Exploration non paramétrique dans les modèles de bandit

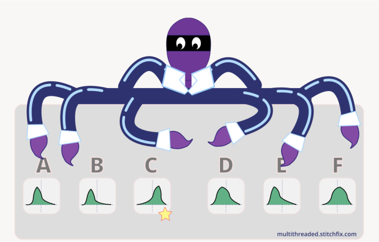
Emilie Kaufmann,

basé sur une collaboration avec
Dorian Baudry et Odalric-Ambrym Maillard

Séminaire du LPSM, septembre 2022

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 $X_t \sim \nu_{A_t}$



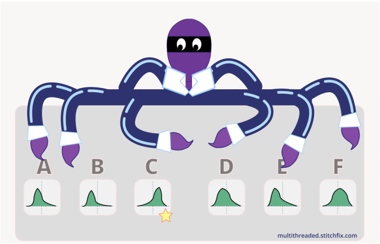
Objective: find a sequential sampling strategy $\mathcal{A} = (A_t)$ that maximizes the sum of rewards $\Leftrightarrow$ minimize the *regret*

$$\mathcal{R}_T(\mathcal{A}) = \mu^\star T - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$$

[Robbins, 1952, Lattimore and Szepesvari, 2019]

The stochastic Multi Armed Bandit (MAB) model

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



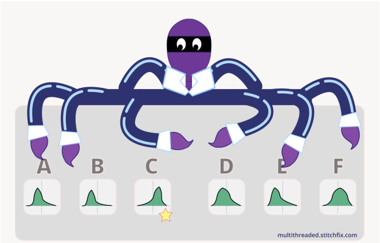
Objective: find a sequential sampling strategy $\mathcal{A} = (A_t)$ that maximizes the sum of rewards $\Leftrightarrow$ minimize the *regret*

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

[Robbins, 1952, Lattimore and Szepesvari, 2019]

The stochastic Multi Armed Bandit (MAB) model

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



Objective: find a sequential sampling strategy $\mathcal{A} = (A_t)$ that maximizes the sum of rewards $\Leftrightarrow$ minimize the *regret*

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

[Robbins, 1952, Lattimore and Szepesvari, 2019]

Examples

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

- movie recommendation → reward: rating (multinomial)

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

Goal: design algorithms that use as little knowledge about the rewards distributions as possible

Outline

(Don't) Follow The Leader

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

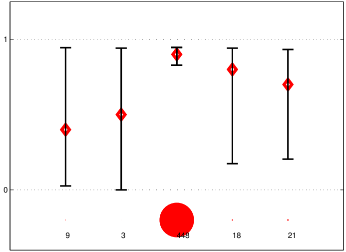
$$\mathbb{E}[N_2(T)] \geq (1 - \mu_1)\mu_2 \times (T - 1)$$

$$\Rightarrow \text{linear regret}$$

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

Smarter algorithms: Two dominant families

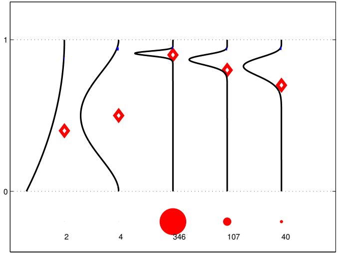
Upper Confidence Bound (UCB)



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

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

Thompson Sampling (TS)

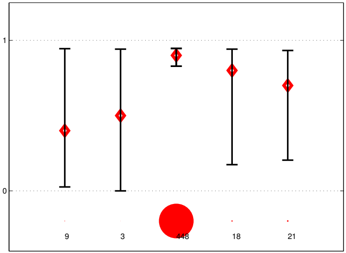


$$A_{t+1} = \underset{a\in[K]}{\text{argmax}}\ \tilde{\mu}_a(t)$$

where $\tilde{\mu}_a(t)$ is a sample from a posterior distribution on μ_a

Smarter algorithms: Two dominant families

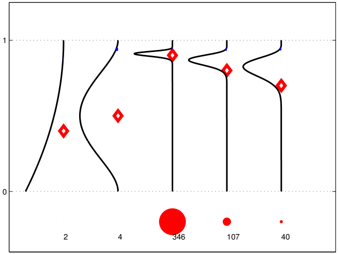
Upper Confidence Bound (UCB)



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

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

Thompson Sampling (TS)



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

where $\tilde{\mu}_a(t)$ is a sample from a posterior distribution on μ_a

➜ both approaches can be tuned to achieve *optimality*

(Problem dependent, asymptotic) optimality

$$\mathcal{R}_T(\mathcal{A}) = \mathbb{E}\left[\sum_{t=1}^{T}(\mu_\star - \mu_{A_t})\right] = \sum_{a:\mu_a<\mu_\star}(\mu_\star - \mu_a)\mathbb{E}[N_a(T)]$$

where $N_a(T)$ is the number of selections of arm a up to round T.

