From Optimality to Robustness: Dirichlet Randomized Exploration in Stochastic Bandits

Patrick Saux[1]
(Joint work with Dorian Baudry[1] and Odalric-Ambrym Maillard[1])

[1] Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000, Lille, France

May 4, 2022

Motivation: recommendations in the real world

Online advertising

- ✔ Huge volume of (meta)data.
- ✔ Limited risks.
- ✔ Easy to model and simulate (Bernoulli, logistic...).

Motivation: recommendations in the real world

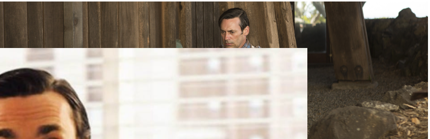
Online advertising

- ✔ Huge volume of (meta)data.
- ✔ Limited risks.
- ✔ Easy to model and simulate (Bernoulli, logistic...).

Farming

- ✘ Slow and scarce data collection process.
- ✘ Risky.
- ? Simulation?

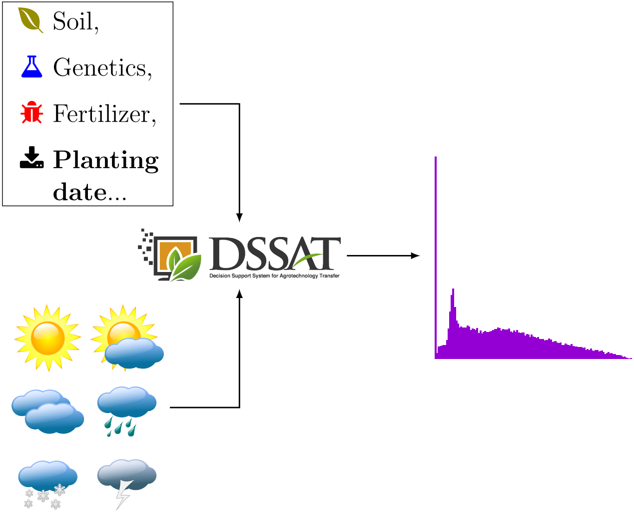
Motivation: recommendations in the real world



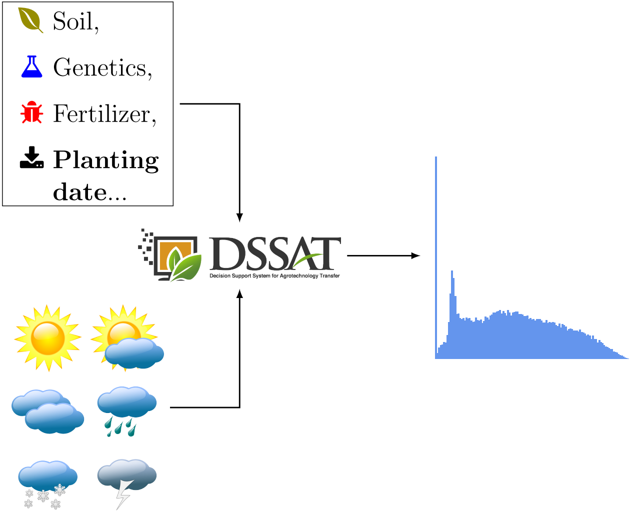
Online advertisi

- Huge volur
- Limited risl
- Easy to mc (Bernoulli,

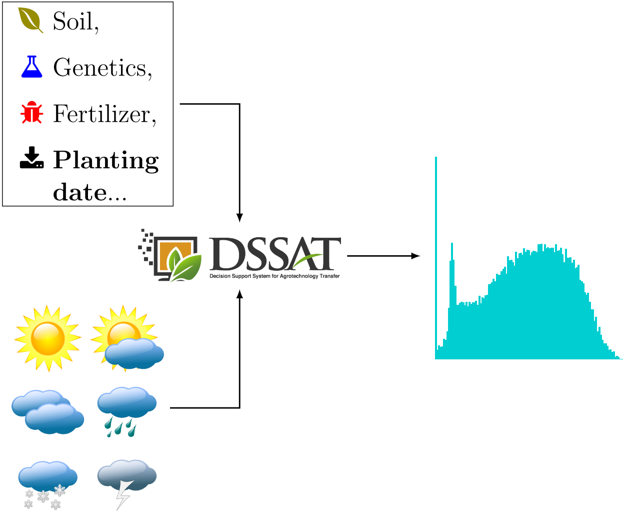
Motivation: recommendations in the real world



