From Optimality to Robustness: Dirichlet Sampling Strategies in Stochastic Bandits

Dorian Baudry[1], Patrick Saux[1], Odalric-Ambrym Maillard [1]

[1]CNRS, Inria, UniversitÃľ de Lille

Motivation: Bandit application in Agriculture

Problem: Recommending a planting date to maize farmers under challenging growing conditions to maximize grain yield.

We use the DSSAT crop model [?] to simulate grain yield distributions under stochastic weather for each planting date. We *in sillico* evaluate different algorithms to make a decision on planting dates in a tailored bandit environment.

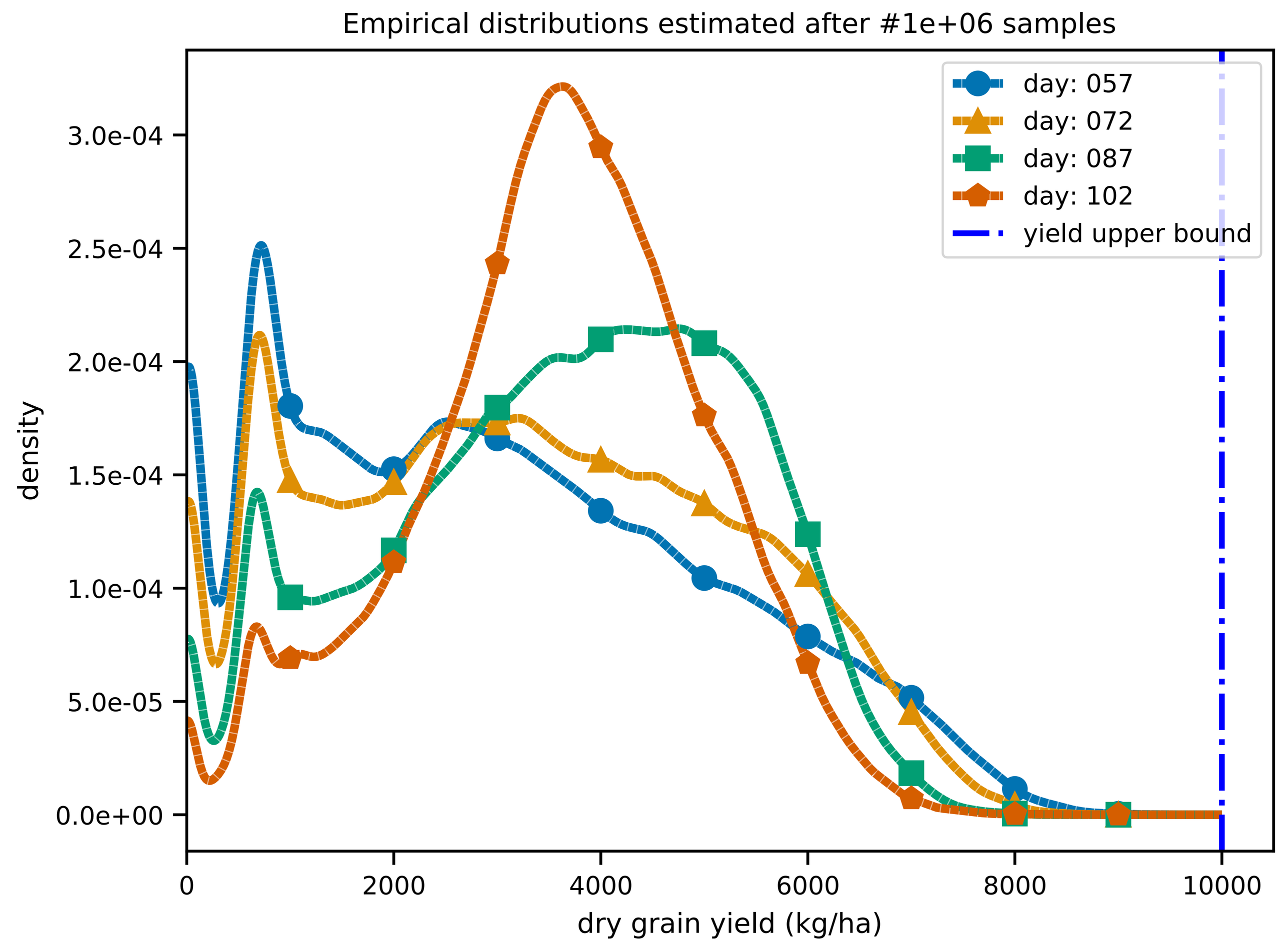


Fig. 1: Example of distributions from the DSSAT simulator

- No simple parametric model for the distributions, but they are bounded due to physical constraints. Expert knowledge can provide a reasonable upper bound.
- Maximizing the expectation is not satisfying (e.g Food Security): we want a **Risk-Aware** policy, accounting for individual farmers' risk aversion levels.

