

REINFORCEMENT LEARNING MULTI-ARMED BANDITS

Odalric-Ambrym Maillard

2021-2022

Inria Scool

- ▷ Why bandits?
- ▷ Exploration-Exploitation trade-off and why max empirical objective is bad.
- ▷ Finite-sample estimation error: Concentration of measure.
- ▷ Optimal strategies:
 - ▶ (KL-)UCB: Optimistic
 - ▶ TS: Bayesian
 - ▶ IMED: Information theory
 - ▶ SDA: Sub-sampling
- ▷ Structure: When actions are informative about other actions
 - ▶ Linear and GP bandits for optimization
 - ▶ IMED-Unimodal
 - ▶ IMED-Lipschitz
 - ▶ IMED-Graph
 - ▶ IMED-Equivalence

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

OPTIMAL STRUCTURE EXPLOITATION

CONCLUSION, PERSPECTIVE

- ▷ Q-learning and co: \simeq MC estimates, never cares about **sampling error**.
How **many samples** used before good policy?

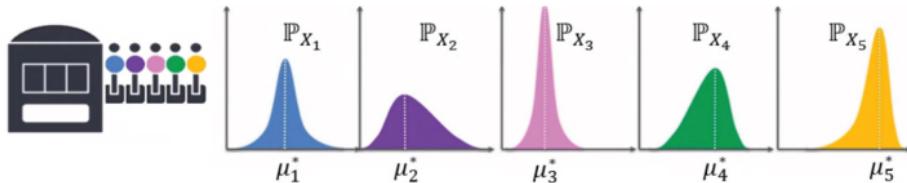
- ▷ **HUGE** amount of **training** data: e.g. for AlphaStar
 - ~ 500 years of cumulated human time on Go,
 - ~ 200 years of cumulated human time on StarCraft
 - ~ 45 years of cumulated human time on Chess,
 - ... not to count all **previously tested** neural architectures

Sample efficiency ? (+ Computation power/efficiency ?)

- ▷ Plus these are perfect simulators
Use samples in an more **optimal** way?

- ▷ **Simulated** world: (Games, Digital models, etc.)
Can make **as many mistakes** as we want and **reset** at any time.
Samples are **cheap/fast** to generate.
- ▷ **Real** world: (Health-care, agriculture, drug-testing, customer-service, etc.)
Every mistake is costly and there is **no reset** possible.
Samples are **costly/slow** to generate.
(Pre-training with a simulator is desirable, but still all models are wrong...)
- ▷ Not only good **final** policy, but at **lowest possible sampling/learning cost!**
Minimize **cumulated errors** while learning.

- ▷ Multi-armed bandits are a simplified version of RL: no dynamics!



- ▷ Used to study and minimize the sampling error, with performance guarantee.
- ▷ Used when too complex system, life-threatening decisions, etc.
- ▷ So far the most applicable part of RL to real-world problems.
- ▷ Promising approach to designing faster and safer RL algorithms

- ▷ Adaptive **clinical trials** (already in 1930's)!
- ▷ **News/Add** recommender systems (last decade)!
- ▷ Audrey Durand's work (Univ. Laval):
 - ▶ Mouse cancer treatment
 - ▶ High-resolution microscope image acquisition
 - ▶ etc.
- ▷ **Agriculture, E-learning, Health-care** (Now).

TABLE OF CONTENTS

WHY BANDITS?

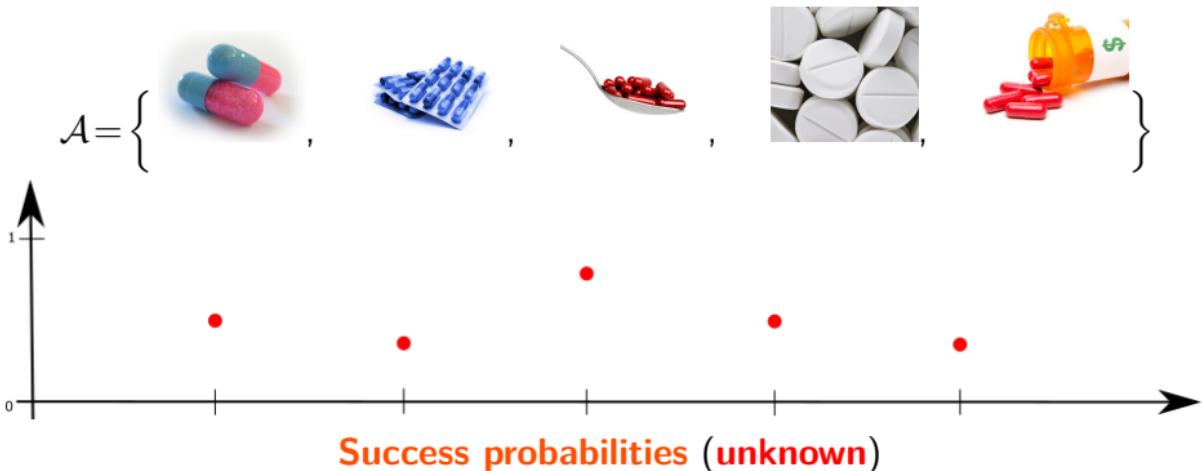
VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

OPTIMAL STRUCTURE EXPLOITATION

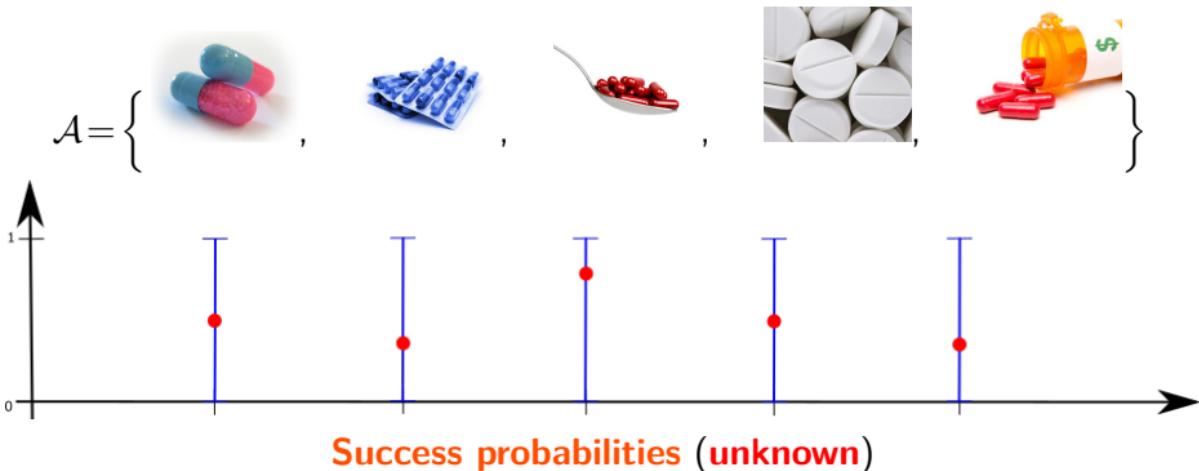
CONCLUSION, PERSPECTIVE



At each time step t :

- ▷ Consider a new patient
- ▷ Choose one treatment $A_t \in \mathcal{A}$
- ▷ Observe output $Y_t = +1$ if cured, 0 else.

We want to maximize number of patients cured (while learning success probabilities).



At each time step t :

- ▷ Consider a new patient
- ▷ Choose one treatment $A_t \in \mathcal{A}$
- ▷ Observe output $Y_t = +1$ if cured, 0 else.

We want to maximize number of patients cured (while learning success probabilities).

The goal of an agent is to **accumulate rewards** over time:

$$\sum_{t=1}^T r_t$$

r_t = reward obtained in state s_t when playing action a_t

We compare to an oracle agent having access to **full knowledge** of the system.

Regret

$$\mathcal{R}_T = \underbrace{\sum_{t=1}^T r_t^*}_{\text{Optimal strategy}} - \underbrace{\sum_{t=1}^T r_t}_{\text{Learner strategy}}$$

- ▷ The sampling strategy (or bandit algorithm) (A_t) is sequential:

$$A_{t+1} = \pi(\underbrace{A_1, Y_1, \dots, A_t, Y_t}_{\text{past history}}).$$

It may depend on **past history**, and be **randomized**.

- ▷ Why minimizing regret?

- ▶ **Optimize while learning**
- ▶ Avoid **mistakes** at each single step (clinical trials: don't kill patients!).
- ▶ Similar to classical **statistical risk** notion.
- ▶ This is **applied** in real-world applications!

We target algorithms with **regret minimization** guarantees.

MULTI-ARMED BANDIT APPLICATION

Basic model (first approximation) for:

- ▶ Clinical trials: (Thompson, 1933)



- ▶ Casino slot machines: (Robbins, 1952)



MULTI-ARMED BANDIT APPLICATION

Basic model (first approximation) for:

- ▶ Clinical trials: (Thompson, 1933)



- ▶ Casino slot machines: (Robbins, 1952)



- ▶ Ad-placement: (2000-2010)



MULTI-ARMED BANDIT APPLICATION

- ▶ Plant health-care:


$$\} \quad \{$$

- ▶ Ground health-care:

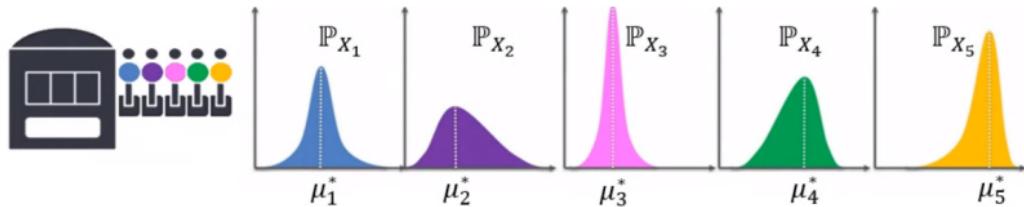

$$\} \quad \{$$

- ▶ Variety choice:


$$: \mathcal{A} = \{$$

$$\} \quad \{$$

MULTI-ARMED STOCHASTIC BANDIT MODEL



Arms: $[n] \triangleq \{1, \dots, n\}$

$(X_t)_t \stackrel{iid}{\sim} \mathbb{P}_X, \mathbb{E}[X] = \mu^* \in \mathbb{R}^n$

$\forall \lambda \in \mathbb{R}, \mathbb{E}[e^{\lambda(X_t - \mu^*)}] \leq e^{\frac{\kappa_i^2 \lambda^2}{2}}$ (sub-Gaussianity)

A FIRST, SIMPLE GAME

Say you have arbitrary complex system but only two strategies: $\mathcal{A} = \{\pi_1, \pi_2\}$.



π_1 :



$r_1 = 0.8$

A FIRST, SIMPLE GAME

Say you have arbitrary complex system but only two strategies: $\mathcal{A} = \{\pi_1, \pi_2\}$.



