Non-Parametric Algorithms for Multi-Armed Bandits

Dorian Baudry

CNRS & Université de Lille & Inria Lille – Nord Europe, France

December 5, 2022

Table of Contents

Motivation: learning problem in agriculture

Objective: Help a community of farmers improve their crop-management practices under challenging conditions.

- Grow maize in a rainfed context and fixed soil conditions.
- Crop-management practice $:=$ set of rules to follow by the farmer (e.g planting date, fertilization,. . .)

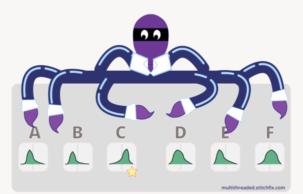
We propose to test **a selected number of policies designed by experts**

Figure: Maize Field in Ghana

Theoretical framework: Multi-Armed Bandits

- K unknown reward distributions $(\nu_1, \ldots, \nu_K)$ called arms.
- At each time t a learner selects an arm and observe a (random) reward.
- **Objective**: maximize the expected sum of rewards.

  $\hookrightarrow$ Exploration/Exploitation trade-off.



Regret and basic notations

Maximizing the expected sum of rewards $\equiv$ minimizing the *regret*.

Consider distributions $(\nu_1, \dots, \nu_K)$ of means $(\mu_1, \dots, \mu_K)$, and $\mu^\star = \max \mu_k$.

The **regret** at round T is

$$\mathcal{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} (\mu^\star - \mu_{A_t})\right] = \sum_{k=1}^{K} \Delta_k \mathbb{E}[N_k(T)] \,,$$

- $\Delta_k = \mu^\star - \mu_k$: "sub-optimality gap" of arm k.
- $N_k(T) = \sum_{t=1} \mathbb{1}(A_t = k)$: Number of selections of arm k.

$\hookrightarrow$ in the presentation we assume that arm 1 is the best.

Objective

- [Burnetas and Katehakis, 1996]: if the arms come from the family of distributions $\mathcal{F}$, for each sub-optimal arm k

$$\text{liminf}_{T\to+\infty} \frac{\mathbb{E}[N_k(T)]}{\log(T)} \geq \frac{1}{C^{\mathcal{F}}(\nu_k, \nu_1)} \, ,$$

  for some function $C^{\mathcal{F}}$.

- Objective:
  1. achieve **logarithmic** regret: $\mathbb{E}[N_k(T)] = \mathcal{O}(\log(T))$.
  2. If possible, match the **optimal** constant:

$$\mathbb{E}[N_k(T)] \leq \frac{\log(T)}{C^{\mathcal{F}}(\nu_k, \nu_1)} + o(\log(T)) \, .$$

Back to agriculture: typical crop yield distributions

We use the Decision Support Systems for Agro-Technological Transfer (DSSAT) simulator [Hoogenboom et al., 2019] to test algorithms *in silico* in a "realistic" environment.

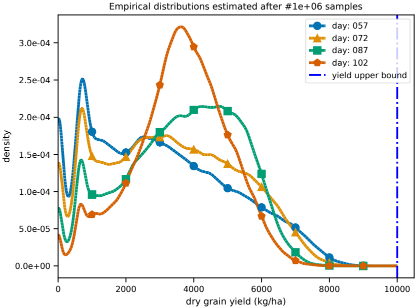


Figure: Yield distribution for different planting dates from the DSSAT simulator

- Reward = **Crop Yield**.
- No simple parametric model for the distributions.

$\hookrightarrow$ We need to design non-parametric algorithms.

Some existing algorithms

- Upper Confidence Bound (UCB)
- Thompson Sampling (TS)
- Index Minimized Empirical Divergence (IMED)

All these methods require some ***knowledge*** on the distributions.
The best algorithms extensively use it (prior/posterior, KL) to be **optimal**

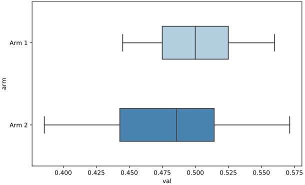


Figure: $5-95\%$ confidence intervals for empirical means, Bernoulli distrib., $(p_1 = 0.5, N_1 = 200,\ p_2 = 0.48, N_2 = 60)$

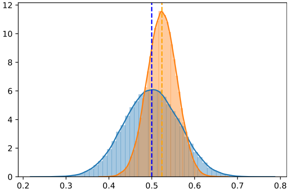


Figure: Densities of two Beta distrib.: $\text{Beta}(30, 30)$ and $\text{Beta}(110, 100)$

Non-exhaustive list of (optimal) bandit algorithms

Algorithm	Scope for optimality	Algorithm parameters
kl-UCB[1] IMED[2] Thompson Sampling[3]	Single Parameter Exponential Family (SPEF) $(\nu_\theta)_{\theta\in\Theta}$	$\mathrm{KL}(\nu_\theta, \nu_{\theta'})$ $\mathrm{KL}(\nu_\theta, \nu_{\theta'})$ Prior/Posterior
KL-UCB[1] IMED[2] Non-Parametric TS[4]	$\mathrm{Supp}(\nu) \subset [b, B]$	Upper bound B

1. [Cappé et al., 2013], 2. [Honda and Takemura, 2015],
3. see e.g [Kaufmann et al., 2012], 4. [Riou and Honda, 2020].

Contributions

Sub-Sampling Dueling Algorithms:

- ***Sub-Sampling Algorithms for Efficient Non-Parametric Bandit Exploration*** (Neurips 2020). DB, Emilie Kaufmann and Odalric-Ambrym Maillard.
- *On Limited-Memory Subsampling Strategies for Bandits* (ICML 2021). DB, Yoan Russac and Olivier Cappé.
- *Efficient Algorithms for Extreme Bandits* (AISTATS 2022). DB, Yoan Russac and Emilie Kaufmann.

Contributions

Non Parametric TS / Dirichlet Sampling:

- ***Optimal Thompson Sampling strategies for support-aware CVaR bandits*** (ICML 2021). DB, Romain Gautron, Emilie Kaufmann and Odalric-Ambrym Maillard.
- From Optimality to Robustness: Dirichlet Sampling Strategies in Stochastic Bandits (Neurips 2021). DB, Patrick Saux and Odalric-Ambrym Maillard.
- Top-Two algorithms revisited (Neurips 2022). Marc Jourdan, Rémy Degenne, DB, Rianne de Heide and Emilie Kaufmann.