For each a, let $\mathcal{D}_a$ be a family of probability distributions.

Lower bound [Lai and Robbins, 1985, Burnetas and Katehakis, 1996]

Under an algorithm achieving small regret for any bandit model $\nu \in \mathcal{D}_1 \times \cdots \times \mathcal{D}_K$, it holds that

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

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.

Matching the lower bound

If $\mathcal{D}$ is a one-dimensional exponential family

$$\mathcal{K}_{\text{inf}}^{\mathcal{D}}(\nu_a, \mu_\star) = \mathrm{kl}(\mu_a, \mu_\star)$$

where $\mathrm{kl}(\mu, \mu') = \mathrm{KL}(\nu_\mu, \nu_{\mu'})$ with $\nu_\mu \in \mathcal{D}$ the unique distribution in $\mathcal{D}$ that has mean μ.

Examples: Bernoulli, Gaussian with known variance σ^2, Poisson...

- kl-UCB [Cappé et al., 2013] uses the $\mathrm{kl}(\cdot, \cdot)$ divergence
- Thompson Sampling using a conjuguate prior

are both matching the lower bound.

➜ can we find a single algorithm that is simultaneously asymptotically optimal for different classes of distributions?

Matching the lower bound

If $\mathcal{D}$ is a one-dimensional exponential family

$$\mathcal{K}_{\text{inf}}^{\mathcal{D}}(\nu_a, \mu_\star) = \frac{(\mu_a - \mu_\star)^2}{2\sigma^2}$$

where $\mathrm{kl}(\mu, \mu') = \mathrm{KL}(\nu_\mu, \nu_{\mu'})$ with $\nu_\mu \in \mathcal{D}$ the unique distribution in $\mathcal{D}$ that has mean μ.

Examples: Bernoulli, Gaussian with known variance σ^2, Poisson...

- kl-UCB [Cappé et al., 2013] uses the $\mathrm{kl}(\cdot, \cdot)$ divergence
- Thompson Sampling using a conjugate prior

are both matching the lower bound.

➜ can we find a single algorithm that is simultaneously asymptotically optimal for different classes of distributions?

A first non-parameteric idea: re-sampling

Non-Parametric Bootstrap

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

where $\tilde{\mu}_a(t)$ average of $N_a(t)$ samples drawn at random with replacement in the history $\mathcal{H}_a(t) = \{Y_{a,1}, \dots, Y_{a,N_a(t)}\}$.

- [Kveton et al., 2019]: vanilla non-parametric bootstrap can have linear regret, a fix adding fake rewards in the history
- logarithmic regret for bounded distributions (*not* optimal)

A first non-parameteric idea: re-sampling

Non-Parametric Bootstrap

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

where $\tilde{\mu}_a(t)$ average of $N_a(t)$ samples drawn at random with replacement in the history $\mathcal{H}_a(t) = \{Y_{a,1}, \dots, Y_{a,N_a(t)}\}$.

- [Kveton et al., 2019]: vanilla non-parametric bootstrap can have linear regret, a fix adding fake rewards in the history
- logarithmic regret for bounded distributions (*not* optimal)

In order to be asymptotically optimal, for potentially unbounded distributions, we rely instead on sub-sampling

[Baransi et al., 2014, Chan, 2020]

Outline

1. Optimal solutions and their limitation
2. Sub-Sampling Duelling Algorithms (SDA)
3. Analysis of RB-SDA
4. A risk-averse non-parametric algorithm

Sub-sampling Duelling Algorithms

A *round-based* approach

1. Find the *leader*: arm with largest number of observations
2. Organize $K - 1$ duels: *leader* vs *challengers*.
3. Draw a set of arms: *winning challengers* xor *leader*.