$\pi_1 :$



$\pi_1 :$



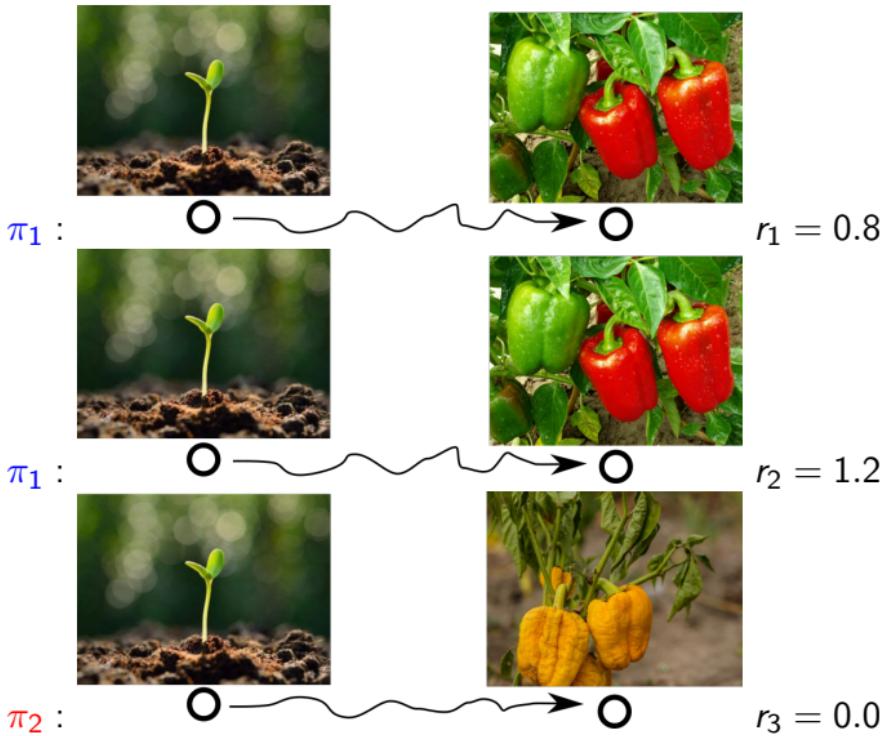
$r_1 = 0.8$



$r_2 = 1.2$

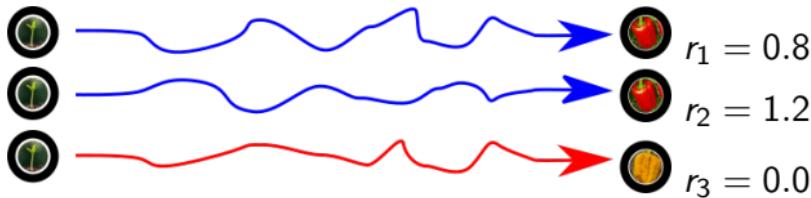
A FIRST, SIMPLE GAME

Say you have arbitrary complex system but only two strategies: $\mathcal{A} = \{\pi_1, \pi_2\}$.



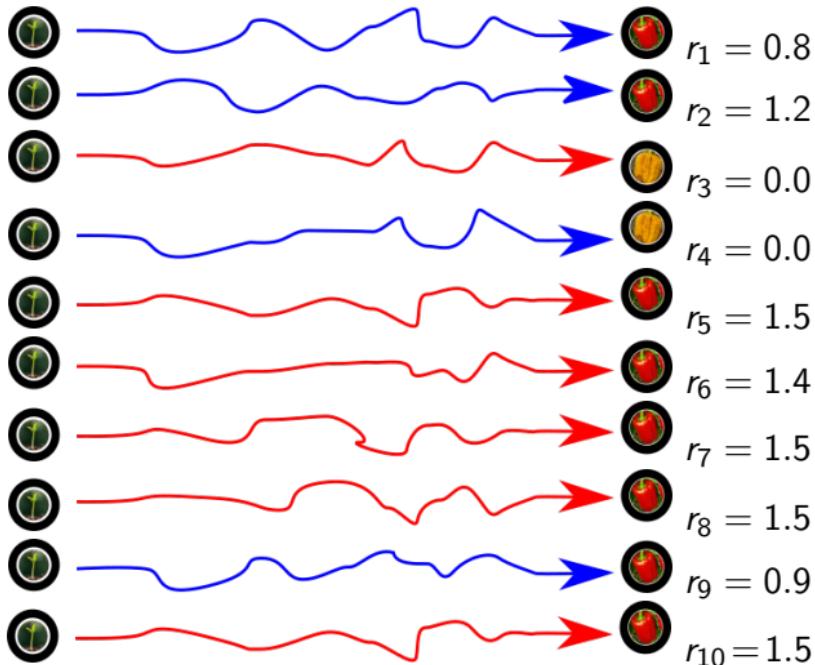
A FIRST, SIMPLE GAME

Say you have arbitrary complex system but only two strategies: $\mathcal{A} = \{\pi_1, \pi_2\}$.



A FIRST, SIMPLE GAME

Say you have arbitrary complex system but only two strategies: $\mathcal{A} = \{\pi_1, \pi_2\}$.



WHAT DO YOU GET?

Out of this naive approach, you get:

- ▷ Total reward: $r_1 + r_2 + \dots + r_{10} = 10.3$
- ▷ Reward of π_1 : 2.9 on $N_1 = 4$ trial; $\hat{\mu}_{\pi_1, N_1} = 0.725$ per trial. $\hat{\sigma}_1 = 0.44$
- ▷ Reward of π_2 : 7.4 on $N_2 = 6$ trial; $\hat{\mu}_{\pi_2, N_2} = 1.233$ per trial, $\hat{\sigma}_2 = 0.55$

We observe that:

- ▷ First 3 rounds indicate π_1 better than π_2 (but few observations).
- ▷ First 10 rounds indicate π_2 better than π_1 .

How far are we from truth?

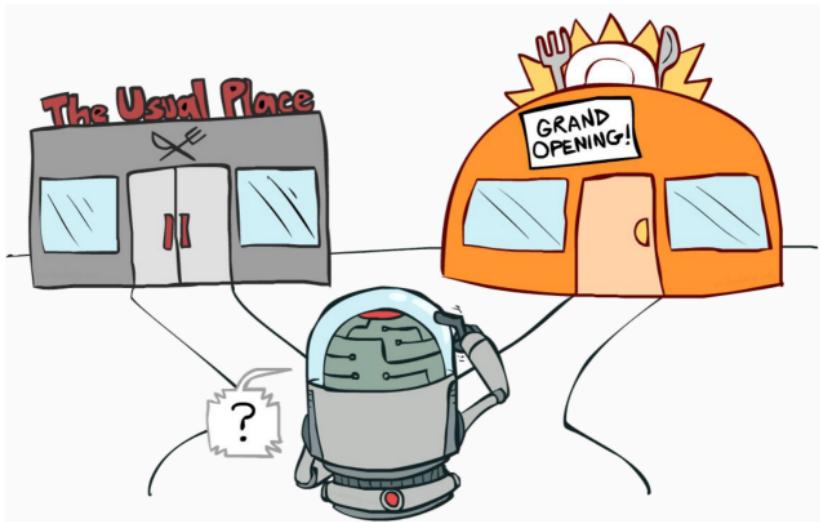
Quantify uncertainty?

Also, it seems that:

- ▷ We should try each policy enough to get accurate estimation (**Exploration**)
- ▷ We should play only the best arm since each mistake is costly (**Exploitation**)

Exploration-Exploitation tradeoff

EXPLORATION EXPLOITATION TRADE-OFF



The **cumulative reward** $\sum_{t=1}^T Y_t$ is a random variable.

▷ **Goal:** find a strategy maximizing its **expectation**

$$\mathbb{E} \left[\sum_{t=1}^T Y_t \right].$$

Oracle (knows the means): always play an arm

$$a^* \in \operatorname{Argmax}_{a \in \mathcal{A}} \mu_a \quad \text{with mean} \quad \mu^* = \max_{a \in \mathcal{A}} \mu_a.$$

▷ Can we be *almost as good as the oracle*?

$$\mathbb{E} \left[\sum_{t=1}^T Y_t \right] = \mathbb{E} \left[\sum_{t=1}^T \mu_{A_t} \right] \simeq T\mu^*?$$

Stochastic regret

Difference between performance of **oracle** policy \star and **learner** policy π in environment ν after T steps:

$$\begin{aligned}\mathcal{R}_T(\nu, \pi) &\stackrel{\text{def}}{=} \mathbb{E}_{\star} \left[\sum_{t=1}^T Y_t \right] - \mathbb{E}_{\pi} \left[\sum_{t=1}^T Y_t \right] \\ &= \sum_{t=1}^T \mu^{\star} - \mathbb{E}_{\pi} \left[\sum_{t=1}^T Y_t \right]\end{aligned}$$

We want the regret to **grow sub-linearly** $\frac{\mathcal{R}_T(\nu, \pi)}{T} \xrightarrow[T \rightarrow \infty]{} 0$ (*consistency*).

- ▶ what rate of regret can we expect?

Lemma (Regret decomposition)

For any policy π and time horizon T ,

$$\mathcal{R}_T(\nu, \pi) = \sum_{a \in \mathcal{A}} (\underbrace{\mu^* - \mu_a}_{\text{Gap}}) \mathbb{E}[N_a(T)]$$

where $N_a(T) = \sum_{t=1}^T \mathbb{I}\{A_t = a\}$.

Gap: $\Delta_a = (\mu^* - \mu_a)$ is problem-dependent, unknown, deterministic.

1. For any t ,
$$\sum_{t=1}^T Y_t = \sum_{t=1}^T \sum_{a \in \mathcal{A}} Y_t \mathbb{I}\{A_t = a\}.$$
2. Hence
$$\mathcal{R}_T(\nu, \pi) = \sum_{t=1}^T \sum_{a \in \mathcal{A}} \mathbb{E}[(\mu^\star - Y_t) \mathbb{I}\{A_t = a\}].$$
3. Now,

$$\begin{aligned}
 \mathbb{E}[(\mu^\star - Y_t) \mathbb{I}\{A_t = a\}] &= \mathbb{E}[\mathbb{E}[\mathbb{I}\{A_t = a\}(\mu^\star - Y_t) | A_t]] \\
 &= \mathbb{E}[\mathbb{I}\{A_t = a\}(\mu^\star - \mathbb{E}[Y_t | A_t])] \\
 &= \mathbb{E}[\mathbb{I}\{A_t = a\}(\mu^\star - \mu_{A_t})] \\
 &= \mu^\star - \mu_a.
 \end{aligned}$$

Setup

- $\nu = \{\nu_a\}_{a \in \mathcal{A}} \subset \mathcal{D}$: unknown real-valued probability distributions.

Setup

- ▶ $\nu = \{\nu_a\}_{a \in \mathcal{A}} \subset \mathcal{D}$: unknown real-valued probability distributions.
- ▶ $\mu_* = \max_{a \in \mathcal{A}} \mu_a$, where $\{\mu_a\}_{a \in \mathcal{A}}$ denote the means.

Game and regret

At each time $t \in \mathbb{N}$, play $a_t \in \mathcal{A}$ from π , receive $Y_t \sim \nu_{a_t}$. Minimize

$$\mathcal{R}_T(\nu, \pi) \stackrel{\text{def}}{=} \mathbb{E}_* \left[\sum_{t=1}^T Y_t \right] - \mathbb{E}_\pi \left[\sum_{t=1}^T Y_t \right] = \sum_{a \in \mathcal{A}} \underbrace{\mu_* - \mu_a}_{\Delta_a} \mathbb{E}_\pi \left[\underbrace{\sum_{t=1}^T \mathbb{I}_{a_t=a}}_{N_a(T)} \right].$$

The environment **does not** reveal the rewards of the other arms.

The expectation summarizes any possible source of randomness.

We want the regret to **grow sub-linearly** $\frac{\mathcal{R}_T(\nu, \pi)}{T} \xrightarrow[T \rightarrow \infty]{} 0$ (*consistency*).

TAKE HOME MESSAGE

To minimize **regret**, **DO NOT** play

$$\hat{\star}_t = \operatorname{argmax}_{a \in \mathcal{A}} \hat{\mu}_a(t) \quad \text{where } \hat{\mu}_a(t) = \frac{1}{N_t(a)} \sum_{t=1}^T Y_t \mathbb{I}\{A_t = a\}$$

TAKE HOME MESSAGE

To minimize **regret**, **DO NOT** play

$$\hat{\star}_t = \operatorname{argmax}_{a \in \mathcal{A}} \hat{\mu}_a(t) \quad \text{where } \hat{\mu}_a(t) = \frac{1}{N_t(a)} \sum_{t=1}^T Y_t \mathbb{I}\{A_t = a\}$$

(yet all MC-based methods Q-learning, (LS)TD, DQN, etc. do so)

To minimize **regret**, **DO NOT** play

$$\hat{\star}_t = \operatorname{argmax}_{a \in \mathcal{A}} \hat{\mu}_a(t) \quad \text{where } \hat{\mu}_a(t) = \frac{1}{N_t(a)} \sum_{t=1}^T Y_t \mathbb{I}\{A_t = a\}$$

(yet all MC-based methods Q-learning, (LS)TD, DQN, etc. do so)

Four type of perturbed strategies instead

- ▷ UCB: $\operatorname{argmax}_{a \in \mathcal{A}} \hat{\mu}_a(t) + B_a(t)$ where $B_a(t) \geq \mu_a - \hat{\mu}_a(t)$ with high probability.
- ▷ TS/DS: $\operatorname{argmax}_{a \in \mathcal{A}} \tilde{\mu}_a(t)$ where $\tilde{\mu}_a \sim \text{Posterior}/\text{Randomly reweighted mean}$.
- ▷ IMED: $\operatorname{argmin}_{a \in \mathcal{A}} N_t(a) \mathbf{D}(\hat{\mu}_a(t), \max_a \hat{\mu}_a(t)) + \ln(N_t(a))$ with divergence D .
- ▷ SDA: All $\{a : \mu_a^\dagger(t) \geq \max_a \hat{\mu}_a(t)\}$ where $\mu_a^\dagger(t)$ mean of $N_t(\hat{\star}_t)$ -many **randomly chosen** observations from a .