Outline

Table of Contents

Why Sub-Sampling?

Simple strategy: Follow The Leader (FTL): $A_t = \text{argmax}\, \widehat{\mu}_k(t)$.

$\hookrightarrow$ bad scenario can happen with fixed probability $\Rightarrow$ linear regret.

Example:

1. Best arm collects a few bad samples $\Rightarrow$ mean under-estimated
2. Another arm pulled a lot $\Rightarrow$ mean concentrates
3. Best arm never pulled again

Core Idea: Comparing the means of sub-samples of the **same size** is a "fair" comparison between two arms!

Fair comparisons: Sub-sampling Dueling Algorithms (SDA)

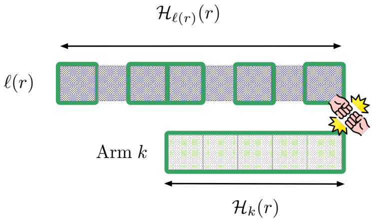
A **round-based** approach [Chan, 2020]:

1. Choose a *leader*: arm with largest number of observations!
2. Perform $K-1$ *duels*: *leader* vs each *challenger*.
3. Draw a set of arms: *winning challengers* (if any) or *leader* (if none).

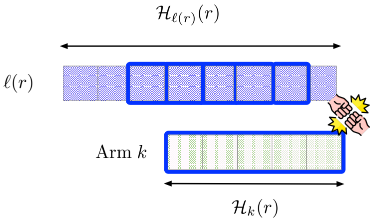
Duel

- Challenger → **empirical mean** $\widehat{\mu}_{k,N_k}$ (full sample size N_k).
- Leader → **mean** $\widehat{\mu}_{\ell,S(N_k,N_\ell)}$ of a ***subsample*** $S(N_k, N_\ell)$ of size N_k from its history.
- Winner: k if $\widehat{\mu}_{k,n} \geq \widehat{\mu}_{\ell,S(N_k,N_\ell)}$, ℓ otherwise.

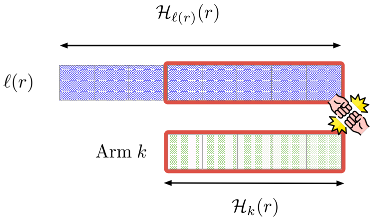
Some sub-sampling algorithms



Sampling without Replacement



Random Block Sampling



Last Block Sampling

Inspirations from the Literature

Keystones

- **Best Empirical Sample Average (BESA)** [Baransi et al., 2014]:
  - Tournament: arms eliminated successively.
  - Sampling Without Replacement (SWR).
- **Sub-Sample Mean Comparison (SSMC)** [Chan, 2020]:
  - Round-based approach ⇒ inspired SDA.
  - Sub-sampling: worst sequence of consecutive observations,

$$\inf_{n \in [1, N-m+1]} \left\{ \bar{Y}_{n:n+m-1} = \frac{1}{m} \sum_{i=n}^{n+m-1} Y_i \right\} .$$

Pros and Cons of BESA and SSMC

BESA:

- \+ Sub-Sampling independent from the history of rewards.
- \+ Works very well in practice for $K = 2$ and usual SPEF distributions.
- \- Tournament does not generalize well the duel principle.

SSMC:

- \+ Leader vs Challenger is more convenient than tournament.
- \- Sub-sampling can be costly and harder to generalize.

$\hookrightarrow$ SDA combines leader vs challenger duels and a reward-independent sub-sampling algorithm, and we introduce novel elements of analysis.

First theoretical guarantees

Assumption 1 (A1): For each arm k, the distributions ν_k (of mean μ_k) admits a good rate function I_k:

$$\forall x > \mu_k, \quad \mathbb{P}\left(\widehat{\mu}_{k,n} \geq x\right) \leq e^{-nI_k(x)} ,$$
$$\forall x < \mu_k, \quad \mathbb{P}\left(\widehat{\mu}_{k,n} \leq x\right) \leq e^{-nI_k(x)} ,$$

$\hookrightarrow$ Satisfied if $\mathbb{E}\left[e^{\lambda|X|}\right] < +\infty$ for some $\lambda > 0$:= light-tailed distributions.

Assumption 2 (A2): The sub-sampling algorithm is a *Block Sampler*

$\hookrightarrow$ e.g Random Block and Last Block.

First theoretical guarantees

Lemma (First upper bound)

Consider ν a bandit problem and SP a sampler satisfying resp. (A1) and (A2). Then, under SP-SDA any sub-optimal arm k satisfies

$$\mathbb{E}[N_k(T)] \leq \frac{1+\epsilon}{I_1(\mu_k)} \log(T) + C_k(\nu, \epsilon)$$

First theoretical guarantees

Lemma (First upper bound)

Consider ν a bandit problem and SP a sampler satisfying resp. (A1) and (A2). Then, under SP-SDA any sub-optimal arm k satisfies

$$\mathbb{E}[N_k(T)] \leq \frac{1+\epsilon}{I_1(\mu_k)} \log(T) + C_k(\nu,\epsilon) + 9\sum_{r=1}^{T} \mathbb{P}\left(N_1(r) \leq C_1 \log(r)\right) \ ,$$

where $C_k(\nu,\epsilon)$ and C_1 are both problem-dependent constants.

Key observation: Under (A1) and (A2), we only need to show that the best arm is **sufficiently explored**.

Ensuring sufficient exploration of the best arm

Two ingredients for exploration under SDA:

1. The sampler provides many *diverse* sub-samples.
2. If it plays many "diverse" duels, the best arm is likely to be pulled.

Key Result: RB-SDA and LB-SDA both provide a sufficient diversity of sub-samples.

$\hookrightarrow$ their theoretical guarantees only depend on the **family of distributions** considered.

What kind of distributions are suitable ?

Definition (Balance function of a distribution)

For two distributions of cdf F_1 and F_k, let $F_{1,j}$ and $F_{k,j}$ be the cdf of the mean of j i.i.d samples. The balance function is defined for any $(M, j) \in \mathbb{N}^2$ as

$$\alpha_{1k}(M, j) := \mathbb{E}_{X \sim F_{1,j}} \left((1 - F_{k,j}(X))^M \right) .$$

$\hookrightarrow$ Interpretation: probability that 1 loses M successive "independent" duels with a fixed sample of size j.

Balance condition: $\alpha_{1k}(M, j)$ needs to be "small enough".