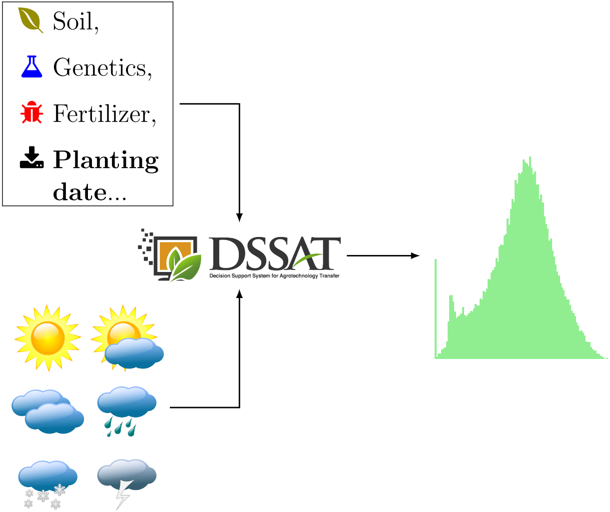
Motivation: recommendations in the real world



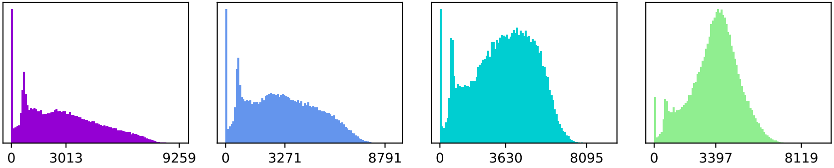
Motivation: recommendations in the real world



Motivation: recommendations in the real world



Motivation: recommendations in the real world



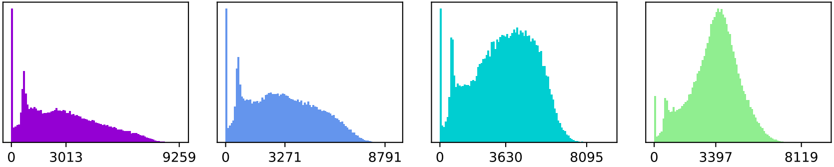
- Observe crop yields (**rewards**) $X_{k,t} \sim \nu_k$ for planting date k (**arm**).
- Minimize **regret** of policy $(\pi_t)_{t=1,\dots,T}$ on a bandit instance $\nu \in \mathcal{F}$:

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

$$\liminf_{T \to +\infty} \frac{\mathbb{E}\left[N_k(T)\right]}{\log T} \geq \frac{1}{\underbrace{\inf\left\{\mathrm{KL}(\nu_k, \widetilde{\nu}) \,\middle|\, \widetilde{\nu} \in \mathcal{F}, \mathbb{E}_{X \sim \widetilde{\nu}}\left[X\right] > \mu^*\right\}}_{\mathcal{K}_{\inf}^{\mathcal{F}}(\nu_k, \mu^*)}}.$$

Motivation: recommendations in the real world

DSSAT
Decision Support System for Agrotechnology Transfer



- Observe crop yields (**rewards**) $X_{k,t} \sim \nu_k$ for planting date k (**arm**).
- Minimize **regret** of policy $(\pi_t)_{t=1,\dots,T}$ on a bandit instance $\nu \in \mathcal{F}$:

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

$$\liminf_{T \to +\infty} \frac{\mathbb{E}\left[N_k(T)\right]}{\log T} \geq \frac{1}{\underbrace{\inf\left\{ \mathrm{KL}(\nu_k, \widetilde{\nu}) \mid \widetilde{\nu} \in \mathcal{F}, \mathbb{E}_{X \sim \widetilde{\nu}}\left[X\right] > \mu^* \right\}}_{\mathcal{K}_{\text{inf}}^{\mathcal{F}}(\nu_k, \mu^*)}}.$$