To minimize **regret**, **DO NOT** play

$$\hat{\star}_t = \operatorname{argmax}_{a \in \mathcal{A}} \hat{\mu}_a(t) \quad \text{where } \hat{\mu}_a(t) = \frac{1}{N_t(a)} \sum_{t=1}^T Y_t \mathbb{I}\{A_t = a\}$$

(yet all MC-based methods Q-learning, (LS)TD, DQN, etc. do so)

Four type of perturbed strategies instead

- ▷ UCB: $\operatorname{argmax}_{a \in \mathcal{A}} \hat{\mu}_a(t) + B_a(t)$ where **Optimistic** $= \hat{\mu}_a(t)$ with high probability.
- ▷ TS/DS: $\operatorname{argmax}_{a \in \mathcal{A}} \tilde{\mu}_a(t)$ where $\tilde{\mu}_a \sim \text{Beta}(\hat{\mu}_a(t), N_t(a))$ /Randomly reweighted **mean**.
- ▷ IMED: $\operatorname{argmin}_{a \in \mathcal{A}} N_t(a) \mathbf{D}(\hat{\mu}_a(t), \max_b \hat{\mu}_b(t)) + \ln(N_t(a))$ with divergence D .
- ▷ SDA: All $\{a : \mu_a^\dagger(t) \geq \max_a \hat{\mu}_a(t)\}$ where $\mu_a^\dagger(t)$ mean of $N_t(\hat{\star}_t)$ -many **randomly chosen** observations from a . **Sub-sampling**

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

First strategies

Confidence bounds

The optimism principle

Performance bounds

Best-achievable regret bounds

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

Problem 1: The environment **does not** reveal the rewards of the arms not pulled by the learner

Problem 1: The environment **does not** reveal the rewards of the arms not pulled by the learner

⇒ the learner should **gain information** by repeatedly pulling all the arms

Problem 1: The environment **does not** reveal the rewards of the arms not pulled by the learner

⇒ the learner should **gain information** by repeatedly pulling all the arms

Problem 2: Whenever the learner pulls a **bad arm**, it suffers some regret

Problem 1: The environment **does not** reveal the rewards of the arms not pulled by the learner

⇒ the learner should **gain information** by repeatedly pulling all the arms

Problem 2: Whenever the learner pulls a **bad arm**, it suffers some regret

⇒ the learner should **reduce the regret** by repeatedly pulling the best arm

Problem 1: The environment **does not** reveal the rewards of the arms not pulled by the learner

⇒ the learner should **gain information** by repeatedly pulling all the arms

Problem 2: Whenever the learner pulls a **bad arm** it suffers some regret

⇒ the learner should **reduce the regret** by repeatedly pulling the best arm

Challenge: The learner should solve two opposite problems!

Problem 1: The environment **does not** reveal the rewards of the arms not pulled by the learner

⇒ the learner should **gain information** by repeatedly pulling all the arms ⇒ **exploration**

Problem 2: Whenever the learner pulls a **bad arm** it suffers some regret

⇒ the learner should **reduce the regret** by repeatedly pulling the best arm

Challenge: The learner should solve two opposite problems!

Problem 1: The environment **does not** reveal the rewards of the arms not pulled by the learner

⇒ the learner should **gain information** by repeatedly pulling all the arms ⇒ **exploration**

Problem 2: Whenever the learner pulls a **bad arm** it suffers some regret

⇒ the learner should **reduce the regret** by repeatedly pulling the best arm ⇒ **exploitation**

Challenge: The learner should solve two opposite problems!

Problem 1: The environment **does not** reveal the rewards of the arms not pulled by the learner

⇒ the learner should **gain information** by repeatedly pulling all the arms ⇒ **exploration**

Problem 2: Whenever the learner pulls a **bad arm** it suffers some regret

⇒ the learner should **reduce the regret** by repeatedly pulling the best arm ⇒ **exploitation**

Challenge: The learner should solve the **exploration-exploitation** dilemma!

SOME (NAIVE) STRATEGIES

- ▷ **Idea 1 :** Pull each arm T/A times

⇒ EXPLORATION

$$\mathcal{R}_T(\nu, \pi) = \left(\frac{1}{A} \sum_{a=2}^A (\mu_1 - \mu_a) \right) T$$

SOME (NAIVE) STRATEGIES

▷ **Idea 1** : Pull each arm T/A times

⇒ EXPLORATION

$$\mathcal{R}_T(\nu, \pi) = \left(\frac{1}{A} \sum_{a=2}^A (\mu_1 - \mu_a) \right) T$$

▷ **Idea 2** : Always pull the **empirical best** arm (cf. Q-learning, etc!)

$$A_{t+1} = \operatorname{argmax}_{a \in \mathcal{A}} \hat{\mu}_a(t)$$

where

$$\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^t Y_s \mathbb{I}_{(A_s=a)}$$

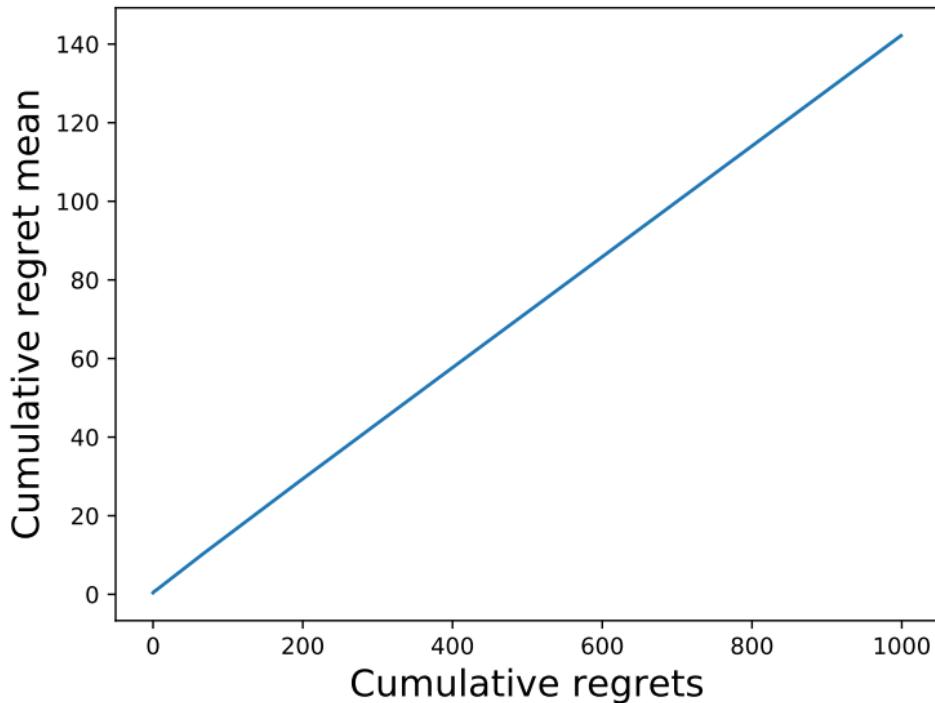
is an estimate of the unknown mean μ_a .

⇒ EXPLOITATION “Follow The Leader”

For Bernoulli bandit $\nu = (\mathcal{B}(\mu_1), \mathcal{B}(\mu_2))$:

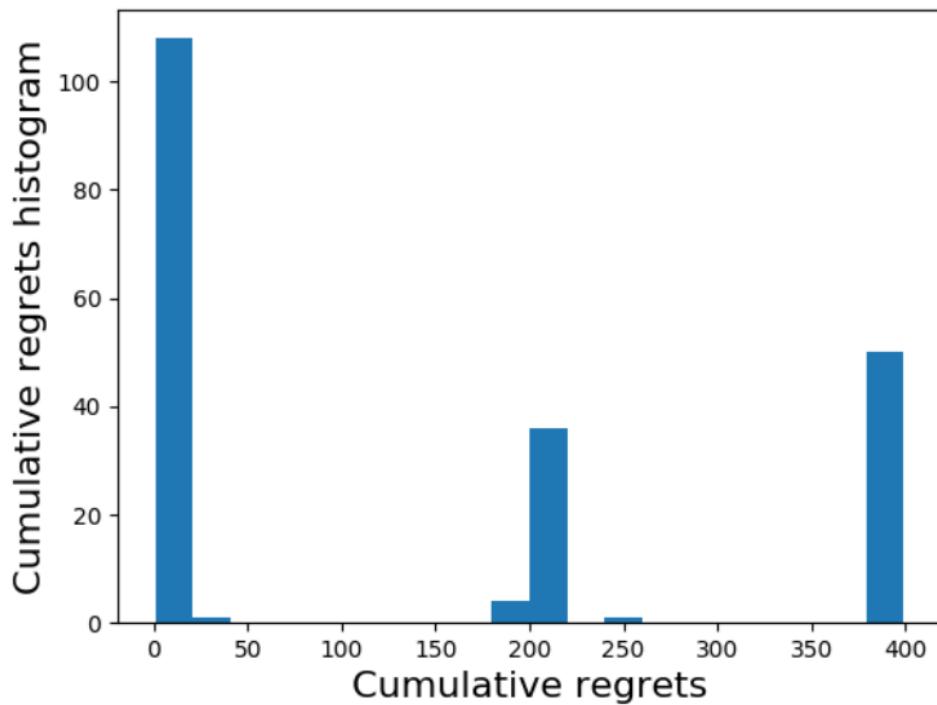
$$\mathcal{R}_T(\nu, \pi) \geq (1 - \mu_1) \times \mu_2 \times (\mu_1 - \mu_2) T$$

REGRET OF FTL FOR A $[\mathcal{B}(0.2), \mathcal{B}(0.4), \mathcal{B}(0.6)]$ -BANDIT



Results averaged over 200 runs.