Suitable families of distributions

Definition (Assumption 3: Dominant left tail)

We say that ν_1 has a dominant left tail if for all $k \geq 2$:

$$\exists y_k \in \mathbb{R},\ c_k \in (0,1) : \forall x \leq y_k\ ,\ \frac{\mathrm{d}\mathbb{P}_{\nu_1}}{\mathrm{d}\mathbb{P}_{\nu_k}}(x) \leq c_k\ .$$

Examples for which the best arm has a dominant left tail:

- all arms come from the same Single Parameter Exponential family (Bernoulli, Gaussian, Poisson, Exponential, ...)
- $\forall k$, if $X \sim \nu_k$ then $X = \mu_k + \eta$, and η is a centered light-tailed noise with the same distribution for all arms.

Illustration of unusual distributions covered by (A3)

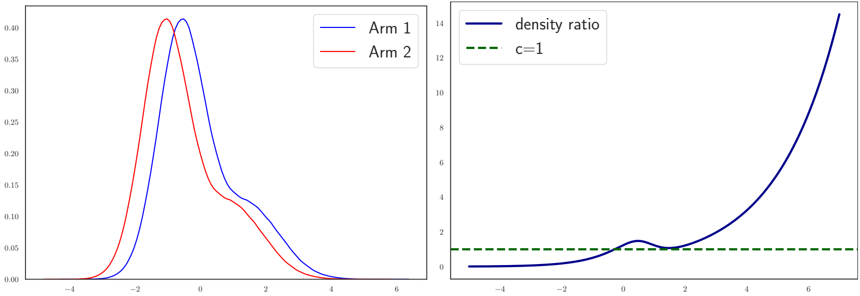


Figure: Two translations of the same Gaussian mixture ($\Delta = 0.5$), and the ratio of their densities with threshold $c = 1$

$\hookrightarrow$ (A3) holds e.g for $y = -0.54$ and $c = 0.8$.

Summary

Theorem

If $\nu = (\nu_1, \ldots, \nu_K) \in \mathcal{F}^K$ is a bandit problem and ***(A1)***, *and* ***(A3)*** *are satisfied, then for all $k \geq 2$* ***LB-SDA*** *and* ***RB-SDA*** *implemented with forced exploration $f_t = \sqrt{\log(t)}$ both satisfy*

$$\mathbb{E}[N_k(T)] \leq \frac{1+\epsilon}{I_1(\mu_k)} \log(T) + \mathcal{O}_\epsilon(1) \,,$$

for any $\epsilon > 0$.

Furthermore, $I_1(\mu_k) = \mathrm{kl}(\mu_k, \mu_1)$ is the **optimal constant** for SPEF: RB-SDA and LB-SDA are even asymptotically optimal.

$\hookrightarrow$ while using no information on the families of distributions!

Empirical results for SDA

Table: Average Regret on 10000 random experiments with Bernoulli Arms (means sampled uniformly)

Horizon	TS	IMED	SSMC	RB	WR
10^2	14	15	17	15	14
10^3	28	32	34	32	31
10^4	46	51	55	51	51
2.10^4	52	58	62	58	57

Table: Average Regret on 10000 random experiments with Gaussian Arms ($\mu_i \sim \mathcal{N}(0,1)$ for each arm i)

Horizon	TS	IMED	RB	WR
10^2	41	45	38	38
10^3	76	82	70	73
10^4	119	124	112	116
2.10^4	133	138	126	130

$\hookrightarrow$ all these results (for any algorithm/time horizon) are very similar ...

... but SDA uses much less knowledge!

Further insights

- We proposed and analyzed two extensions of LB-SDA:
  - A natural extension to non-stationary environment.
  - An adaptation for *Extreme Bandits* with robust comparisons of ”tails”.
- However, there are some cases where SDA does not work: Gaussian with different variances, general bounded distributions . . .

  $\hookrightarrow$ Motivation to continue exploring alternative families of non-parametric algorithms.

Table of Contents

Back to the DSSAT environment

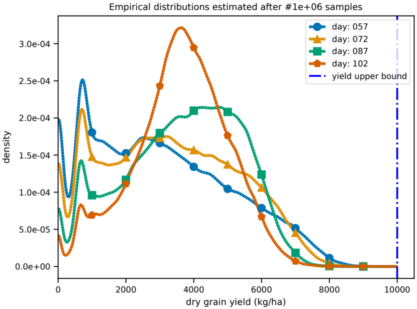


Figure: Yield distribution for different planting dates from the DSSAT simulator

- No simple parametric model for the distributions.

  ↪ the yield may be **bounded** by a yield potential.

- Maximizing the expected yield may not be suitable for the farmers.

  ↪ we want an alternative **risk-aware** metric.

Conditional Value at Risk (CVaR)

Definition: For a distribution ν and $\alpha \in (0, 1]$,

$$\mathrm{CVaR}_\alpha(\nu) = \sup_{x\in\mathbb{R}} \left\{ x - \frac{1}{\alpha}\mathbb{E}_\nu\left[(x-X)^+\right] \right\} \approx \mathbb{E}_{X\sim\nu}[X|X \leq q_\alpha] \, .$$

$\hookrightarrow$ **average** of the fraction α of the **worst possible outcomes**.

We use CVaR to model different farmers' preferences:

- small $\alpha \to$ *food security*. If $\alpha \approx 0$: "worst-case analysis".
- larger $\alpha \to$ *market-oriented* farming. $\alpha = 1$: standard setting.

Back to the DSSAT environment

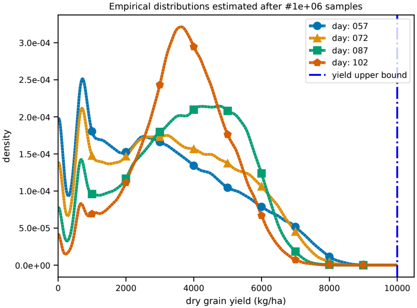


Figure: Yield distribution for different planting dates from the DSSAT simulator

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

α	5%	20%	80%	100%
Blue	0	448	2238	3016
Yellow	46	627	2570	3273
Green	287	1059	3074	**3629**
Red	**538**	**1515**	**3120**	3586

$\neq$ best arm according to $\alpha \Rightarrow$ we need specific **CVaR bandit algorithms**!