Optimal bandit algorithms: SPEF

$$\mathcal{F} = \left\{ \nu \text{ with density } p_\theta(x) = h(x) e^{\theta F(x) - \mathcal{L}(\theta)},\ \theta \in \Theta \subseteq \mathbb{R} \right\}.$$

Algorithm	Scope for optimality	Algorithm parameters
kl-UCB[1]	Single Parameter Exponential Family (SPEF) $(\nu_\theta)_{\theta\in\Theta}$	$\mathrm{KL}(\nu_\theta, \nu_{\theta'})$
IMED[2]		$\mathrm{KL}(\nu_\theta, \nu_{\theta'})$
Thompson Sampling[3]		Prior/Posterior
SDA[4]		Non-parametric

1. Cappé et al. (2013), 2. Honda and Takemura (2015), 3. Korda et al. (2013), 4. Baudry et al. (2020).

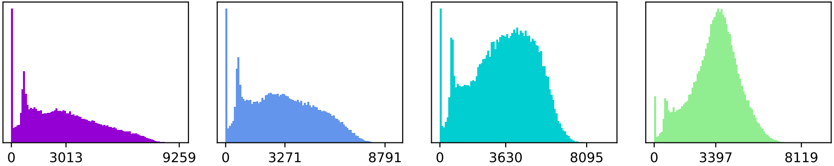
Optimal bandit algorithms: bounded

$$\mathcal{F}_B = \{\nu \text{ such that } \mathbb{P}_{X\sim\nu}(X \in [b, B]) = 1\}\,.$$

Algorithm	Scope for optimality	Algorithm parameters
Empirical IMED[2]	$\text{Supp}(\nu) \subset (-\infty, B]$ ν is light-tailed*	Upper bound B
Empirical KL-UCB[1] NPTS[5]	$\text{Supp}(\nu) \subset [b, B]$	

1. Cappé et al. (2013), 2. Honda and Takemura (2015), 5. Riou and Honda (2020).

Motivation: which setting should we use?



- SPEF ? Definitely not ✖.
- Bounded ? Which choice for B **?**

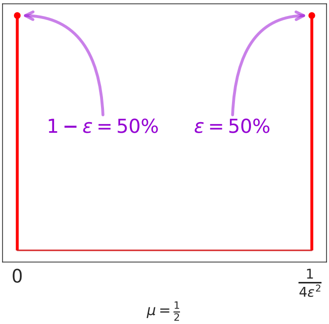
  $$B_1 \leq B_2 \implies \mathcal{K}_{\text{inf}}^{\mathcal{F}_{B_1}}(\nu_k, \mu^*) \geq \mathcal{K}_{\text{inf}}^{\mathcal{F}_{B_2}}(\nu_k, \mu^*).$$
- Light-tailed ? Reasonable assumption ✔

  $\hookrightarrow$ $\exists \lambda_0 > 0 : \forall \lambda \in [-\lambda_0, \lambda_0]$, $\mathbb{E}\left[e^{\lambda X}\right] < +\infty$.

Can we find algorithms assuming only that the distributions are light-tailed, without strong parametric assumptions on the tails?

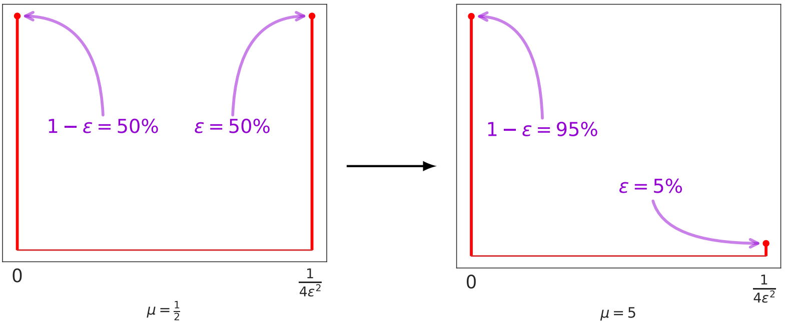
What can we expect?