REGRET OF FTL FOR A $[\mathcal{B}(0.2), \mathcal{B}(0.4), \mathcal{B}(0.6)]$ -BANDIT



Given $m \in \{1, \dots, T/A\}$,

- ▶ draw each arm m times
- ▶ compute the empirical best arm $\hat{a} = \operatorname{argmax}_a \hat{\mu}_a(Am)$
- ▶ keep playing this arm until round T

$$A_{t+1} = \hat{a} \text{ for } t \geq Am$$

⇒ EXPLORATION followed by EXPLOITATION

Given $m \in \{1, \dots, T/A\}$,

- ▶ draw each arm m times
- ▶ compute the empirical best arm $\hat{a} = \operatorname{argmax}_a \hat{\mu}_a(Am)$
- ▶ keep playing this arm until round T

$$A_{t+1} = \hat{a} \text{ for } t \geq Am$$

⇒ EXPLORATION followed by EXPLOITATION

Analysis: 2 arms, $\mu_1 > \mu_2$. $\Delta = \mu_1 - \mu_2$.

$$\mathcal{R}_T(\nu, \pi) = \Delta \times \mathbb{E}[N_2(T)]$$

$$\begin{aligned} N_2(T) &= m + (T - 2m)\mathbb{I}_{(\hat{a}=2)} \\ \mathbb{E}[N_2(T)] &\leq m + (T - 2m)\mathbb{P}(\hat{\mu}_1(2m) < \hat{\mu}_2(2m)) \\ &\leq m + T \exp\left(-\frac{m\Delta^2}{2}\right) \quad (\text{Hoeffding's inequality}) \end{aligned}$$

Given $m \in \{1, \dots, T/A\}$,

- ▶ draw each arm m times

Given $m \in \{1, \dots, T/A\}$,

- ▶ draw each arm m times
- ▶ compute the empirical best arm $\hat{a} = \operatorname{argmax}_a \hat{\mu}_a(Am)$

Given $m \in \{1, \dots, T/A\}$,

- ▶ draw each arm m times
- ▶ compute the empirical best arm $\hat{a} = \operatorname{argmax}_a \hat{\mu}_a(Am)$
- ▶ keep playing this arm until round T

$$A_{t+1} = \hat{a} \text{ for } t \geq Am$$

⇒ EXPLORATION followed by EXPLOITATION

Given $m \in \{1, \dots, T/A\}$,

- ▶ draw each arm m times
- ▶ compute the empirical best arm $\hat{a} = \operatorname{argmax}_a \hat{\mu}_a(Am)$
- ▶ keep playing this arm until round T

$$A_{t+1} = \hat{a} \text{ for } t \geq Am$$

⇒ EXPLORATION followed by EXPLOITATION

Analysis: 2 arms, $\mu_1 > \mu_2$. $\Delta = \mu_1 - \mu_2$.

$$\mathcal{R}_T(\nu, \pi) \leq \underbrace{\Delta m}_{\text{increases with } m} + \underbrace{\Delta T \exp\left(-\frac{m\Delta^2}{2}\right)}_{\text{decreases with } m}$$

A good choice: $m = \left\lfloor \frac{2}{\Delta^2} \ln \left(\frac{T\Delta^2}{2} \right) \right\rfloor$

Given $m \in \{1, \dots, T/A\}$,

- ▶ draw each arm m times
- ▶ compute the empirical best arm $\hat{a} = \operatorname{argmax}_a \hat{\mu}_a(Am)$
- ▶ keep playing this arm until round T

$$A_{t+1} = \hat{a} \text{ for } t \geq Am$$

⇒ EXPLORATION followed by EXPLOITATION

Analysis: 2 arms, $\mu_1 > \mu_2$. $\Delta = \mu_1 - \mu_2$.

$$\mathcal{R}_T(\nu, \pi) \leq \underbrace{\Delta m}_{\text{increases with } m} + \underbrace{\Delta T \exp\left(-\frac{m\Delta^2}{2}\right)}_{\text{decreases with } m}$$

A good choice: $m = \left\lfloor \frac{2}{\Delta^2} \ln \left(\frac{T\Delta^2}{2} \right) \right\rfloor$

⇒ requires the knowledge of $\Delta = \mu_1 - \mu_2$!

- We want to play $\text{Argmax}\{\mu_a, a \in \mathcal{A}\}$ but μ_a is **unknown**.

$$\mu_a = \hat{\mu}_a(t) + \underbrace{(\mu_a - \hat{\mu}_a(t))}_{\text{error term}} .$$

- **Bound** the error term and play a **penalized** strategy instead.

- We want to play $\text{Argmax}\{\mu_a, a \in \mathcal{A}\}$ but μ_a is **unknown**.

$$\mu_a = \hat{\mu}_a(t) + \underbrace{(\mu_a - \hat{\mu}_a(t))}_{\text{error term}}.$$

- **Bound** the error term and play a **penalized** strategy instead.
- Tool: Use Hoeffding inequality (plus Union bound).

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

First strategies

Confidence bounds

The optimism principle

Performance bounds

Best-achievable regret bounds

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

Lemma (Hoeffding's inequality)

For n i.i.d. random variables $X_i \in [0, 1]$ with mean μ , we have

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n X_i - \mu \geq \sqrt{\frac{\ln(1/\delta)}{2n}}\right) \leq \delta$$

$$\mathbb{P}\left(\mu - \frac{1}{n} \sum_{i=1}^n X_i \geq \sqrt{\frac{\ln(1/\delta)}{2n}}\right) \leq \delta.$$

- ▷ Non-asymptotic version of the Law of Large Numbers.
- ▷ Suggests to use $B_t(a) = \sqrt{\frac{\ln(1/\delta)}{2N_t(a)}}$ for arm a .

CONCENTRATION INEQUALITIES

Finite sample guarantee I:

$$\mathbb{P}\left[\underbrace{\frac{1}{n} \sum_{t=1}^n X_t - \mathbb{E}[X_1]}_{\text{deviation}} > \underbrace{\varepsilon}_{\text{accuracy}}\right] \leq \underbrace{\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right)}_{\text{confidence}}$$

Finite sample guarantee II:

$$\mathbb{P}\left[\frac{1}{n} \sum_{t=1}^n X_t - \mathbb{E}[X_1] > (b-a)\sqrt{\frac{\ln 1/\delta}{2n}}\right] \leq \delta$$

Finite sample guarantee III:

$$\mathbb{P}\left[\frac{1}{n} \sum_{t=1}^n X_t - \mathbb{E}[X_1] > \varepsilon\right] \leq \delta \text{ if } n \geq \frac{(b-a)^2 \ln 1/\delta}{2\varepsilon^2}$$

▶ skip

Proposition(Chernoff-Hoeffding Inequality)

Let $X_i \in [a_i, b_i]$ be n **independent** r.v. with mean $\mu_i = \mathbb{E}X_i$. Then

$$\mathbb{P}\left[\left|\sum_{i=1}^n (X_i - \mu_i)\right| \geq \varepsilon\right] \leq 2 \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Hoeffding Lemma

Let $X_i \in [a_i, b_i]$ be n **independent** r.v. with mean $\mu_i = \mathbb{E}X_i$. Then

$$\forall \lambda, \quad \mathbb{E}[e^{\lambda(X_i - \mu_i)}] \leq e^{\lambda^2(b_i - a_i)^2/8}$$

For any $\lambda > 0$

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n X_i - \mu_i \geq \varepsilon\right) &= \mathbb{P}(e^{\lambda \sum_{i=1}^n X_i - \mu_i} \geq e^{\lambda \varepsilon}) \\ &\leq e^{-\lambda \varepsilon} \mathbb{E}[e^{\lambda \sum_{i=1}^n X_i - \mu_i}], \quad \text{Markov inequality} \\ &= e^{-\lambda \varepsilon} \prod_{i=1}^n \mathbb{E}[e^{\lambda(X_i - \mu_i)}], \quad \text{Independence} \\ &\leq e^{-\lambda \varepsilon} \prod_{i=1}^n e^{\lambda^2(b_i - a_i)^2/8}, \quad \text{Hoeffding lemma} \\ &= e^{-\lambda \varepsilon + \lambda^2 \sum_{i=1}^n (b_i - a_i)^2 / 8}\end{aligned}$$

Similar arguments hold for $\mathbb{P}(\sum_{i=1}^n X_i - \mu_i \leq -\varepsilon)$.

Choosing $\lambda = 4\varepsilon / \sum_{i=1}^n (b_i - a_i)^2$ concludes.

- ▷ Consider $Y_t \in [0, 1] = [a, b]$. Suggests to use $\operatorname{argmax}_{a \in \mathcal{A}} \hat{\mu}_a(t) + \sqrt{\frac{\ln(1/\delta)}{2N_t(a)}}$.

- ▷ Consider $Y_t \in [0, 1] = [a, b]$. Suggests to use $\operatorname{argmax}_{a \in \mathcal{A}} \hat{\mu}_a(t) + \sqrt{\frac{\ln(1/\delta)}{2N_t(a)}}$.

Caveat: $N_t(a)$ is a **random variable** depending on past, not deterministic.

$$\mathbb{P}\left(\mu_a - \hat{\mu}_a(t) \geq \sqrt{\frac{\ln(1/\delta)}{2N_t(a)}}\right) \leq \delta.$$

- ▷ Consider $Y_t \in [0, 1] = [a, b]$. Suggests to use $\operatorname{argmax}_{a \in \mathcal{A}} \hat{\mu}_a(t) + \sqrt{\frac{\ln(1/\delta)}{2N_t(a)}}$.

Caveat: $N_t(a)$ is a **random variable** depending on past, not deterministic.

$$\mathbb{P}\left(\mu_a - \hat{\mu}_a(t) \geq \sqrt{\frac{\ln(1/\delta)}{2N_t(a)}}\right) \leq \delta.$$

- ▷ Fix:
 - ▶ Simple: **Union bound** argument over all possible values of $N_t(a) \in [0, t]$.
 - ▶ More elaborate: **time-uniform** confidence bounds.

We introduce

- Time of k^{th} pull of arm a : $\tau_{a,k} = \min\{t \in \mathbb{N} : N_t(a) = k\}$

We introduce

- ▶ Time of k^{th} pull of arm a : $\tau_{a,k} = \min\{t \in \mathbb{N} : N_t(a) = k\}$
- ▶ k^{th} sample from arm a : $X_{a,k} = Y_{\tau_{a,k}}$

We introduce

- ▶ Time of k^{th} pull of arm a : $\tau_{a,k} = \min\{t \in \mathbb{N} : N_t(a) = k\}$
- ▶ k^{th} sample from arm a : $X_{a,k} = Y_{\tau_{a,k}}$
- ▶ **Mean** of n first samples from arm a : $\mu_{a,n} = \frac{1}{n} \sum_{k=1}^n X_{a,k}$

Then

- ▶ For each k , $\tau_{a,k}$ is a (predictable) stopping time.

We introduce

- ▶ Time of k^{th} pull of arm a : $\tau_{a,k} = \min\{t \in \mathbb{N} : N_t(a) = k\}$
- ▶ k^{th} sample from arm a : $X_{a,k} = Y_{\tau_{a,k}}$
- ▶ **Mean** of n first samples from arm a : $\mu_{a,n} = \frac{1}{n} \sum_{k=1}^n X_{a,k}$

Then

- ▶ For each k , $\tau_{a,k}$ is a (predictable) stopping time.
- ▶ The $(X_{a,k})_k$ are **i.i.d.** according to ν_a .

We introduce

- ▶ Time of k^{th} pull of arm a : $\tau_{a,k} = \min\{t \in \mathbb{N} : N_t(a) = k\}$
- ▶ k^{th} sample from arm a : $X_{a,k} = Y_{\tau_{a,k}}$
- ▶ **Mean** of n first samples from arm a : $\mu_{a,n} = \frac{1}{n} \sum_{k=1}^n X_{a,k}$

Then

- ▶ For each k , $\tau_{a,k}$ is a (predictable) stopping time.
- ▶ The $(X_{a,k})_k$ are **i.i.d.** according to ν_a .
- ▶ Further $\hat{\mu}_a(t) = \mu_{a,N_t(a)}$ almost surely.

Finally, by a union bound over all possible values of $N_t(a) \in [0, t]$.

$$\begin{aligned} & \mathbb{P}\left(\mu_a - \hat{\mu}_a(t) \geq \sqrt{\frac{\ln(1/\delta)}{2N_t(a)}}\right) \\ & \leq \mathbb{P}\left(\exists n \in \{0, \dots, t\}, \mu_a - \mu_{a,n} \geq \sqrt{\frac{\ln(1/\delta)}{2n}}\right) \leq t\delta. \end{aligned}$$

using that $\mathbb{P}(\mu_a - \mu_{a,0} \geq \infty) = \mathbb{P}(\mu_a \geq \infty) = 0$.

WARNING

Caveat to remember

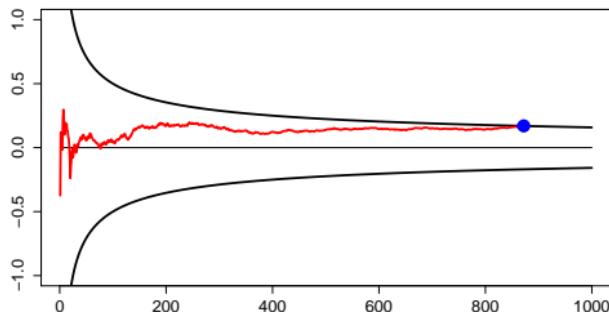
$$\mathbb{P}\left(\mu_a \geq \hat{\mu}_a(t) + \sqrt{\frac{\ln(1/\delta)}{2N_t(a)}}\right) \not\leq \delta.$$

$$\mathbb{P}\left(\mu_a \geq \hat{\mu}_a(t) + \sqrt{\frac{\ln(1/\delta)}{2N_t(a)}}\right) \leq t\delta.$$

SEQUENTIAL EXPLORE-THEN-EXPLOIT (2 ARMS)

- ▶ Explore uniformly until the **random time**

$$\tau = \inf \left\{ t \in \mathbb{N} : |\hat{\mu}_1(t) - \hat{\mu}_2(t)| > \sqrt{\frac{4 \ln(T/t)}{t}} \right\}$$



- ▶ $A_{t+1} = \operatorname{argmax}_a \hat{\mu}_a(\tau)$ for $t \in \{\tau, \dots, T\}$

$$\mathcal{R}_T(\nu, \pi) \leq \frac{2}{\Delta} \ln(T) + C \sqrt{\ln(T)}.$$

- ⇒ same regret rate, without knowing Δ
- ⇒ but requires the knowledge of T ...