CVaR Bandits

A good strategy **pulls the arm with the best CVaR most often.**

At time T, for a bandit $\nu = (\nu_1, \dots, \nu_K)$ we define the α-CVaR regret by

$$\mathcal{R}_\nu^\alpha(T) = \sum_{k=1}^{K} \Delta_k^\alpha \mathbb{E}[N_k(T)] \ ,$$

where Δ_k^α its α-CVaR gap,

$$\Delta_k^\alpha = \max_{i \in [K]} \mathrm{CVaR}_\alpha(\nu_i) - \mathrm{CVaR}_\alpha(\nu_k) \ .$$

Best possible asymptotic performance

Theorem (Regret Lower Bound in CVaR bandits)

Let $\alpha \in (0, 1]$, $\nu = (\nu_1, \dots, \nu_K) \in \mathcal{F}^K$ for some family of distributions $\mathcal{F}$. Then, under any uniformly efficient algorithm it holds for any sub-optimal arm k that

$$\lim_{T \to +\infty} \frac{\mathbb{E}_\nu[N_k(T)]}{\log T} \geq \frac{1}{C_\alpha^{\mathcal{F}}(\nu_k, \nu_1)}.$$

↪ we still target logarithmic regret.

↪ $C_\alpha^{\mathcal{F}}$ extends the notion of asymptotic optimality to CVaR-bandits.

Non-Parametric TS (NPTS) for $\alpha = 1$

- From [Riou and Honda, 2020], generalizes Beta/Bernoulli TS.
- Uses upper bound B, *Dirichlet* distribution $\mathcal{D}_n = \text{Dir}(1, \dots, 1)$.

Consider observations $\mathcal{X} = (X_1, \dots, X_n)$, a step of NPTS computes

$$\boxed{\widetilde{\mu}(\mathcal{X}) = \textstyle\sum_{i=1}^n w_i X_i + w_{n+1} B \ , \quad w \sim \mathcal{D}_{n+1} \ .}$$

Choose $A_t \in \text{argmax}_{k \in [K]} \ \widetilde{\mu}(\mathcal{X}_t^k) \Rightarrow$ **asymptotically optimal** regret.

Motivation: Strong theoretical and empirical performance when $\alpha = 1$, no need for tight concentration inequalities for the CVaR.

B-CVTS for $\alpha \in (0, 1]$

Intuition: re-weighted mean $\rightarrow$ CVaR of a noisy empirical distribution.

Details: given B, α and history $(X_1, \dots, X_n)$:

1. Draw $w = (w_1, \dots, w_{n+1}) \sim \mathcal{D}_{n+1}$, define $\widetilde{\nu}_n$ the distribution with density
$$\widetilde{\nu}_n(x) = \sum_{i=1}^{n} \underbrace{w_i \mathbb{1}(X_i = x)}_{\text{random re-weighting}} + \underbrace{w_{n+1} \mathbb{1}(B = x)}_{\text{exploration bonus}} .$$
2. Return $\widetilde{c}_\alpha := \text{CVaR}_\alpha(\widetilde{\nu}_n)$.

Arm selection: At round t, given the histories $(\mathcal{X}_t^1, \dots, \mathcal{X}_t^k)$ choose

$$A_t = \text{argmax}\, \widetilde{c}_\alpha^k .$$

Theoretical Guarantees

Theorem (Optimality of B-CVTS)

*For any parameter $\alpha \in (0,1]$, if all the distributions are continuous then B-CVTS is **asymptotically optimal**, i.e for any sub-optimal arm k it satisfies*

$$\mathbb{E}[N_k(T)] \leq \frac{\log(T)}{C_\alpha^{\mathcal{F}}(\nu_k, \nu_1)} + o(\log(T)) \,.$$

$\hookrightarrow$ First (provably) asymptotically optimal algorithm in CVaR bandits.

The proof follows [Riou and Honda, 2020], but required technical results for **boundary crossing probabilities**, i.e

$$\mathbb{P}_{w \sim \mathcal{D}_{n+1}}(\mathrm{CVaR}_\alpha(\widetilde{\nu}_{k,n}) \geq c)).$$

Experiments with the DSSAT environment

B-CVTS vs U-UCB (UCB on the CVaR) and CVaR-UCB (CVaR of "optimistic" cdf), same upper bound, $\alpha = 5\%$ and $\alpha = 80\%$.

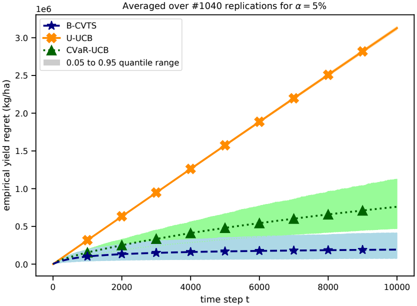


Figure: Averaged CVaR regret and $5\% - 95\%$ CI for 1040 replications with horizon $T = 10^4$ and $\alpha = 5\%$

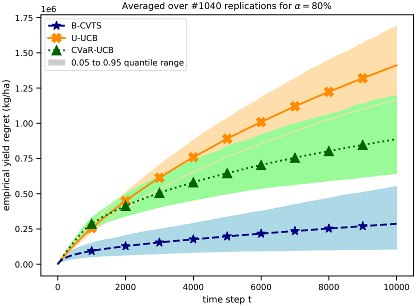


Figure: Averaged CVaR regret and $5\% - 95\%$ CI for 1040 replications with horizon $T = 10^4$ and $\alpha = 80\%$

More experiments: 7 arms, $\alpha = 5\%$

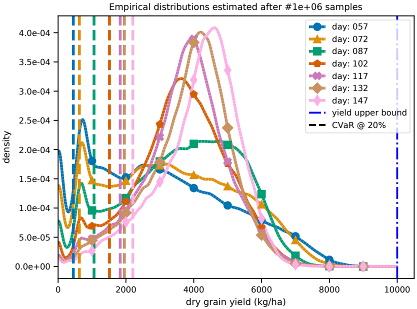


Figure: 7 arms from DSSAT, empirical distributions ; 10^6 samples.

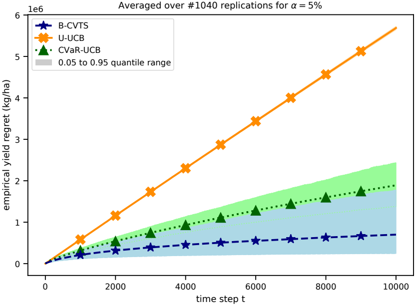


Figure: DSSAT 7 armed bandit, $\alpha = 5\%$; 1040 replications.