$\hookrightarrow$ We consider the CVaR bandits framework, and make the hypothesis of a known bounded support.

Theoretical formalism: CVaR Bandits

CVaR is a *coherent risk measure* [?], defined for a parameter α and a distribution ν as

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

- K unknown reward distributions called *arms.*
- The learner sequentially collects rewards and updates her policy.
- Objective: **Minimizing the α-CVaR regret**

$$\mathcal{R}_T^\alpha = \mathbb{E}\left[\sum_{t=1}^T (\max_k \mathrm{CVaR}_\alpha(\nu_k) - \mathrm{CVaR}_\alpha(\nu_{A_t}))\right].$$

Lower bound in CVaR bandits

Our first contribution is an extension of the Burnetas & Katehakis lower bound [?] for CVaR bandits.

Theorem 1: Regret Lower Bound in CVaR bandits Let $\alpha \in (0,1]$. Let $\mathcal{F} = \mathcal{F}_1 \times \dots \times \mathcal{F}_K$ be a set of bandit models and $\nu = (\nu_1, \dots, \nu_K)$ where each ν_k belongs to the class of distribution $\mathcal{F}_k$. Then, under any *uniformly efficient* strategy the expected number of pulls of a sub-optimal arm k satisfies

$$\lim_{T\to+\infty} \frac{\mathbb{E}_\nu[N_k(T)]}{\log T} \geq \frac{1}{\mathcal{K}_{\inf}^{\alpha,\mathcal{F}_k}(\nu_k, c^\star)},$$

where $c^\star = \max_{i\in\{1,\dots,K\}} \mathrm{CVaR}_\alpha(\nu_i)$, and

$$\mathcal{K}_{\inf}^{\alpha,\mathcal{F}_k}(\nu_k, c^\star) = \inf_{\nu'\neq\nu_k\in\mathcal{F}_k} \left\{\mathrm{KL}(\nu_k,\nu') : \mathrm{CVaR}_\alpha(\nu') \geq c^\star\right\}.$$

$\hookrightarrow$ Any algorithm matching this lower bound is called *asymptotically optimal*

Algorithms: M-CVTS and B-CVTS

- Inspired by NPTS [?]

$$\tilde{\mu}(\mathcal{X}, B) = \sum_{i=1}^n w_i X_i + w_{n+1} B,$$

- Pairwise comparisons (duels) inspired by [?, ?]
  (i) Choose a **leader**: arm with largest number of observations!
  (ii) Perform $K-1$ **duels**: *leader* vs each *challenger*.
  (iii) Draw a set of arms: *winning challengers* (if any) or *leader* (if none).

Theorem 1: Generic Regret Decomposition For any light-tailed bandit problem $\nu = (\nu_1, \dots, \nu_K)$ and any bonus $\mathfrak{B}(\ell, k)$, for any suboptimal arm k it holds that

$$\mathbb{E}[N_k(T)] \leq \underbrace{n_k(T)}_{\substack{\text{Sample size needed}\\ \text{to "separate" arm } k\\ \text{from the best arm}}} + \underbrace{B_T^\nu}_{\substack{\text{Capacity of DS}\\ \text{strategy to "recover" from a}\\ \text{bad scenario for the best arm}}} + \underbrace{\mathcal{O}(1)}_{\substack{\text{Constant terms}\\ \text{from light-tailed}\\ \text{concentration}}}.$$

The analysis of boundary crossing probabilities for Dirichlet distributions suggests the following exploration bonus, with tunable **leverage** ρ:

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

Three Dirichlet Sampling instances

Bounded Dirichlet Sampling (BDS)

- Case 1: $X \leq B$ with **known** B: $\mathfrak{B}(\ell,k) = B$.
- Case 2: $\mathbb{P}\left(X \in [B-\gamma, B]\right) \geq p$ with **known** γ, p (but not B!):

$$\mathfrak{B}(\ell,k) = \max\left(B(\mathcal{X}_k, \hat{\mu}_\ell, \rho), \max_{i=1,\dots,n} X_{k,i} + \gamma\right).$$

Theorem 2 BDS is optimal in case 1 (NPTS) and 2 ($\rho \geq -1/\log(1-p)$).

Quantile Dirichlet Sampling (QDS)

✂ Truncate after quantile $1-\alpha$ and summarize the right tail by its CVaR.

Theorem 3 For any $\rho \geq (1+\alpha)/\alpha^2$, QDS has **logarithmic regret** for the family of semi-bounded distributions that are "dense enough" after their quantile $1-\alpha$.

Robust Dirichlet Sampling (RDS)

✖ Only assuming light tails is incompatible with $\log T$ regret.