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

First strategies

Confidence bounds

The optimism principle

Performance bounds

Best-achievable regret bounds

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

The Upper Confidence Bound algorithm

Auer, P.; Cesa-Bianchi, N. & Fischer, P.

Finite-time analysis of the multiarmed bandit problem *Machine Learning*,
Springer, 2002, 47, 235-256

Optimism in Face of Uncertainty Learning (OFUL)

Whenever we are **uncertain** about the outcome of an arm, we consider the **best possible world** and choose the **best arm**.

Optimism in Face of Uncertainty Learning (OFUL)

Whenever we are **uncertain** about the outcome of an arm, we consider the **best possible world** and choose the **best arm**.

$$\operatorname{argmax}_{a \in \mathcal{A}} \max \left\{ \mathbb{E}_{\tilde{\nu}_a}[X] : \tilde{\nu}_a \text{ compatible with obs. on arm } a \right\}$$

Optimism in Face of Uncertainty Learning (OFUL)

Whenever we are **uncertain** about the outcome of an arm, we consider the **best possible world** and choose the **best arm**.

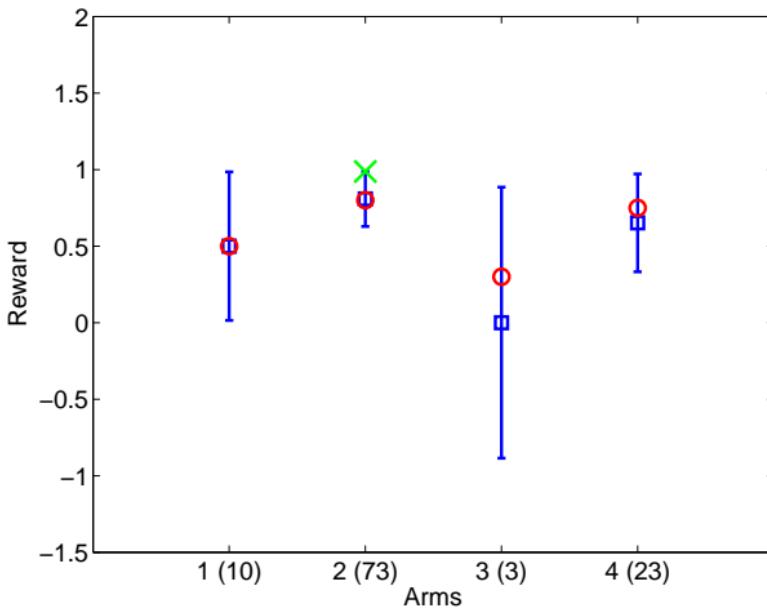
$$\operatorname{argmax}_{a \in \mathcal{A}} \max \left\{ \mathbb{E}_{\tilde{\nu}_a}[X] : \tilde{\nu}_a \text{ compatible with obs. on arm } a \right\}$$

Why it works:

- ▶ If the *best possible world* is correct \Rightarrow **no regret**
- ▶ If the *best possible world* is wrong \Rightarrow **the reduction in the uncertainty is maximized**

THE UPPER-CONFIDENCE BOUND (UCB) ALGORITHM

The idea



The **Upper Confidence Bound** algorithm (Auer et al. 2002)

Choose $A_{t+1} = \text{Argmax}\{\mu_{a,t}^+, a \in \mathcal{A}\}$ where

$$\mu_{a,t}^+ = \tilde{\mu}_{a,t} + \sqrt{\frac{\ln(1/\delta_t)}{2N_a(t)}}.$$

with δ_t such that $\sum_t t\delta_t < \infty$.

Intuition: UCB should pull the suboptimal arms

- ▶ **Enough:** so as to understand which arm is the best
- ▶ **Not too much:** so as to keep the regret as small as possible

Intuition: UCB should pull the suboptimal arms

- ▶ **Enough:** so as to understand which arm is the best
- ▶ **Not too much:** so as to keep the regret as small as possible

The confidence $1 - \delta$ has the following impact (similar for α)

- ▶ **Big $1 - \delta$:** high level of **exploration**
- ▶ **Small $1 - \delta$:** high level of **exploitation**

Intuition: UCB should pull the suboptimal arms

- ▶ **Enough:** so as to understand which arm is the best
- ▶ **Not too much:** so as to keep the regret as small as possible

The confidence $1 - \delta$ has the following impact (similar for α)

- ▶ **Big $1 - \delta$:** high level of **exploration**
- ▶ **Small $1 - \delta$:** high level of **exploitation**

Solution: depending on the time horizon, we can tune how to trade-off between exploration and exploitation

The **U**pper **C**onfidence **B**ound algorithm (Auer et al. 2002)

Choose $A_{t+1} = \text{Argmax}\{\mu_{a,t}^+, a \in \mathcal{A}\}$ where

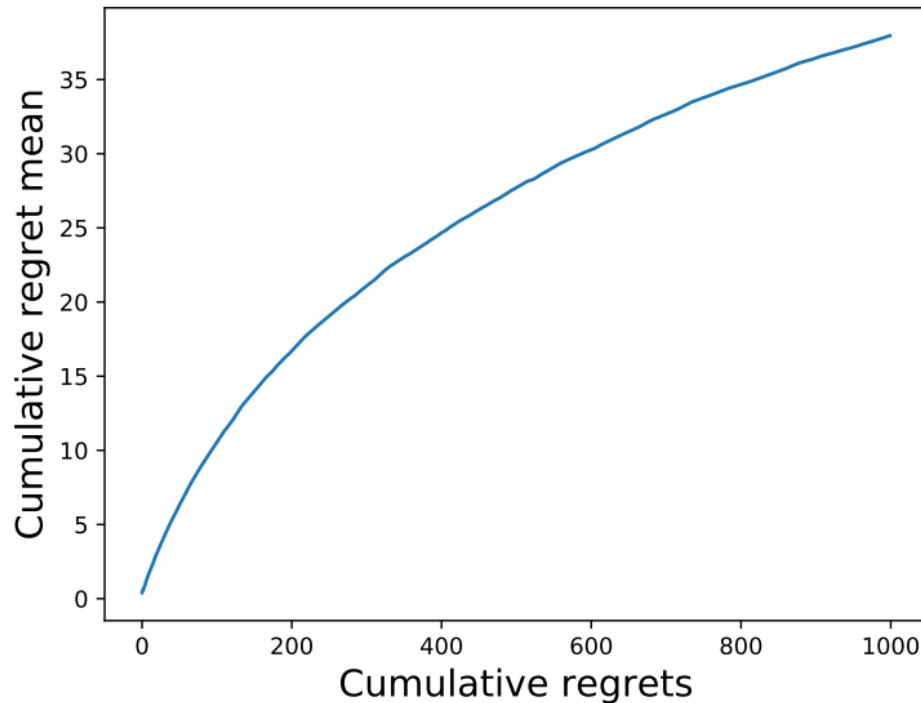
$$\mu_{a,t}^+ = \hat{\mu}_a(t) + \sqrt{\frac{\ln(1/\delta_t)}{2N_a(t)}}.$$

with δ_t such that $\sum_t t\delta_t < \infty$.

- ▶ Choice $\delta_t = t^{-3}$ gives for each $a \in \mathcal{A}$, $t > A$,

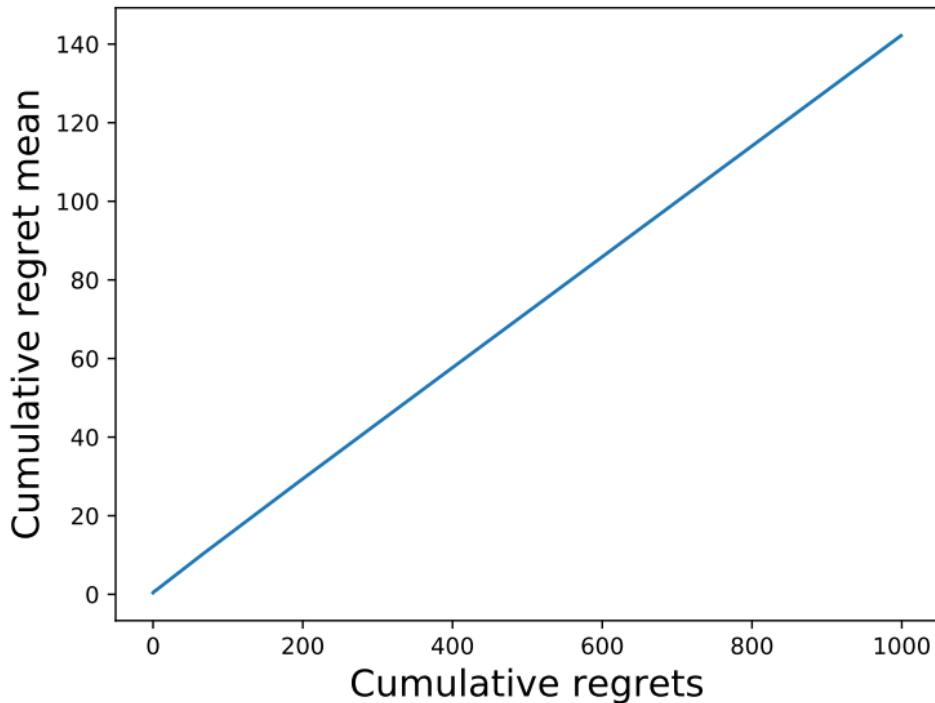
$$\mathbb{P}\left(\mu_a - \hat{\mu}_a(t) \geq \sqrt{\frac{\ln(1/\delta_t)}{2N_t(a)}}\right) \leq \frac{1}{t^2}.$$

REGRET OF UCB FOR A $[\mathcal{B}(0.2), \mathcal{B}(0.4), \mathcal{B}(0.6)]$ -BANDIT



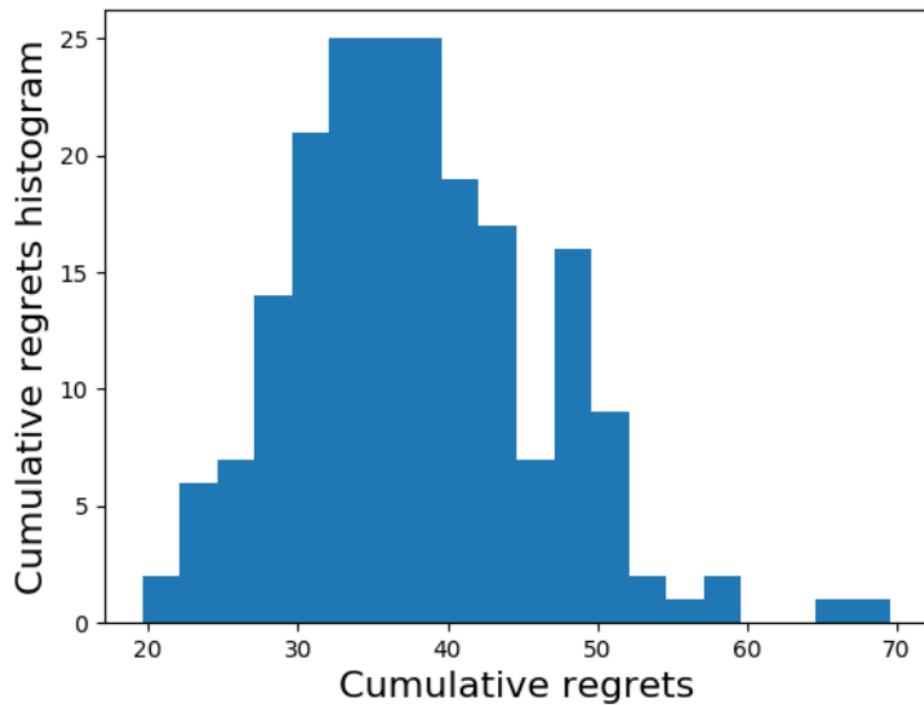
Results averaged over 200 runs.