More experiments: over-estimating the upper bound

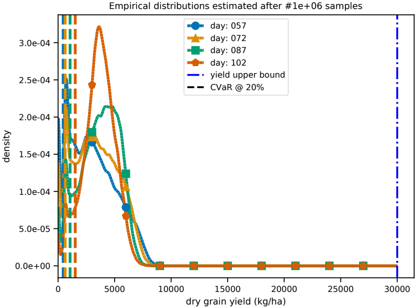


Figure: Initial distributions with over-estimated support ; 10^6 samples.

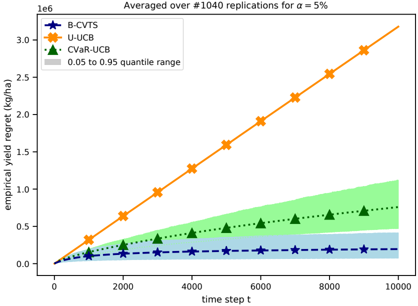


Figure: $\alpha = 5\%$; 1040 replications.

$\hookrightarrow$ Same results as with "exact" upper bound!

$\hookrightarrow$ we can use a conservative upper bound provided by experts.

Further Theoretical Guarantees

Theorem (Logarithmic regret with $B = +\infty$)

If $\alpha < 1$, B is unknown, and B-CVTS runs with $B = +\infty$ it holds that

$$\mathbb{E}[N_k(T)] \leq \frac{\log(T)}{\min\{\log(1/\alpha), C_\alpha^{\mathcal{F}}(\nu_k, \nu_1)\}} + o(\log(T)) \,.$$

$\hookrightarrow$ Optimal if $\log(1/\alpha) \leq C_\alpha^{\mathcal{F}}(\nu_k, \nu_1)$, bounded by $\frac{\log(T)}{\log(1/\alpha)}$ otherwise.

$\hookrightarrow$ the price to pay is small in risk-averse setting:

$$\frac{\log(T)}{\log(1/\alpha)} = 4 \quad \text{for } \alpha = 10\%,\ T = 10^4.$$

Brief overview of Dirichlet Sampling

- For $\alpha = 1$, another strategy is needed if B unknown
- We propose a ***data-dependent*** exploration bonus inside a round-based algorithm.

1. *Bounded Dirichlet Sampling* (BDS): **logarithmic (but close to optimal)** regret for bounded distributions under a *detectability* assumption.
2. *Quantile Dirichlet Sampling* (QDS): **logarithmic regret** for unbounded distributions satisfying a mild quantile condition.
3. *Robust Dirichlet Sampling* (RDS): **slightly larger than logarithmic** regret ($\mathcal{R}_T = \mathcal{O}(\log(T) \log\log(T))$), under (A1) only !

$\hookrightarrow$ Theoretical trade-off between generality and regret guarantees.

Table of Contents

Summary of the contributions

Focus	SDA	NPTS/DS
Non-Parametric assumptions	Concentration (A1) Dominant left tail (A3) ↪ **Logarithmic regret**, **optimal** for SPEF.	From bounded to light-tailed (A1). ↪ **trade-off** theoretical guarantees/assumptions
Alternative metric	**Extreme Bandits** (≈ CVaR for $\alpha \to 0$)	**CVaR Bandits**, $\alpha \in (0, 1]$ for bounded distributions
Extensions	**Limited memory** (LB-SDA) **Non-Stationary** environments (SW-LB-SDA)	**Batched** Feedback

Perspectives

- Extending the sub-sampling idea to structured settings (e.g linear bandits) is non-trivial:
  - Equalizing sample size is not the right thing to do.
  - Equalizing some "information criterion" instead?
- Building optimal NPTS algorithms for unbounded distributions (e.g for sub-gaussian distributions), making SDA work when (A3) does not hold.
- Other interests: bridging the gap between the simulator and the real-world in the use-case in agriculture: taking in account context, spatial/temporal correlations, weather predictions, ...

Thank you for your attention !

SDA for structured/contextual bandits

- Examples: linear bandits, kernel bandits, GP bandits, ...
- Sample size do not reflect the information collected. Linear bandits:

$$r_t = \theta_\star^t x_{A_t} + \eta_t \ , \quad V_t = X_t^T X_t + \lambda I_d \ , \quad .$$

For actions $(x_k)_{k\in[K]}$ we could e.g compare $||x_k||_{V_t^{-1}}^{-1}$.

- Idea:
  1. Leader: $\ell = \text{argmax}_{k\in[K]} ||x_k||_{V_t^{-1}}^{-1}$
  2. Compute estimator $\widehat{\theta}_t$ (all observations), for $k \neq \ell$ compute $\widetilde{\theta}_{k,t}$ (e.g go back in time until the metrics match)
  3. Duel : ${\widehat{\theta}_t}^T x_k$ vs $\widetilde{\theta}_t^T x_\ell$

Challenges: concentration tools, balance condition...

Why Block Samplers?

Lemma (concentration of a sub-sample)

Consider a round $s \leq r$, two distributions ν_a and ν_b under the event $\mathcal{M}_s = \{n_0 \leq N_a(s) \leq N_b(s) \leq r\}$. If $\mathcal{S}_b^s(.,.)$ is a block sampler, for any $\xi \in (\mu_a, \mu_b)$ it holds that

$$\sum_{s=1}^{r} \mathbb{P}\left(\widehat{\mu}_{a,N_a(s)} \geq \widehat{\mu}_{b,\bar{\mathcal{S}}_b^s(N_b(s),N_a(s))}, \mathcal{M}_s\right) \leq \sum_{j=n_0}^{r} \mathbb{P}\left(\widehat{\mu}_{a,j} \geq \xi\right) + r \sum_{j=n_0}^{r} \mathbb{P}\left(\widehat{\mu}_{b,j} \leq \xi\right)$$

Elements of Proof

1. $\{X \leq Y\} \subset \{X \leq \xi\} \cup \{Y \geq \xi\}$
2. $\{N_a = n, a \text{ is pulled}\}$ can only happen once
3. **Union bound on the blocks**, and $\mathbb{P}(\widehat{\mu}_{b,j+1:j+n} \geq \xi) = \mathbb{P}(\widehat{\mu}_{b,j} \geq \xi)$ for any n, and if $N_b < r$ there are at most r blocks.

More on the diversity condition

Diversity = calling the sampler multiple times ensures a variety of sub-samples.