A *round-based* approach

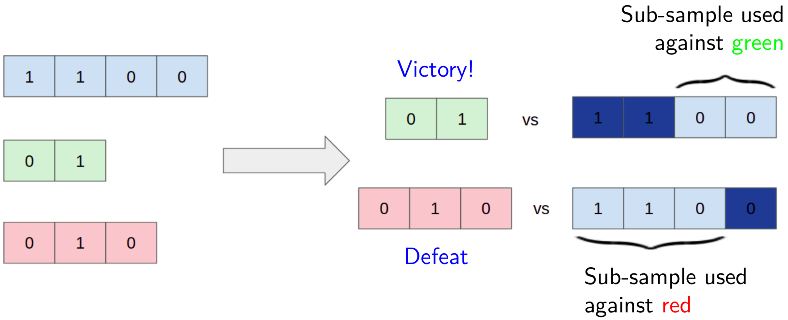
1. Find the *leader*: arm with largest number of observations
2. Organize $K - 1$ duels: *leader vs challengers*.
3. Draw a set of arms: *winning challengers* xor *leader*.

How do duels work?

Idea: a *fair comparison* of two arms with different history size

- challenger: compute $\hat{\mu}_c$, the empirical mean
- leader: compute $\tilde{\mu}_\ell$, the mean of a *sub-sample* of the same size as the history of the challenger.
- challenger wins if $\hat{\mu}_c \geq \tilde{\mu}_\ell$

Illustration of a round



In this example the leader is *blue*: *green* wins against *blue*, *red* loses
$\Rightarrow$ only *green* is drawn at the end of the round.

Possible Sub-Sampling Schemes

Input of SDA: how to sub-sample n elements from N?

- Sampling Without Replacement (SW-SDA): pick a random subset of size n in $[1, N]$

  (as in BESA [Baransi et al. 14], analyzed for 2 arms)

- Random-Block Sampling (RB-SDA): return a block of size n starting from random $n_0 \sim \mathcal{U}([1, N-n])$

7.6	-4	0.7	1.4	3.1	0.1	-1.2

- Last Block Sampling (LB-SDA): return $\{N-n, \dots, N\}$

7.6	-4	0.7	1.4	3.1	0.1	-1.2

- SSMC [Chan 20] uses data-dependent sub-sampling

Outline

Regret of SDA algorithms

SDA algorithms are round-based

- $\mathcal{A}_r$: set of arms that are sampled in round r
- r_T (random) number of rounds before T samples are collected

$\widetilde{N}_a(r) = \sum_{s=1}^{r} \mathbb{1}(a \in \mathcal{A}_s)$: number of selections of a in r rounds

$$
\begin{aligned}
\mathcal{R}_T(\mathcal{A}) &= \sum_{a=1}^{K} (\mu_\star - \mu_a)\mathbb{E}[N_a(T)] \\
&\leq \sum_{a=1}^{K} (\mu_\star - \mu_a)\mathbb{E}\left[\widetilde{N}_a(r_T)\right] \\
&\leq \sum_{a=1}^{K} (\mu_\star - \mu_a)\mathbb{E}\left[\widetilde{N}_a(T)\right]
\end{aligned}
$$

First ingredient: Concentration

Definition (Block Sampler)

A *block sampler* outputs a sequence of **consecutive observations** in the rewards history.

$\hookrightarrow$ Random Block and Last Block are block samplers, not SWR.

- $Y_{a,n}$: n-th observation from arm a
- $\overline{Y}_{a,\mathcal{S}} = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} Y_{a,i}$ for a subset $\mathcal{S}$
- $\mathcal{S}^s_{a,b} \subseteq [N_a(s)]$ sub-sample used in round s for the leader a against the challenger b, $\left|\mathcal{S}^s_{a,b}\right| = N_b(s)$

Lemma (concentration of a sub-sample)

Under a block sampler, for any $\mu_a < \xi < \mu_b$,

$$\sum_{s=1}^{r} \mathbb{P}\Big(\bar{Y}_{a,\mathcal{S}^s_{a,b}} \geq \bar{Y}_{b,N_b(s)}, n_0 \leq N_b(s) \leq N_a(s)\Big) \leq \sum_{j=n_0}^{r} \mathbb{P}\left(\bar{Y}_{a,j} \geq \xi\right) + r\sum_{j=n_0}^{r} \mathbb{P}\left(\bar{Y}_{b,j} \leq \xi\right)$$