How about not knowing the bound B? Not knowing the variance?



What can we expect?

How about not knowing the bound B? Not knowing the variance?



$\hookrightarrow$ Mass leakage at infinity!
Hadiji and Stoltz (2020),
Ashutosh et al. (2021).

Nonparametric Thompson Sampling

- From Riou and Honda (2020)
- Pull arm with best **resampled** mean, denoting $\mathcal{X} = (X_1, \dots, X_n)$ an arms' history,

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

- $w \sim \mathcal{D}_{n+1}(1, \dots, 1)$ (Dirichlet distribution),
- B: upper bound of the support of the arms' distribution.

✔ optimal for a large class of distributions...

✖ ... upper bounded by a **known** B.

Nonparametric Thompson Sampling

- From Riou and Honda (2020)
- Pull arm with best **resampled** mean, denoting $\mathcal{X} = (X_1, \dots, X_n)$ an arms' history,

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

- $w \sim \mathcal{D}_{n+1}(1, \dots, 1)$ (Dirichlet distribution),
- B: upper bound of the support of the arms' distribution.

✔ optimal for a large class of distributions...

✘ ... upper bounded by a **known** B.

We generalize to **Dirichlet Sampling**, comparing two arms k and ℓ with

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

Using data-dependent bonus in pairwise comparisons

A **round-based** approach Chan (2020); Baudry et al. (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).

$\hookrightarrow$ possibly several arms drawn per round.

Pairwise comparison **(Duel)** step:

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

Intuition: After r rounds, the leader has at least r/K data, its sample mean should be an accurate estimation. On the other hand, DS ensures enough exploration for the challengers!

Technical tool #1: duel-based regret decomposition

Theorem (Regret decomposition)

For any light-tailed bandit problem $\nu = (\nu_1, \ldots, \nu_K)$ and any bonus $\mathfrak{B}(\ell, k)$, for any suboptimal arm k it holds that

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

- $n_k(T)$ will be the first-order term, and mostly depend on the family of distributions $\mathcal{F}$.
- The bonus $\mathfrak{B}(\cdot, \cdot)$ will be essentially designed to obtain $B_T^\nu = \mathcal{O}(1)$ for the family $\mathcal{F}$.

Dirichlet Randomized Exploration

$$\widetilde{\mu}(k, \ell, \mathfrak{B}) = \sum_{i=1}^{n} w_i X_i + w_{n+1} \mathfrak{B}(k, \ell)$$

↓

Exploration bonus $\mathfrak{B}(k, \ell)$

Technical tool #2: boundary crossing probability

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

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

$\mathcal{X}_{n+1} = (X_1, \dots, X_{n+1})$ is a collection of *fixed* data and $w \sim \mathcal{D}_n(1, \dots, 1)$.

Lemma (BCP bounds)

Let $\bar{\mathcal{X}}_{n+1} = \max \mathcal{X}_{n+1} \geq g(n)$ *and* $\bar{\Delta}_n^+ = \frac{1}{n} \sum_{X_i < \bar{\mathcal{X}}_{n+1}} (\mu - X_i)^+$, *then*

$$-\frac{n\bar{\Delta}_n^+}{g(n) - \mu} \leq \log [BCP] \leq -(n+1)\mathcal{K}_{\text{inf}}^{\mathcal{F}_{\bar{\mathcal{X}}_{n+1}}} (\widehat{\nu}_{\mathcal{X}_{n+1}}, \mu) .$$

$\hookrightarrow$ motivates the following exploration bonus with **leverage** $\rho > 0$:

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

Dirichlet Randomized Exploration

$$\widetilde{\mu}(k, \ell, \mathfrak{B}) = \sum_{i=1}^{n} w_i X_i + w_{n+1} \mathfrak{B}\,(k, \ell)$$