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

Diversity with Block Samplers: An upper bound on $X_{m,H,j}$ is obtained by upper bounding the number of **unique** *starting elements*.

Proofs of the "diversity property" for RB, LB

- RB: drawing random starts allows to cover most of the history with high probability (Lemma 4.3 in [Baudry et al., 2020])
- the leader is pulled sufficiently enough to "move" the sub-sample in a sliding window fashion (Lemma 3 in [Baudry et al., 2021a])

More on the Balance condition

Definition (Balance Condition)

Let $M_t = \mathcal{O}(t/\log t)$, $n_t = \mathcal{O}(\log t)$, and consider some sequence f_t. The balance condition holds between F_1 and F_2 ($\mu_1 > \mu_2$) if

$$\sum_{t=1}^{T} \sum_{j=f_t}^{n_t} \alpha(M_t, j) = o(\log T) \,.$$

- ↪ M_t is the number of "diverse" duels that we are sure to obtain with RB and LB sub-sampling.
- ↪ f_t is an amount of *forced exploration* introduced in SDA, i.e: if some arm satisfies $N_k(t) < f_t$ it is automatically pulled.
- ↪ this is the property that restrains the family of distributions for which SDA works.

Some problems for which sub-sampling requires adaptation

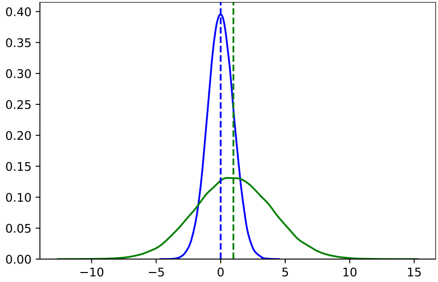


Figure: pdf of distributions $\nu_1 = \mathcal{N}(1, 3)$ and $\nu_2 = \mathcal{N}(0, 1)$

- The best arm has higher variance
- $\mathbb{P}_{\nu_1}(X \leq -5) \approx 10^{-1}$, while $\mathbb{P}_{\nu_2}(X \leq -5) \approx 10^{-7}$

  $\hookrightarrow$ if $X_{11} \leq -5$, arm 1 may be "stuck" for a long time.
- SSTC [Chan, 2020]: compare t-stats,

$$\frac{\widehat{\mu}_{k,n_k} - \widehat{\mu}_{\ell,n_\ell}}{\widehat{\sigma}_{k,n_k}} \text{ vs } \frac{\widehat{\mu}_{\ell,S(n_k,n_\ell)} - \widehat{\mu}_{\ell,n_\ell}}{\widehat{\sigma}_{\ell,S(n_k,n_\ell)}}$$

Some problems for which sub-sampling requires adaptation

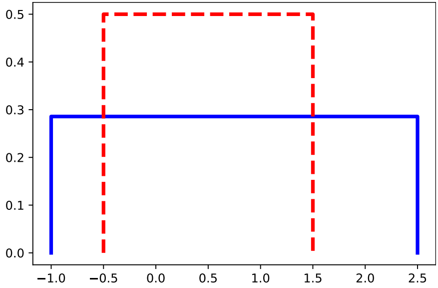


Figure: pdf of distributions $\nu_1 = \mathcal{U}([-1, 2.5])$ and $\nu_2 = \mathcal{U}([-0.5, 1.5])$

- Worst-cases of the best arm cannot be reached by the other arm
- Additional forced exploration/data processing to apply SDA?

$\hookrightarrow$ for known bounded supports the **binarization trick** can be used.

Upper bound on the balance function under (A3)

1. (A3) $\Rightarrow \forall x \leq y_k\ , F_{1,j}(x) \leq c^j F_{k,j}(x)$.

2. $\forall u \leq y_k$:

$$\begin{aligned}\alpha_{1k}(M,j) &\leq F_{1,j}(u) + (1 - F_{k,j}(u))^M \\ &\leq c^j F_{k,j}(u) + e^{-MF_{k,j}(u)} \quad \text{(using (A3) and } \log(1-u) \leq -u\text{)} \\ &\leq \frac{c^j}{M}(1 + \log(M) - j\log(c)) \quad \text{(Optimizing over } F_{k,j}(u)\text{)}\end{aligned}$$

$\hookrightarrow$ sufficient in our proofs with asymptotically negligible forced exploration.

$\hookrightarrow$ If $\alpha_{1k}(M,j) = \mathcal{O}\left(\frac{1}{M\log(M)^a}\right)$ for any a no forced exploration needed.

Very sketchy proof sketch for the regret upper bound

We upper bound $\mathbb{E}[N_k(T)]$ as follow:

- Dominant log term = sum all the events "$\ell(r) = 1$, k is pulled and $N_k(T) \leq \frac{\log T}{\mathrm{kl}(\mu_k, \mu^\star)}$"
  $\hookrightarrow$ Additional constant terms under $\ell(r) \neq 1$

- $\mathbb{E}[\sum_{r=1}^{T} \ell(r) \neq 1]$, decomposed for each r as
  - 1 has already been leader but has been overtaken : highly un-probable:
    $\hookrightarrow$ it must have lost a duel with at least sample size r/K!
  - 1 has never been leader, itself decomposed in
    - Never been leader but relatively large number of samples $N_1(r) = \Omega(\log r)$ $\rightarrow$ very un-likely too
    - Never been leader and "stuck" with a small sample size $N_1(r) = \mathcal{O}(\log r)$:
      this is where we need diversity and balance condition!

Motivation for LB-SDA with limited memory

Theorem (Asymptotic Optimality LB-SDA-LM)

Just as LB-SDA, LB-SDA-LM is asymptotically optimal when arms belong to the same Single-Parameter Exponential Family (SPEF).

Table: Storage/computational cost at round T for some subsampling algorithms.

Algorithm	Storage	Comp. cost: Best-Worst case
SSMC [Chan, 2020]	$O(T)$	$O(1)$-$O(T)$
RB-SDA	$O(T)$	$O(\log T)$
LB-SDA	$O(T)$	$O(1)$-$O(\log T)$
LB-SDA-LM	$O((\log T)^2)$	$O(1)$-$O(\log T)$

LB-SDA-LM with Bernoulli arms

$\mu_1 = 0.05$
$\mu_2 = 0.15$

Memory:

$$m_r = \log(r)^2 + 50$$

$\hookrightarrow$ Between 50 and 150 samples kept for each arm.