First ingredient: Concentration

Assumption 1: (*arm concentration*)

$$\begin{aligned} \forall x > \mu_a, \quad \mathbb{P}\left(\bar{Y}_{a,n} \geq x\right) &\leq e^{-nI_a(x)} \\ \forall x < \mu_a, \quad \mathbb{P}\left(\bar{Y}_{a,n} \leq x\right) &\leq e^{-nI_a(x)} . \end{aligned}$$

for some rate function $I_a(x)$ (1-d exp. families: $I_a(x) = \mathrm{kl}(x, \mu_a)$)

Lemma (for SDA using a block sampler)

Under Assumption 1, for every $\varepsilon > 0$, there exists a constant $C_k(\boldsymbol{\nu}, \epsilon)$ with $\boldsymbol{\nu} = (\nu_1, \dots, \nu_k)$ such that

$$\mathbb{E}\left[\widetilde{N}_a(T)\right] \leq \frac{1+\epsilon}{I_1(\mu_a)} \log(T) + 32 \sum_{r=1}^{T} \mathbb{P}\left(\widetilde{N}_1(r) \leq (\log(r))^2\right) + C_a(\boldsymbol{\nu}, \epsilon)$$

Proof: exploits only concentration (and how the algorithm works)

First ingredient: Concentration

Assumption 1: (*arm concentration*)

$$\forall x > \mu_a, \quad \mathbb{P}\left(\bar{Y}_{a,n} \geq x\right) \quad \leq \quad e^{-nI_a(x)}$$
$$\forall x < \mu_a, \quad \mathbb{P}\left(\bar{Y}_{a,n} \leq x\right) \quad \leq \quad e^{-nI_a(x)} \; .$$

for some rate function $I_a(x)$ (1-d exp. families: $I_a(x) = \mathrm{kl}(x, \mu_a)$)

Lemma (for SDA using a block sampler)

Under Assumption 1, for every $\varepsilon > 0$, there exists a constant $C_k(\boldsymbol{\nu}, \epsilon)$ with $\boldsymbol{\nu} = (\nu_1, \dots, \nu_k)$ such that

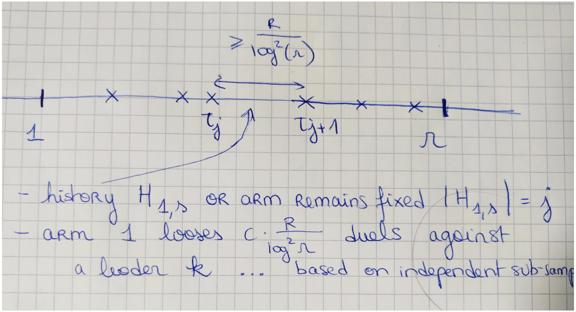
$$\mathbb{E}\left[\widetilde{N}_a(T)\right] \leq \frac{1+\epsilon}{I_1(\mu_a)} \log(T) + 32 \sum_{r=1}^{T} \mathbb{P}\left(\widetilde{N}_1(r) \leq (\log(r))^2\right) + C_a(\boldsymbol{\nu}, \epsilon)$$

Proof: exploits only concentration (and how the algorithm works)

$$\left(N_1(r) \leq \log^2(r)\right)$$

$$\subseteq \bigcup_{j=0}^{\lfloor \log^2(r) \rfloor} \left(\tau_{j+1} - \tau_j \geq \frac{r}{\log^2(r)}\right) \bigcap \left\{\text{arm 1 is not the leader}\right\}$$

τ_j: instant in of the j-th selection of arm 1



Two extra ingredients

To upper bound $\sum_{r=1}^{T} \mathbb{P}\left(N_1(r) \leq (\log(r))^2\right)$, we further need:

① **Diversity**: the sub-sampler produces a variety of *independent* sub-samples when being called a lot of times

$X_{m,H,j}$:= number of mutually non-overlapping sets when we draw m sub-samples of size j in a history of size H.