Exploration bonus $\mathfrak{B}(k, \ell)$

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

Algorithm #1: Bounded Dirichlet Sampling (BDS)

Case 1: known upper bound

$X \leq B$ with **known** B:

$$\mathfrak{B}(\ell, k) = B \quad \text{(NPTS, Riou and Honda (2020))} \,.$$

Case 2: unknown but detectable bound

$\mathbb{P}(X \in [B-\gamma, B]) \geq p$ with **known** γ, p (but not B!):

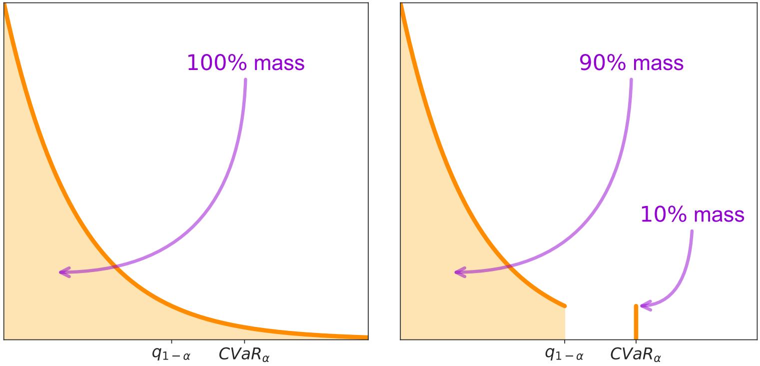
$$\mathfrak{B}(\ell, k) = \max\left(B(\mathcal{X}_k, \widehat{\mu}_\ell, \rho), \max_{i=1,\dots,n} X_{k,i} + \gamma \right) \,.$$

Theorem

For any $\rho \geq -1/\log(1-p)$, BDS is optimal in case 2 for the family $\mathcal{F}_{\gamma,p} = \{\nu : \exists B_\nu : \mathbb{P}(X \leq B_\nu) = 1$ and $\mathbb{P}(X \in [B_\nu - \gamma, B_\nu]) \geq p\}$.

Algorithm #2: Quantile Dirichlet Sampling (QDS)

? What about unbounded distributions?



✄ ... truncate them!

Algorithm #2: Quantile Dirichlet Sampling (QDS)

- n_α: exact number of observations smaller than the $1-\alpha$ quantile.
- $\widehat{C}_{k,\alpha} = \frac{1}{n-n_\alpha} \sum_{i=n_\alpha+1}^{n} X_{k,(i)}$: empirical CVaR ($\approx \mathbb{E}_\nu[X|X > q_{1-\alpha}(\nu)]$).
- Non-uniform Dirichlet sampling $w \sim \mathcal{D}(\underbrace{1,\dots,1}_{n_\alpha}, n-n_\alpha, 1)$ and:

$$\sum_{i=1}^{n_\alpha} w_i X_{k,(i)} + w_{n_\alpha+1}\widehat{C}_{k,\alpha} + w_{n_\alpha+2} B(\mathcal{X}_k, \widehat{\mu}_\ell, \rho) \; .$$

Theorem

*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 $1-\alpha$.*

Algorithm #3: Robust Dirichlet Sampling (RDS)

Can we have no assumption at all?

- ✖ Not with $\log T$ regret: Hadiji and Stoltz (2020), Ashutosh et al. (2021)
- 💡 Intuition: $\rho = \rho_n$ must grow to ∞ to eventually capture all possible settings:

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

Theorem

*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!