REGRET OF FTL FOR A $[\mathcal{B}(0.2), \mathcal{B}(0.4), \mathcal{B}(0.6)]$ -BANDIT

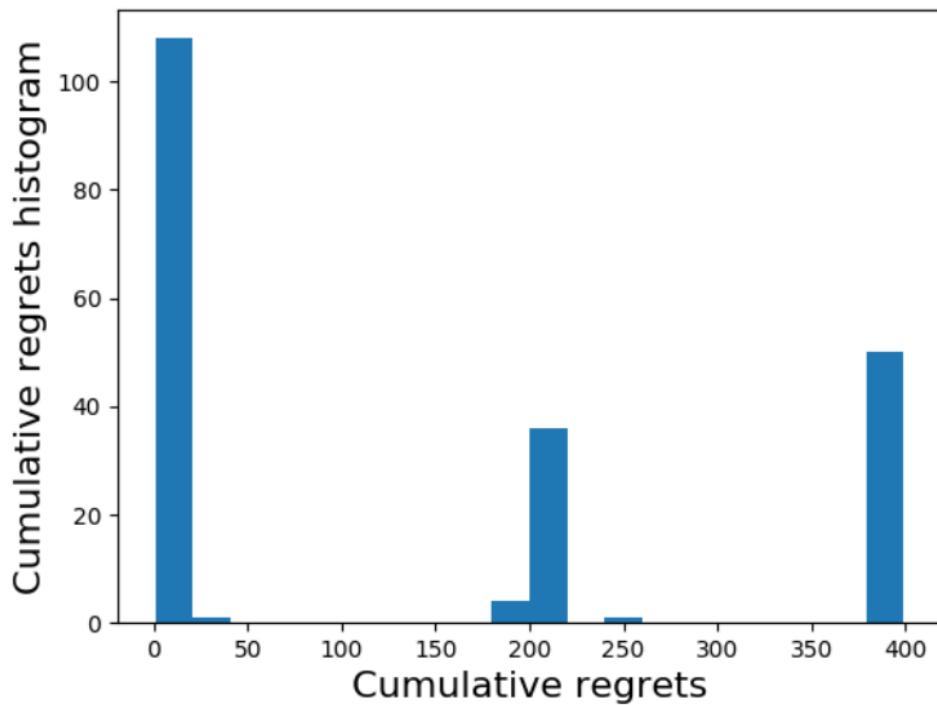


Results averaged over 200 runs.

REGRET OF UCB FOR A $[\mathcal{B}(0.2), \mathcal{B}(0.4), \mathcal{B}(0.6)]$ -BANDIT



REGRET OF FTL FOR A $[\mathcal{B}(0.2), \mathcal{B}(0.4), \mathcal{B}(0.6)]$ -BANDIT



$$\mu_{a,t}^+ = \hat{\mu}_a(t) + \sqrt{\frac{\ln(1/\delta_t)}{2N_a(t)}}.$$

Exploitation: "Follow current knowledge"

Choose arm with highest $\hat{\mu}_a(t)$

Exploration: Maximally improve current knowledge

Choose least known arm: arm with smallest $N_a(t)$.

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

First strategies

Confidence bounds

The optimism principle

Performance bounds

Best-achievable regret bounds

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

- Assume rewards generated by ν are **bounded** in $[0, 1]$.

Theorem (Distribution-dependent regret bounds for UCB)

In the stochastic multi-armed bandit game, the UCB strategy with $\delta_t = t^{-2}(t+1)^{-1}$ satisfies the following performance bound.

$$\mathcal{R}_T(\nu, \text{UCB}) \leq \sum_{a; \Delta_a > 0} \left[\frac{6}{\Delta_a} \ln(T) + 3\Delta_a \right]$$

Scaling in $\sum_{a; \Delta_a > 0} \frac{\ln(T)}{\Delta_a}$ without knowing Δ, T .

▶ skip proof

1. Concentration inequalities

From $\hat{\mu}_a(t) = \mu_{a,N_t(a)}$ and $N_t(a) \leq t$, we deduce

$$\begin{aligned} \mathbb{P}\left(\hat{\mu}_a(t) - \mu_a \geq \sqrt{\frac{\ln(1/\delta_t)}{2N_t(a)}}\right) &\leq \mathbb{P}\left(\exists n \leq t, \mu_{a,n} - \mu_a \geq \sqrt{\frac{\ln(1/\delta_t)}{2n}}\right) \\ &\stackrel{(a)}{\leq} \sum_{n=1}^t \mathbb{P}\left(\mu_{a,n} - \mu_a \geq \sqrt{\frac{\ln(1/\delta_t)}{2n}}\right) \\ &\stackrel{(b)}{\leq} \sum_{n=1}^t \frac{1}{t^2(t+1)} = \frac{1}{t(t+1)}. \end{aligned}$$

(a) is by a union bound argument, (b) is by Hoeffding inequality.

2. Algorithm mechanism

If $A_{t+1} = a$ is sub-optimal, by definition of the chosen arm

$$\hat{\mu}_a(t) + \sqrt{\frac{\ln(1/\delta_t)}{2N_t(a)}} \geq \hat{\mu}_*(t) + \sqrt{\frac{\ln(1/\delta_t)}{2N_t(*)}}.$$

Now, on an event of probability higher than $1 - \frac{1}{t(t+1)}$,

$$\hat{\mu}_*(t) + \sqrt{\frac{\ln(1/\delta_t)}{2N_t(*)}} \geq \mu_*$$

Likewise, on an event of probability higher than $1 - \frac{1}{t(t+1)}$,

$$\mu_a + 2\sqrt{\frac{\ln(1/\delta_t)}{2N_t(a)}} \geq \hat{\mu}_a(t) + \sqrt{\frac{\ln(1/\delta_t)}{2N_t(a)}}.$$

By a union bound argument, we deduce that for each $a \in \mathcal{A}$, with probability higher than $1 - \frac{2}{t(t+1)}$, if $A_{t+1} = a$ is sub-optimal, then

$$\mu_a + 2\sqrt{\frac{\ln(1/\delta_t)}{2N_t(a)}} \geq \mu_* \quad \text{that is } \boxed{N_t(a) \leq \frac{2\ln(1/\delta_t)}{(\mu_* - \mu_a)^2}}.$$

3. Control of sub-optimal arm pull

Now, let us consider some integer u_a .

$$\begin{aligned}N_T(a) &= \sum_{t=1}^T \mathbb{I}\{A_t = a \cap N_{t-1}(a) \leq u_a\} + \sum_{t=1}^T \mathbb{I}\{A_t = a \cap N_{t-1}(a) > u_a\} \\&= \sum_{t=1}^T \mathbb{I}\{A_t = a \cap N_{t-1}(a) \leq u_a\} + \sum_{t=u_a+2}^T \mathbb{I}\{A_t = a \cap N_{t-1}(a) > u_a\} \\&\leq u_a + 1 + \sum_{t=u_a+2}^T \mathbb{I}\{A_t = a \cap N_{t-1}(a) > u_a\}\end{aligned}$$

4. Tuning threshold

Let us choose

$$u_a = \left\lceil \frac{2 \ln(1/\underline{\delta}_T)}{\Delta_a^2} \right\rceil$$

where $\underline{\delta}_T = \min_{t \leq T} \delta_{t-1} = \delta_{T-1}$. Thus,

$$\begin{aligned}\mathbb{E}[N_T(a)] &\leq u_a + 1 + \sum_{t=u_a+2}^T \mathbb{P}(A_t = a \cap N_{t-1}(a) > u_a) \\ &\leq u_a + 1 + \sum_{t=3}^T \mathbb{P}(A_t = a \cap N_{t-1}(a) > \frac{2 \ln(1/\delta_{t-1})}{\Delta_a^2}) \\ &\leq u_a + 1 + \sum_{t=3}^T \frac{2}{(t-1)t} \text{ for } \delta_t = \frac{1}{t^2(t+1)}.\end{aligned}$$

5. Summing all terms

Finally,

$$\begin{aligned}\mathcal{R}_T(\nu, UCB) &= \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[N_T(a)] \\ &\leq \sum_{a \in \mathcal{A}} \frac{2 \ln((T-1)^2 T)}{\Delta_a} + 2\Delta_a + \Delta_a \sum_{t=2}^{\infty} \frac{2}{(t+1)t} \\ &\leq \sum_{a \in \mathcal{A}} \frac{6 \ln(T)}{\Delta_a} + 2\Delta_a(1 + 1/2).\end{aligned}$$

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

First strategies

Confidence bounds

The optimism principle

Performance bounds

Best-achievable regret bounds

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

Definition (**Uniformly good** strategy)

A strategy is uniformly good on \mathcal{D} if for any stochastic bandit $\nu = (\nu_a)_{a \in \mathcal{A}} \in \mathcal{D}$,

$$a \notin \mathcal{A}_*(\nu) \implies \forall \alpha \in (0, 1) \quad \mathbb{E}_\nu[N_a(T)] = o(T^\alpha).$$

Theorem (Lai & Robbins, 1985)

Any **uniformly good** strategy on the set of **Bernoulli** bandit $\nu = (\mathcal{B}(\theta_1), \dots, \mathcal{B}(\theta_A))$ with means $\theta_a < 1$ must satisfy:

$$a \notin \mathcal{A}_*(\nu) \implies \liminf_{T \rightarrow \infty} \frac{\mathbb{E}[N_a(T)]}{\ln(T)} \geq \frac{1}{\text{KL}(\theta_a, \theta_*)}.$$

Since $\mathcal{R}_\nu(T) = \sum_{a: \Delta_a > 0} \Delta_a \mathbb{E}[N_a(T)]$,
 Thus $\liminf_{T \rightarrow \infty} \frac{\mathcal{R}_\nu(T)}{\ln(T)} \geq \sum_{a: \Delta_a > 0} \frac{\mu_* - \mu_a}{\text{KL}(\theta_a, \theta_*)}.$

We only need to study the **expected number of pulls** of **suboptimal** arms

▷ **Most confusing** environment:

For $a \notin \mathcal{A}_*(\nu)$, find $\tilde{\nu}$ such that $a = \mathcal{A}_*(\tilde{\nu})$ but $\nu_{a_*} = \tilde{\nu}_{a_*}$

▷ **Change of measure** argument:

$$(\text{Probability}) \quad \forall \Omega, \forall c \in \mathbb{R}, \quad \mathbb{P}_\nu \left(\Omega \cap \left\{ \ln \left(\frac{d\nu}{d\tilde{\nu}}(X) \right) \leq c \right\} \right) \leq \exp(c) \mathbb{P}_{\tilde{\nu}}(\Omega).$$

$$(\text{Expectation}) \quad \mathbb{E}_\nu \left[\ln \left(\frac{d\nu}{d\tilde{\nu}}(X) \right) \right] \geq \sup_{g: \mathcal{X} \rightarrow [0,1]} \text{kl} \left(\mathbb{E}_\nu[g(X)], \mathbb{E}_{\tilde{\nu}}[g(X)] \right).$$

 skip proof

1. Reduction

$$\frac{\mathbb{E}[N_a(T)]}{\ln(T)} \geq c \mathbb{P}_\nu(N_a(T) \geq c \ln(T)) \quad (\text{Markov inequality})$$

Study $\Omega = \{N_T(a) < c \ln(T)\}$. Show that $\mathbb{P}_\nu(\Omega) \rightarrow 0$ with T .