Under Random Block sampling,

$$\sum_{r=1}^{T} \sum_{j=1}^{(\log r)^2} \mathbb{P}\left(X_{N_r,N_r,j} < \gamma \frac{r}{(\log r)^2}\right) = o(\log T)\,.$$

for $N_r = O(r/\log^2(r))$ and some $\gamma \in (0,1)$

Two extra ingredients

To upper bound $\sum_{r=1}^{T} \mathbb{P}\left(N_1(r) \leq (\log(r))^2\right)$, we further need:

② a **Balance condition**: the optimal arm (arm 1) is not likely to loose many (M) duels based on *independent* sub-samples of a sub-optimal arm (arm a)

Balance function of arm $a \neq 1$:

$$\begin{aligned} \alpha_a(M,j) &:= \mathbb{E}_{X \sim \nu_{1,j}}\left[(1 - F_{\nu_{a,j}}(X))^M\right] \\ &= \mathbb{P}\left(\bigcap_{m=1}^{M} \left(\overline{Y}_{1,j} < \overline{Y}_{a,\mathcal{S}_m}\right)\right) \quad |\mathcal{S}_m| = j, \mathcal{S}_m \cap \mathcal{S}_{m'} = \emptyset \end{aligned}$$

The balance condition for arm a is

$$\forall \beta \in (0,1), \ \sum_{r=1}^{T} \sum_{j=g_r}^{\lfloor (\log r)^2 \rfloor} \alpha_a\left(\left\lfloor \beta \frac{r}{(\log r)^2} \right\rfloor, j\right) = o(\log T)$$

g_r: amount of forced exploration added to the algorithm

Final results

General Theorem [Baudry et al., 2020]

If all arms satisfy Assumption 1 and the sub-optimal arms satisfy the balance condition, RB-SDA satisfies, for all sub-optimal arm a,

$$\mathbb{E}\left[\widetilde{N}_a(T)\right] \leq \frac{1+\varepsilon}{I_1(\mu_a)} \log(T) + o_\varepsilon(\log T) \,.$$

One-parameter exponential families:

- satisfy Assumption 1 and $I_1(x) = \mathrm{kl}(x, \mu_1)$
- satisfy the balance condition with $g_r = \sqrt{\log(r)}$ (and $g_r = 1$ for Bernoulli, Gaussian and Poisson distributions)

➜ RB-SDA is asymptotically optimal for *different* exponential family bandit models (possibly with unbounded support)

Works very well in practice!

Average Regret on $N = 10000$ random instances with $K = 10$

- **Bernoulli arms**

T	TS	IMED	PHE	SSMC	RB-SDA
100	**13.8**	**15.1**	16.7	16.5	**14.8**
1000	**27.8**	**31.9**	39.5	34.2	**31.8**
10000	**45.8**	**51.2**	72.3	55.0	**51.1**
20000	**52.2**	**57.6**	85.6	61.9	**57.7**

- **Gaussian arms**

T	TS	IMED	SSMC	RB-SDA
100	41.2	45.1	40.6	**38.1**
1000	76.4	82.1	76.2	**70.4**
10000	118.5	124.0	120.1	**111.8**
20000	132.6	138.1	135.1	**125.7**

more experiments in [Baudry et al. 20]

RB-SDA has logarithmic regret for any class of distributions that concentrate and satisfy the balance condition.

(same result for LB-SDA, see [Baudry et al., 2021b])

Sufficient condition: if there exists x_0 and $C < 1$ such that

$$\forall x \leq x_0, \quad f_1(x) < C f_a(x),$$

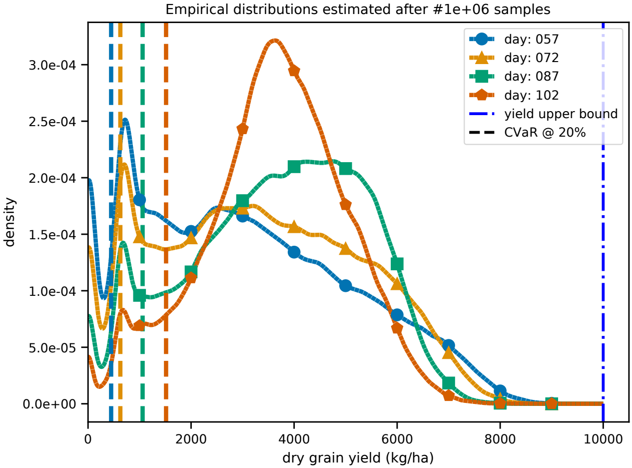
the balance condition is satisfied with $g_r = \sqrt{\log(r)}$

- 👍 interpretation: SDA works when the best arm has the "lightest left tail"
- 👎 this condition does not always hold for Gaussian with unknown variances, or multinomial distributions

Outline

1. Optimal solutions and their limitation
2. Sub-Sampling Duelling Algorithms (SDA)
3. Analysis of RB-SDA
4. A risk-averse non-parametric algorithm

Motivation: recommending planting dates to farmers



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

A risk-averse bandit problem

Specifics of our application:

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

→ quality of an arm is 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 $n_\alpha = \inf\left\{n : \sum_{i=1}^{n} p_i x_i \geq \alpha\right\}$.