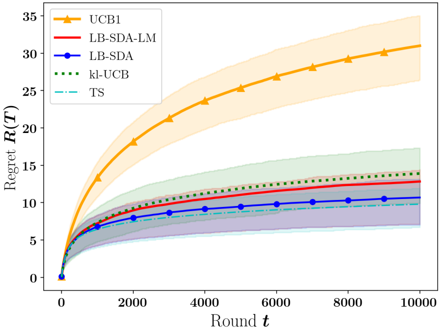


Figure: Cost of storage limitation on a Bernoulli instance. The reported regret are averaged over 2000 independent replications.

$\rightarrow$ Limiting memory does not have a significant cost in this example!

Abruptly Changing Environments: SW-LB-SDA

Sliding Window LB-SDA

- Natural adaptation of LB-SDA with a sliding window of size τ
- Additional mechanisms to ensure sufficient exploration
- Non-parametric nature $\Rightarrow$ potential for new settings

Theorem (Regret Guarantees)

If the time horizon T and number of breakpoints Γ_T are known, and that between each breakpoints the arms are from the same SPEF, choosing $\tau = \mathcal{O}(\sqrt{T \log(T)/\Gamma_T})$ ensures that the dynamic regret of SW-LB-SDA satisfies

$$\mathcal{R}_T = \mathcal{O}(\sqrt{T\Gamma_T \log T}) \, .$$

Example: SW-LB-SDA with Gaussian arms

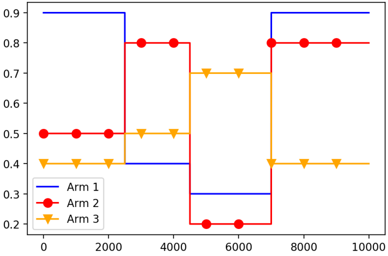


Figure: Time-dependent means, associated with standard deviations $\sigma = \{0.25, 0.5, 1, 0.25\}$

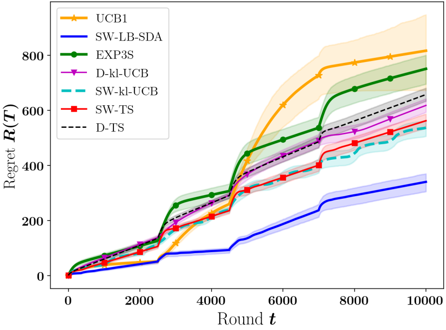


Figure: Performance on this Gaussian instance, averaged on 2000 independent replications.

→ SW-LB-SDA naturally adapts to the variance changes!

SDA for Extreme Bandits (very short introduction)

- Extreme Bandits: maximize $\mathbb{E}[\max_t X_t] \Rightarrow$ find arm with heavier tail
- Non-parametric approaches are appealing, but hard to derive theoretical guarantees.
- Compare **Quantile of Maxima** $\Rightarrow$ nice concentration properties
- Two algorithms: **QoMax-ETC** (needs horizon T), and **QoMax-SDA** (anytime).
- Strong theoretical guarantees under mild assumptions, strong empirical performance.

Intuition: why Dirichlet re-sampling works in B-CVTS?

For distributions bounded by B, it holds that for $c \geq \mathrm{CVaR}_\alpha(\widetilde{\nu}_n)$ and any $n \in \mathbb{N}$

$$-\frac{1}{n} \log\left(\mathbb{P}_{w \sim \mathcal{D}_{n+1}}(\mathrm{CVaR}_\alpha(\widetilde{\nu}_n) \geq c)\right) = C_\alpha^{\mathcal{F}}(\widehat{\nu}_n, c) + o(1) \,.$$

$\hookrightarrow$ Dirichlet Sampling implicitly samples with a rate related to the $C_\alpha^{\mathcal{F}}$.

- **Upper bound:** Chernoff method, Dirichlet weights as a normalized sum of independent exponential r.v, and properties of the CVaR.
- **Lower bound:** discretization argument as in [Riou and Honda, 2020] + working directly on the integral.

Highlights of the analysis

2 regimes in the analysis: **Post-Convergence** and **Pre-Convergence** (arm is sampled more (resp. less) than the optimal rate).

- Post-CV: The empirical distribution will eventually get "close enough" to the true (DKW inequality), so that

$$C_\alpha^{\mathcal{F}}(\widehat{\nu}_n, c) \approx C_\alpha^{\mathcal{F}}(\nu, c) \ .$$

$\hookrightarrow$ we use the continuity of $\mathcal{K}_{\text{inf}}^{\alpha}$ in both arguments.

- Pre-CV: Adding the upper bound B in the history allows to balance all "bad scenarios". Illustration with multinomial distributions,

$$\frac{\mathbb{P}(\widehat{\nu}_n)}{\mathbb{P}_{w\sim\mathcal{D}_{n+1}}(\text{CVaR}_\alpha(\widetilde{\nu}_n) \geq c|\widehat{\nu}_n)} \leq \exp(-n\delta_c)$$

for some universal constant $\delta > 0$, if $c \leq \text{CVaR}_\alpha(\nu)$.

Dirichlet Sampling (DS)

Another way to perform duels

- Leader → **empirical mean** $\widehat{\mu}_\ell$.
- Challenger → **Dirichlet Sampling** with a bonus $\mathfrak{B}(k, \ell)$.
- Winner: largest of the two.

Inspired by the Non-Parametric TS of [Riou and Honda, 2020], DS computes a "biased re-weighted mean"

$$\widetilde{\mu}(k, \ell, \mathfrak{B}) = \sum_{i=1}^{n} w_i X_i + w_{n+1} \underbrace{\mathfrak{B}(k, \ell)}_{\substack{\textbf{data-dependent} \\ \textbf{exploration bonus} \\ \textbf{arm } k \textbf{ vs arm } \ell}} \quad , \text{ with } ||w||_1 = 1.$$

where $w \sim \mathcal{D}_{n+1}(1, \ldots, 1)$ (Dirichlet distribution with param 1 for each item)

First theoretical guarantees

Theorem (Generic regret decomposition of DS)