💡 Intuition: $\rho = \rho_n$ must grow to ∞ to eventually capture all possible settings:

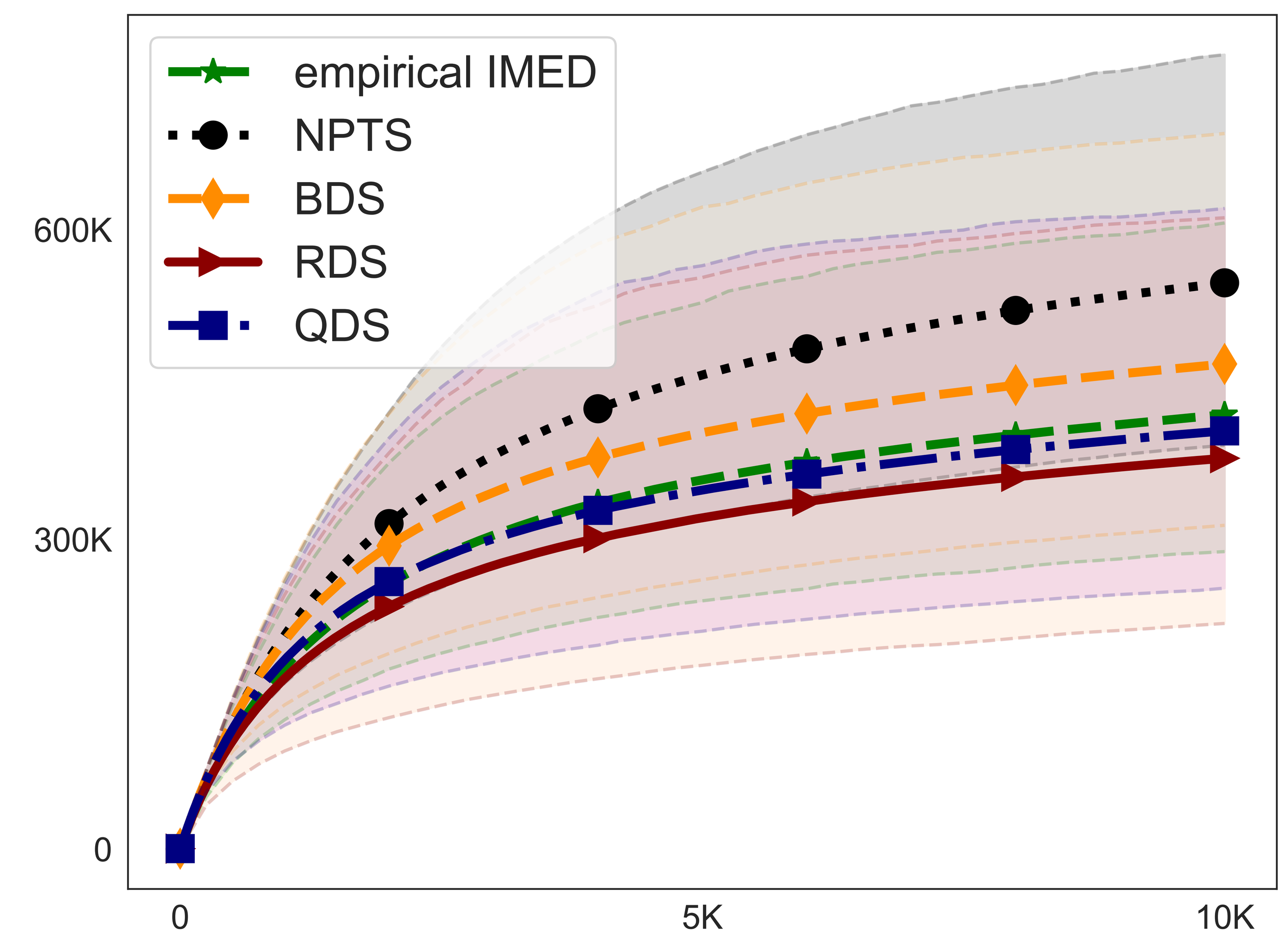
$$\sum_{i=1}^n w_i X_{k,i} + w_{n+1} B(\mathcal{X}_k, \hat{\mu}_\ell, \rho_n).$$

Theorem 4 Let $\rho_n \to +\infty, \rho_n = o(n)$. For **light-tailed distributions**, RDS satisfies

$$\mathcal{R}_T = \mathcal{O}\left(\log(T)\log\log(T)\right).$$

$\hookrightarrow$ We recommend $\rho_n = \sqrt{\log n}$ as a baseline.

Numerical Experiments



Regrets on the 7 armed DSSAT bandit of Figure 1, 5000 replications. Empirical IMED and NPTS are run with 50% larger upper bound to replicate conservative expert knowledge. The overall winner is **RDS**.

Aknowledgments: The authors acknowledge the funding of the French National Research Agency under projects BADASS and I-Site ULNE.