→ 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 is 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

A risk-averse bandit problem

Specifics of our application:

- ➜ **bounded** distributions, with known upper bound B
- ➜ quality of an arm is 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_α			
	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

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

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

Lower bound [Baudry et al., 2021a]

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

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

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: $\nu_a \in \mathcal{B}_a = \{\text{distributions supported in } [0, B_a]\}$.

➜ We propose an index policy, **B-CVTS**:

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

Index of arm a after t rounds

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

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

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

$\alpha = 1 \rightarrow$ Non Parametric Thompson Sampling [Riou and Honda 20]

B-CVTS is asymptotically optimal for bounded distributions.

Theorem [Baudry et al., 2021a]

On an instance ν such that $\nu \in \mathcal{B}_1 \times \cdots \times \mathcal{B}_K$, we have

$$\mathcal{R}_T(\text{B-CVTS}) \leq \sum_{a: c_a^\alpha < c_\star^\alpha} \frac{(c_\star^\alpha - c_a^\alpha) \log T}{\mathcal{K}_{\inf}^{\alpha, \mathcal{B}_a}(\nu_a, c_1^\alpha)} + o(\log T) .$$

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

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

where

- $\mathcal{D}_n$ is a $\text{Dir}(1, \dots, 1)$ 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 $\hat{F}_{a,t}$

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

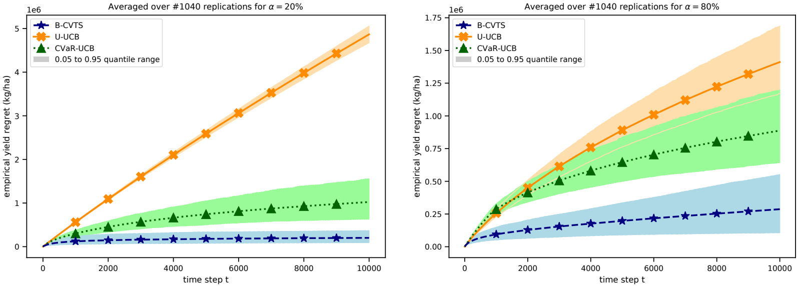
- CVaR-UCB: [Tamkin et al., 2020] buids an optimistic cdf $\overline{F}_{a,t}$

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

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)

Conclusion

Two non-parameteric exploration methods that can be good alternative to the standard UCB or Thompson Sampling:

- for bounded rewards, Non Parametric Thompson Sampling is optimal and can be naturally extended to tackle risk aversion
- Subsampling Duelling Algorithms can be simultaneously optimal in several bounded and unbounded parametric families
- ... but do not work for "any" distributions

Follow-up work:

- duelling with median-of-means instead of empirical means can make SDA work for heavy tailed distributions [Baudry et al., 2022]
- NPTS can be also be useful for pure exploration [Jourdan et al., 2022]

Baransi, A., Maillard, O., and Mannor, S. (2014).
Sub-sampling for multi-armed bandits.
In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML / PKDD*.

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

Baudry, D., Kaufmann, E., and Maillard, O.-A. (2020).
Sub-sampling for Efficient Non-Parametric Bandit Exploration.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Baudry, D., Russac, Y., and Cappé, O. (2021b).
On limited-memory subsampling strategies for bandits.
In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Baudry, D., Russac, Y., and Kaufmann, E. (2022).
Efficient algorithms for extreme bandits.
In *AISTATS*.

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

Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. (2013).
Kullback-Leibler upper confidence bounds for optimal sequential allocation.
Annals of Statistics, 41(3):1516–1541.

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*.

Chan, H. P. (2020).
The multi-armed bandit problem: An efficient nonparametric solution.
The Annals of Statistics, 48(1).

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)*.

Kveton, B., Szepesvári, C., Ghavamzadeh, M., and Boutilier, C. (2019).
Perturbed-history exploration in stochastic multi-armed bandits.
In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*.

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.

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

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.*