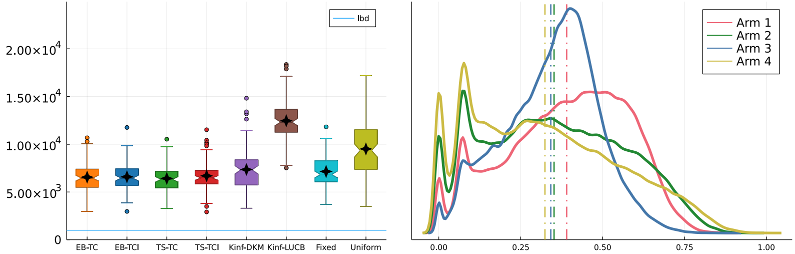


Figure: Empirical stopping time (a) on scaled DSSAT instances with their density and mean (b). Lower bound is $T^\star(\boldsymbol{\nu}) \ln(1/\delta)$.

Top Two algorithms beyond Fixed Confidence

EB-TC$_{\varepsilon_0}$

$$B_t = \arg\max_{a \in [K]} \hat{\mu}_a(t)$$

$$C_t = \arg\min_{a \neq B_t} \left[\frac{\hat{\mu}_{B_t}(t) - \hat{\mu}_a(t) + \varepsilon_0}{\sqrt{\frac{1}{N_{B_t}(t)} + \frac{1}{N_a(t)}}} \right]$$

[Jourdan et al., 2023b]

- motivated by the lower bound for (ε_0, δ)-PAC identification
- can be used for (ε, δ)-PAC identification[1] for $\varepsilon \neq \epsilon_0$
- first guarantees in the anytime setting...

[1] $\mathbb{P}\left(\mu_{\hat{a}_\tau} > \mu_\star - \varepsilon\right) \geq 1 - \delta$

Top Two algorithms Beyond Fixed Confidence

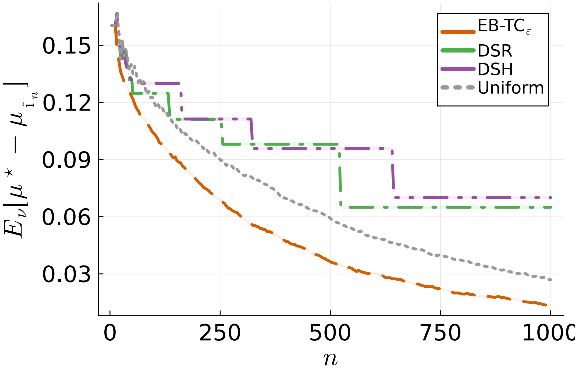


Figure: Simple regret as a function of time on an instance $\mu \in \{0.4, 0.6\}^{10}$ with 2 best arms

(... but the theory is just saying that the algorithm is not too much worse than uniform sampling...)

Conclusion

Thompson Sampling for maximizing rewards:

- is asymptotically optimal for simple parametric distributions
- can be extended to some non-parametric settings

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:

- Understand better the good anytime performance
- Top Two for more complex pure exploration problems?

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

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.

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. (2023a).
Dealing with unknown variances in best-arm identification.
In *Algorithmic Learning Theory (ALT)*.

Jourdan, M., Degenne, R., and Kaufmann, E. (2023b).
An ε-best-arm identification algorithm for fixed confidence and beyond.
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*.

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

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)*.

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