Dirichlet Randomized Exploration

$$\widetilde{\mu}(k,\ell,\mathfrak{B}) = \sum_{i=1}^{n} w_i X_i + w_{n+1}\mathfrak{B}\,(k,\ell)$$

Exploration bonus $\mathfrak{B}(k,\ell)$

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

BDS

$\rho \geq \frac{-1}{\log(1-p)}$

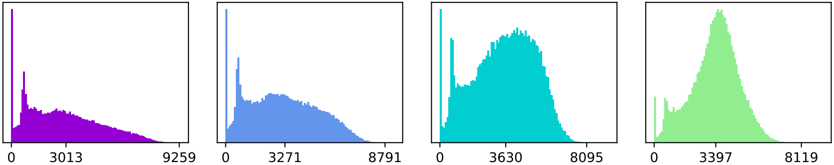
QDS

$\rho \geq \frac{1+\alpha}{\alpha^2}$

RDS

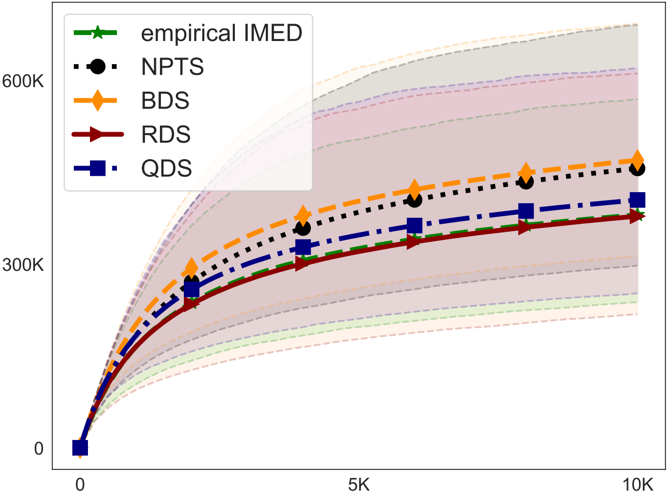
$\rho_n = \sqrt{\log(n)}$

Experiments: recommendations in agriculture



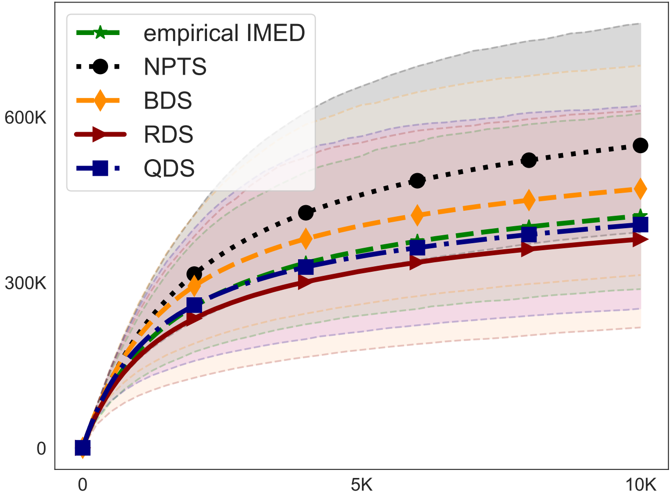
We compare DS algorithms with optimal algorithms considering bounded distributions with known B.

Experiments: recommendations in agriculture



(Oracle upper bound).

Experiments: recommendations in agriculture



(Conservative expert upper bound, 50% larger than oracle).

Conclusion

- ✔ Generic regret analysis of round-based index policies,
- ✔ Analysis of BCP and empirical $\mathcal{K}_{\inf}$,
- ✔ Three instances of Dirichlet Randomized Exploration with strong guarantees.

Future works?

- ? Heavy-tails? Reweighting of median of means?
- ? Resampling in contextual bandits?
- ? Deployment *in vivo* for agricultural recommendations?

Ashutosh, K., Nair, J., Kagrecha, A., and Jagannathan, K. (2021). Bandit algorithms: Letting go of logarithmic regret for statistical robustness. In International Conference on Artificial Intelligence and Statistics, pages 622–630. PMLR.

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

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.

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

Hadiji, H. and Stoltz, G. (2020). Adaptation to the range in k-armed bandits. arXiv preprint arXiv:2006.03378.

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.

Korda, N., Kaufmann, E., and Munos, R. (2013). Thompson sampling for one-dimensional exponential family bandits. In Advances in Neural Information Processing Systems.

Riou, C. and Honda, J. (2020). Bandit algorithms based on thompson sampling for bounded reward distributions. In Algorithmic Learning Theory, pages 777–826. PMLR.