2. Confusing instance

Let $\tilde{\nu} = (\tilde{\theta}_1, \dots, \tilde{\theta}_A)$ be a maximally confusing instance for $a \notin \mathcal{A}^*(\nu)$

$$\begin{cases} \tilde{\theta}_{a'} = \theta_{a'} & \text{if } a' \neq a \\ \tilde{\theta}_a = \lambda & \text{where } \lambda > \mu_* \text{ (hence } a \in \mathcal{A}_*(\tilde{\nu})) \end{cases}$$

3. (Bernoulli) **log-Likelihood** threshold

Let $\mathcal{E} = \{\mathcal{L}_{N_a(T)} \leq (1 - \alpha) \ln(T)\}$
where $\mathcal{L}_m = \sum_{j=1}^m \ln \left(\frac{d\nu_{\theta_a}}{d\nu_\lambda}(X_{a,j}) \right)$ with $d\nu_\theta(x) = \theta^x(1-\theta)^{1-x}$.

$$\begin{aligned}\mathbb{P}_\nu(\Omega \cap \mathcal{E}) &= \mathbb{E}_\nu \left(e^{\ln \left(\frac{d\nu}{d\nu}(Y) \right)} \mathbb{I}\{\Omega \cap E\} \right) \\ &\leq T^{1-\alpha} \mathbb{P}_{\tilde{\nu}}(\Omega \cap \mathcal{E}) \quad (\text{Change of measure})\end{aligned}$$

$$\begin{aligned}\mathbb{P}_\nu(\Omega \cap \mathcal{E}) &\leq T^{1-\alpha} \mathbb{P}_{\tilde{\nu}} \left(\sum_{a' \neq a} N_{a'}(T) > T - c \ln(T) \right) \quad \left(\sum_{a'} N_{a'}(T) = T \right) \\ &\leq T^{1-\alpha} \frac{\sum_{a' \neq a} \mathbb{E}_{\tilde{\nu}}[N_{a'}(T)]}{T - c \ln(T)} \quad (\text{Markov inequality}) \\ &= o(1) \quad (\text{Consistency for } \tilde{\nu})\end{aligned}$$

4. (Maximal) concentration inequality

$$\begin{aligned}\mathbb{P}_\nu(\Omega \cap \mathcal{E}^c) &\leq \mathbb{P}_\nu\left(\exists m < c \ln(T) : \sum_{j=1}^m \underbrace{\ln\left(\frac{d\nu_{\theta_a}(X_{a,j})}{d\nu_\lambda(X_{a,j})}\right)}_{Z_j} > (1-\alpha)\ln(T)\right). \\ &= \mathbb{P}_\nu\left(\frac{\max_{m < c \ln(T)} \sum_{j=1}^m Z_j}{c \ln(T)} > \frac{1-\alpha}{c \text{kl}(\theta_a, \lambda)} \underbrace{\text{kl}(\theta_a, \lambda)}_{\mathbb{E}_\theta[Z_j]}\right)\end{aligned}$$

Lemma (Asymptotic maximal Hoeffding inequality)

For any i.i.d. bounded Z_j with **positive** mean μ ,

$$\forall \eta > 0, \lim_{n \rightarrow \infty} \mathbb{P}_\nu\left(\frac{\max_{m < n} \sum_{j=1}^m Z_j}{n} > (1 + \eta)\mu\right) = 0.$$

\implies e.g. $c = \frac{1 - 2\alpha}{\text{kl}(\theta_a, \lambda)}$ to conclude.

We make use of the **fundamental lemma** for change of measure:

(Kaufmann, PhD), (Garivier et al. 2016), (Wald 1945)

For a (random) sequence generated by a sequential sampling policy,

$$\text{KL}(\nu_{\mathbf{a}}, \tilde{\nu}_{\mathbf{a}}) = \sum_{a' \in \mathcal{A}} \mathbb{E}_{\nu}[N_{a'}(T)] \text{KL}(\nu_{a'}, \tilde{\nu}_{a'}) \geq \sup_{\Omega} \text{kl}(\mathbb{P}_{\nu}[\Omega], \mathbb{P}_{\tilde{\nu}}[\Omega]).$$

where $\text{kl}(x, y) = \text{KL}(\mathcal{B}(x), \mathcal{B}(y))$.

► Interpretation:

The divergence between two processes is always as high as the divergence of any test that can be built from these processes.

Hence $\forall a \notin \mathcal{A}^*(\nu)$

$$\mathbb{E}_\nu[N_a(T)] \geq \sup_{\Omega, \tilde{\nu}} \frac{\text{kl}(\mathbb{P}_\nu[\Omega], \mathbb{P}_{\tilde{\nu}}[\Omega]) - \sum_{a' \neq a} \text{KL}(\nu_{a'}, \tilde{\nu}_{a'}) \mathbb{E}_\theta[N_{a'}(T)]}{\text{KL}(\nu_a, \tilde{\nu}_a)}.$$

Hence $\forall a \notin \mathcal{A}^*(\nu)$

$$\mathbb{E}_\nu[N_a(T)] \geq \sup_{\Omega, \tilde{\nu}} \frac{\text{kl}(\mathbb{P}_\nu[\Omega], \mathbb{P}_{\tilde{\nu}}[\Omega]) - \sum_{a' \neq a} \text{KL}(\nu_{a'}, \tilde{\nu}_{a'}) \mathbb{E}_\theta[N_{a'}(T)]}{\text{KL}(\nu_a, \tilde{\nu}_a)}.$$

Choose $\tilde{\nu}$ such that $\mathcal{A}^*(\tilde{\nu}) = \{a\}$, $\Omega = \{N_a(T) > T^\alpha\}$:

- ▶ $\mathbb{P}_\nu[\Omega] \leq \mathbb{E}_\nu[N_a(T)] T^{-\alpha} = o(1)$
- ▶ $\text{kl}(\mathbb{P}_\nu[\Omega], \mathbb{P}_{\tilde{\nu}}[\Omega]) \simeq \ln\left(\frac{1}{\mathbb{P}_{\tilde{\nu}}(N_T(a) \leq T^\alpha)}\right) \geq \ln\left(\frac{T - T^\alpha}{\sum_{a' \neq a} \mathbb{E}_{\tilde{\nu}}[N_T(a')]} \right) \simeq \ln(T).$
- ▶ Choose $\tilde{\nu}_{a'}$ for $a' \neq a$: $\tilde{\nu}_{a'} = \nu_{a'}$ (no constraint)

Hence $\forall a \notin \mathcal{A}^*(\nu)$

$$\mathbb{E}_\nu[N_a(T)] \geq \sup_{\Omega, \tilde{\nu}} \frac{\text{kl}(\mathbb{P}_\nu[\Omega], \mathbb{P}_{\tilde{\nu}}[\Omega]) - \sum_{a' \neq a} \text{KL}(\nu_{a'}, \tilde{\nu}_{a'}) \mathbb{E}_\theta[N_{a'}(T)]}{\text{KL}(\nu_a, \tilde{\nu}_a)}.$$

Choose $\tilde{\nu}$ such that $\mathcal{A}^*(\tilde{\nu}) = \{a\}$, $\Omega = \{N_a(T) > T^\alpha\}$:

- ▶ $\mathbb{P}_\nu[\Omega] \leq \mathbb{E}_\nu[N_a(T)] T^{-\alpha} = o(1)$
- ▶ $\text{kl}(\mathbb{P}_\nu[\Omega], \mathbb{P}_{\tilde{\nu}}[\Omega]) \simeq \ln\left(\frac{1}{\mathbb{P}_{\tilde{\nu}}(N_T(a) \leq T^\alpha)}\right) \geq \ln\left(\frac{T - T^\alpha}{\sum_{a' \neq a} \mathbb{E}_{\tilde{\nu}}[N_T(a')]} \right) \simeq \ln(T).$
- ▶ Choose $\tilde{\nu}_{a'}$ for $a' \neq a$: $\tilde{\nu}_{a'} = \nu_{a'}$ (no constraint)

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\ln(T)} \geq \frac{1 - 0}{\inf_{\tilde{\nu}_a} \{\text{KL}(\nu_a, \tilde{\nu}_a) : \tilde{\mu}_a > \mu_\star(\nu)\}}$$

For the proof the Weyl's lemma, consider $\theta, \theta' \in \Theta$:

$$\hat{\mathcal{L}}_T = \sum_{s=1}^T \ln \left(\frac{\nu_{\theta'_{A_s}}(Y_s)}{\nu_{\theta_{A_s}}(Y_s)} \right) = \sum_{a \in \mathcal{A}} \sum_{s=1}^T \mathbb{I}\{A_s = a\} \ln \left(\frac{\nu_{\theta'_a}(Y_s)}{\nu_{\theta_a}(Y_s)} \right)$$

For the proof the Weyl's lemma, consider $\theta, \theta' \in \Theta$:

$$\hat{\mathcal{L}}_T = \sum_{s=1}^T \ln \left(\frac{\nu_{\theta'_{A_s}}(Y_s)}{\nu_{\theta_{A_s}}(Y_s)} \right) = \sum_{a \in \mathcal{A}} \sum_{s=1}^T \mathbb{I}\{A_s = a\} \ln \left(\frac{\nu_{\theta'_a}(Y_s)}{\nu_{\theta_a}(Y_s)} \right)$$

For any event Ω it holds (**Change of measure**) (from Weyl 1940)

$$\begin{aligned} \mathbb{P}_{\theta'}[\Omega] &= \mathbb{E}_\theta[\exp(\hat{\mathcal{L}}_T)\mathbb{I}\{\Omega\}] = \mathbb{E}_\theta\left[\exp(\hat{\mathcal{L}}_T)|\Omega\right]\mathbb{P}_\theta[\Omega] \\ &\stackrel{\text{Jensen}}{\geqslant} \exp\left(\mathbb{E}_\theta[\hat{\mathcal{L}}_T|\Omega]\right)\mathbb{P}_\theta[\Omega] = \exp\left(\frac{\mathbb{E}_\theta[\hat{\mathcal{L}}_T\mathbb{I}\{\Omega\}]}{\mathbb{P}_\theta[\Omega]}\right)\mathbb{P}_\theta[\Omega], \end{aligned}$$

Reorganizing the terms, we get $-\mathbb{E}_\theta[\hat{\mathcal{L}}_T\mathbb{I}\{\Omega\}] \geqslant \mathbb{P}_\theta[\Omega] \ln \left(\frac{\mathbb{P}_\theta[\Omega]}{\mathbb{P}_{\theta'}[\Omega]} \right)$.

For the proof the Weyl's lemma, consider $\theta, \theta' \in \Theta$:

$$\widehat{\mathcal{L}}_T = \sum_{s=1}^T \ln \left(\frac{\nu_{\theta'_A_s}(Y_s)}{\nu_{\theta_A_s}(Y_s)} \right) = \sum_{a \in \mathcal{A}} \sum_{s=1}^T \mathbb{I}\{A_s = a\} \ln \left(\frac{\nu_{\theta'_a}(Y_s)}{\nu_{\theta_a}(Y_s)} \right)$$

For any event Ω it holds (**Change of measure**) (from Weyl 1940)

$$\begin{aligned} \mathbb{P}_{\theta'}[\Omega] &= \mathbb{E}_\theta[\exp(\widehat{\mathcal{L}}_T) \mathbb{I}\{\Omega\}] = \mathbb{E}_\theta \left[\exp(\widehat{\mathcal{L}}_T) | \Omega \right] \mathbb{P}_\theta[\Omega] \\ &\stackrel{\text{Jensen}}{\geq} \exp \left(\mathbb{E}_\theta[\widehat{\mathcal{L}}_T | \Omega] \right) \mathbb{P}_\theta[\Omega] = \exp \left(\frac{\mathbb{E}_\theta[\widehat{\mathcal{L}}_T \mathbb{I}\{\Omega\}]}{\mathbb{P}_\theta[\Omega]} \right) \mathbb{P}_\theta[\Omega], \end{aligned}$$