Consider a bandit model satisfying (A1). Then, for any re-weighted mean depending only on the empirical mean of ℓ, it holds for any $\epsilon \in [0, \Delta_k)$ that

$$\mathbb{E}\left[N_k(T)\right] \leq n_k(T) + B^k_{T,\epsilon} + C_{\nu,\epsilon} \ ,$$

where $C_{\nu,\epsilon}$ is independent on T and

$$n_k(T) = \mathbb{E}\left[\sum_{r=1}^{T-1} \mathbb{1}(k \in \mathcal{A}_{r+1}, \ell(r) = 1)\right] \ ,$$

where $\ell(r)$ is the leader at round r, and

$$B^k_{T,\epsilon} = \sum_{k'=2}^{K} \sum_{n=1}^{\lceil 2\log(T)/I_1(\mu_k+\epsilon)\rceil} \sup_{\widehat{\mu}\in[\mu_{k'}-\epsilon, \mu_{k'}+\epsilon]} \mathbb{E}\left[\frac{\mathbb{1}\left(\mu_{1,n} \leq \widehat{\mu}\right)}{\mathbb{P}(\widetilde{\mu}(1, k', \mathfrak{B}) \geq \widehat{\mu})}\right] \ .$$

Choice of the Exploration Bonus $\mathfrak{B}(k, \ell)$

Lemma (Necessary condition with a data-independent bonus)

Consider a fixed bonus B_μ, and denote by F_1 the cdf of ν_1. Then, $B^k_{T,\epsilon}$ can converge only if

$$B_\mu > \mu + \frac{1}{1 - F_1(\mu)} \mathbb{E}_{X \sim F_1} \left[(\mu - X)_+ \right] \ .$$

This result motivates a bonus of the form

$$\mathfrak{B}(k, \ell) := B\left(X, \widehat{\mu}_\ell, \rho\right) := \widehat{\mu}_\ell + \rho \times \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{\mu}_\ell - X_{k,i}\right)^+ \ ,$$

for some parameter ρ that will be tuned under different assumptions (not necessarily on $F_1(\mu)$).

Boundary Crossing Probability

We call "Boundary Crossing Probability" (BCP) the quantity

$$[\text{BCP}] := \mathbb{P}_{w \sim \mathcal{D}_{n+1}} \left(\sum_{i=1}^{n+1} w_i X_i \geq \mu \right) ,$$

where $(X_1, \ldots, X_n)$ is a collection of *fixed* data and $w \sim \mathcal{D}_{n+1}(1, \ldots, 1)$.

$\hookrightarrow$ the design of DS algorithms is guided by upper and lower bounds on the BCP.

Three algorithms to relax the bounded support assumption

- Bounded DS (BDS):
  - $\mathfrak{B}(k,\ell) = B$ if it is known (= NPTS [Riou and Honda, 2020]).
  - $\mathfrak{B}(k,\ell) = \max\{\max X_i + \gamma, B(X, \widehat{\mu}_\ell, \rho)\}$ for $\rho \geq \frac{-1}{\log(1-p)}$ if B is unknown but $\exists \gamma, p$: $\mathbb{P}([B-\gamma, B]) \geq p \Rightarrow$ upper bound unknown but detectable.
- Quantile DS (QDS): replace the fraction α of best outcomes of arm k by their mean (un-biased truncation), use $\mathfrak{B}(k,\ell) = B(X, \widehat{\mu}_\ell, \rho)$ with $\rho \geq \frac{1+\alpha}{\alpha^2} \Rightarrow$ enough information before the quantile so that the best arm can be identified.
- Robust DS (RDS): use $\mathfrak{B}(k,\ell) = B(X, \widehat{\mu}_\ell, \rho_{n_k}) \Rightarrow$ no assumption at all.

Theoretical Results: from optimality to robustness

- *Bounded Dirichlet Sampling* (BDS) is **optimal** for bounded distributions with known upper bound, and has **logarithmic (but close to optimal)** regret under the detectability assumption.
- *Quantile Dirichlet Sampling* (QDS) has a **logarithmic regret** for distributions satisfying a mild quantile condition.
- *Robust Dirichlet Sampling* (RDS) has **slightly larger than logarithmic** regret ($\mathcal{R}_T = \mathcal{O}(\log(T) \log\log(T))$, but for all *light-tailed distributions*.

$\hookrightarrow$ the choice of the algorithm depends on the quantity of information we have on the distributions. In any case, RDS can be used.

$\hookrightarrow$ Theoretical trade-off between generality and regret guarantees, but in practice all algorithms perform very well.

Look back: SDA vs DS

Question: In a roud-based algorithm, what can we do to give a fair chance to the challenger?

- **Penalizing the leader** by using a subset of its observations, ***Sub-Sampling Dueling Algorithms*** [Baudry et al., 2020].

  $\hookrightarrow$ works because the leader's sample size is large.

- **Boosting the challenger** by randomly re-sampling its observation and an exploration bonus based on the leader's history: ***Dirichlet Sampling*** [Baudry et al., 2021b].

  $\hookrightarrow$ works because with appropriate assumptions on the distributions and because the mean of leader concentrates.

Baransi, A., Maillard, O.-A., and Mannor, S. (2014). Sub-sampling for multi-armed bandits. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 115–131. Springer.

Baudry, D., Kaufmann, E., and Maillard, O.-A. (2020). Sub-sampling for efficient non-parametric bandit exploration. Advances in Neural Information Processing Systems, 33.

Baudry, D., Russac, Y., and Cappé, O. (2021a). On limited-memory subsampling strategies for bandits. In Meila, M. and Zhang, T., editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 727–737. PMLR.

Baudry, D., Saux, P., and Maillard, O. (2021b). From optimality to robustness: Dirichlet sampling strategies in stochastic bandits. CoRR, abs/2111.09724.

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

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

Chan, H. P. (2020). The multi-armed bandit problem: An efficient nonparametric solution. The Annals of Statistics, 48(1):346–373.

Honda, J. and Takemura, A. (2015).
Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards.
Journal of Machine Learning Research, 16:3721–3756.

Hoogenboom, G., Porter, C., Boote, K., Shelia, V., Wilkens, P., Singh, U., White, J., Asseng, S., Lizaso, J., Moreno, L., et al. (2019).
The dssat crop modeling ecosystem.
Advances in crop modelling for a sustainable agriculture, pages 173–216.

Kaufmann, E., Korda, N., and Munos, R. (2012).
Thompson sampling: An asymptotically optimal finite-time analysis.
In Algorithmic Learning Theory - 23rd International Conference, ALT.

Riou, C. and Honda, J. (2020).
Bandit algorithms based on thompson sampling for bounded reward distributions.
In Algorithmic Learning Theory - 31st International Conference (ALT) 2012.