Reorganizing the terms, we get $-\mathbb{E}_\theta[\widehat{\mathcal{L}}_T \mathbb{I}\{\Omega\}] \geq \mathbb{P}_\theta[\Omega] \ln \left(\frac{\mathbb{P}_\theta[\Omega]}{\mathbb{P}_{\theta'}[\Omega]} \right)$. Likewise for the complement Ω^c . Summing up the terms, we obtain

$$\begin{aligned} -\mathbb{E}_\theta[\widehat{\mathcal{L}}_T] &= \sum_{a \in \mathcal{A}} \mathbb{E}_\theta[N_T(a)] \mathbf{KL}(\theta_a, \theta'_a) \\ &\geq \mathbb{P}_\theta[\Omega] \ln \left(\frac{\mathbb{P}_\theta[\Omega]}{\mathbb{P}_{\theta'}[\Omega]} \right) + (1 - \mathbb{P}_\theta[\Omega]) \ln \left(\frac{1 - \mathbb{P}_\theta[\Omega]}{1 - \mathbb{P}_{\theta'}[\Omega]} \right). \end{aligned}$$

For the proof the Weyl's lemma, consider $\theta, \theta' \in \Theta$:

$$\widehat{\mathcal{L}}_T = \sum_{s=1}^T \ln \left(\frac{\nu_{\theta'_{A_s}}(Y_s)}{\nu_{\theta_{A_s}}(Y_s)} \right) = \sum_{a \in \mathcal{A}} \sum_{s=1}^T \mathbb{I}\{A_s = a\} \ln \left(\frac{\nu_{\theta'_a}(Y_s)}{\nu_{\theta_a}(Y_s)} \right)$$

For any event Ω it holds (**Change of measure**) (from Weyl 1940)

$$\begin{aligned} \mathbb{P}_{\theta'}[\Omega] &= \mathbb{E}_\theta[\exp(\widehat{\mathcal{L}}_T) \mathbb{I}\{\Omega\}] = \mathbb{E}_\theta \left[\exp(\widehat{\mathcal{L}}_T) | \Omega \right] \mathbb{P}_\theta[\Omega] \\ &\stackrel{\text{Jensen}}{\geq} \exp \left(\mathbb{E}_\theta[\widehat{\mathcal{L}}_T | \Omega] \right) \mathbb{P}_\theta[\Omega] = \exp \left(\frac{\mathbb{E}_\theta[\widehat{\mathcal{L}}_T \mathbb{I}\{\Omega\}]}{\mathbb{P}_\theta[\Omega]} \right) \mathbb{P}_\theta[\Omega], \end{aligned}$$

Reorganizing the terms, we get $-\mathbb{E}_\theta[\widehat{\mathcal{L}}_T \mathbb{I}\{\Omega\}] \geq \mathbb{P}_\theta[\Omega] \ln \left(\frac{\mathbb{P}_\theta[\Omega]}{\mathbb{P}_{\theta'}[\Omega]} \right)$. Likewise for the complement Ω^c . Summing up the terms, we obtain

$$\boxed{\sum_{a \in \mathcal{A}} \mathbb{E}_\theta[N_T(a)] \text{KL}(\theta_a, \theta'_a) \geq \text{kl}(\mathbb{P}_\theta[\Omega], \mathbb{P}_{\theta'}[\Omega])}$$

Set of optimal arms for $\nu = (\nu_a)_{a \in \mathcal{A}}$: $\mathcal{A}_*(\nu) = \text{Argmax}_{a \in \mathcal{A}} \mu_a(\nu)$.

Definition (**Uniformly Good strategies**)

A bandit strategy is **uniformly-good** on \mathcal{D} if

$$\forall \nu = (\nu_a)_{a \in \mathcal{A}} \in \mathcal{D}, \forall a \notin \mathcal{A}_*(\nu), \quad \mathbb{E}[N_T(a)] = o(T^\alpha) \quad \text{for all } \alpha \in (0, 1].$$

Theorem ((Lai, Robbins 85) “Price for being uniformly-good”)

Any uniformly good strategy on $\mathcal{D} = \text{Bern}^{\mathcal{A}}$ must satisfy

$$\forall a \notin \mathcal{A}_*(\nu) \quad \liminf_{T \rightarrow \infty} \frac{\mathbb{E}_\nu[N_T(a)]}{\ln(T)} \geq \frac{1}{\text{kl}(\mu_a(\nu), \mu_*(\nu))}.$$

This generalizes beyond Bernoulli distributions:

Lower bound (Burnetas & Katehakis, 96)

Any uniformly good strategy on a product set $\mathcal{D} \in \otimes_{a \in \mathcal{A}} \mathcal{D}_a$ of distributions (under mild assumptions) must satisfy

$$\forall a : \mu_a < \mu_\star \quad \liminf_{T \rightarrow \infty} \frac{\mathbb{E}[N_a(T)]}{\ln T} \geq \frac{1}{\mathcal{K}_a(\nu_a, \mu_\star)},$$

$$\mathcal{K}_a(\nu_a, \mu_\star) = \inf \{ \text{KL}(\nu_a, \nu) : \nu \in \mathcal{D}, \mathbb{E}_\nu[X] > \mu_\star \}$$

$$\liminf_{T \rightarrow \infty} \frac{\mathfrak{R}_T}{\ln T} \geq \sum_{a \in \mathcal{A}} \frac{\Delta_a}{\mathcal{K}_a(\nu_a, \mu_\star)}$$

- ▶ Even though the initial problem involves **means only**, the lower bound depend on the full **distributions**.

- ▷ Insight from lower bound: Any **uniformly-good** strategy on \mathcal{D} must satisfy:

$$\forall a \notin \mathcal{A}_*(\nu), \liminf_T \frac{\mathbb{E}[N_T(a)]}{\ln(T)} \geq \sup \left\{ \frac{1}{\text{KL}(\nu_a, \tilde{\nu}_a)} : \underbrace{\tilde{\nu} = (\nu_1, \dots, \tilde{\nu}_a, \dots, \nu_A), \mathcal{A}_*(\tilde{\nu}) = \{a\}}_{\text{most confusing (unstructured)}} \right\}$$

THE OPTIMISTIC PRINCIPLE REVISITED

- ▷ Insight from lower bound: Any **uniformly-good** strategy on \mathcal{D} must satisfy:

$$\forall a \notin \mathcal{A}_*(\nu), \liminf_T \frac{\mathbb{E}[N_T(a)]}{\ln(T)} \geq \sup \left\{ \frac{1}{\underbrace{\text{KL}(\nu_a, \tilde{\nu}_a)}_{\text{most confusing (unstructured)}} : \tilde{\nu} = (\nu_1, \dots, \tilde{\nu}_a, \dots, \nu_A), \mathcal{A}_*(\tilde{\nu}) = \{a\}} \right\}$$

- ▷ KL-UCB plays arms **not pulled enough** for being **uniformly-good**:

$$a_{t+1} \in \operatorname{argmax}_{a \in \mathcal{A}} \max \left\{ \mathbb{E}_{\tilde{\nu}_a}[X] : N_T(a) \leq \frac{\ln(T)}{\text{KL}(\hat{\nu}_{t,a}, \tilde{\nu}_a)}, \tilde{\nu} \text{ most confusing for } a \right\}$$

- ▷ Insight from lower bound: Any **uniformly-good** strategy on \mathcal{D} must satisfy:

$$\forall a \notin \mathcal{A}_*(\nu), \liminf_T \frac{\mathbb{E}[N_T(a)]}{\ln(T)} \geq \sup \left\{ \frac{1}{\underbrace{\text{KL}(\nu_a, \tilde{\nu}_a)}_{\text{most confusing (unstructured)}} : \tilde{\nu} = (\nu_1, \dots, \tilde{\nu}_a, \dots, \nu_A), \mathcal{A}_*(\tilde{\nu}) = \{a\}} \right\}$$

- ▷ KL-UCB plays arms **not pulled enough** for being **uniformly-good**:

$$a_{t+1} \in \operatorname{argmax}_{a \in \mathcal{A}} \max \left\{ \mathbb{E}_{\tilde{\nu}_a}[X] : N_T(a) \leq \frac{\ln(T)}{\text{KL}(\hat{\nu}_{t,a}, \tilde{\nu}_a)}, \tilde{\nu} \text{ most confusing for } a \right\}$$

Play an arm in order to
rule-out a most confusing instance

(Selects one causing maximal regret if not played.)

- ▷ Different from “expecting the best reward in the best world”: testing.

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

OPTIMAL STRUCTURE EXPLOITATION

CONCLUSION, PERSPECTIVE

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

KL-UCB

Thompson Sampling

IMED

Sub/Re-sampling strategy

Advanced concentration tools

EXPLOITING STRUCTURE

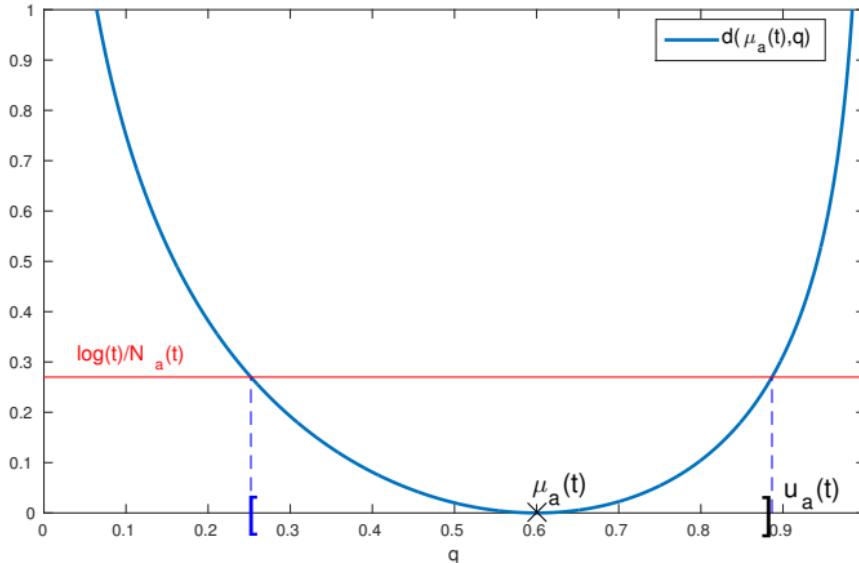
Kullback-Leibler Upper Confidence Bound

Tze Leung Lai. "Adaptive treatment allocation and the multi-armed bandit problem."
The Annals of Statistics , pages 1091-1114, 1987.

THE kl-UCB ALGORITHM (BERNOULLI ARMS)

- A UCB-type algorithm: $A_{t+1} = \operatorname{argmax}_a U_a(t)$
- ... associated to **the right upper confidence bound**:

$$U_a(t) = \max \{q : N_a(t) \text{kl}(\hat{\mu}_a(t), q) \leq \ln(t)\},$$



$$\text{kl}(p, q) = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}$$

Use empirical **distributions**: $\hat{\nu}_a(t) = \frac{1}{N_a(t)} \sum_{s=1}^t \delta_{Y_s} \mathbb{I}_{\{a_s=a\}}$.

- ▶ Family \mathcal{D} : Bernoulli, Poisson, Exponential, Gaussian, etc.
- ▶ $\Pi_{\mathcal{D}} : \mathcal{P}(\mathbb{R}) \rightarrow \mathcal{D} : \Pi_{\mathcal{D}}(\hat{\nu}_a(t)) = \text{Maximum Likelihood}$

KL-UCB for a family \mathcal{D} (generic form)

Choose $a_{t+1} \in \operatorname{Argmax}_{a \in \mathcal{A}} U_a(t)$ where

$$U_a(t) = \sup \left\{ \mathbb{E}_{\nu}[X] : \nu \in \mathcal{D} \text{ and } N_a(t) \text{KL}\left(\Pi_{\mathcal{D}}(\hat{\nu}_a(t)), \nu\right) \leq \ln(t \ln^c(t)) \right\}.$$

For Bernoulli distributions

$$[\text{Capp\'e et al. 13}]: \quad \mathbb{E}_\nu[N_a(T)] \leq \frac{\ln T}{\text{kl}(\mu_a, \mu^*)} + O(\sqrt{\ln(T)}).$$

For exponential families of dimension D (i.e. $\nu(x) \propto \exp(\langle \theta, \psi(x) \rangle)$, $\theta \in \mathbb{R}^D$), and parameter $c > \max(D/2 - 1, 0)$,

$$[\text{Maillard 17}]: \quad \mathbb{E}_\nu[N_a(T)] \leq \frac{\ln T}{\mathcal{K}(\mu_a, \nu^*)} + O(\sqrt{\ln(T)}).$$

- ▶ KL-UCB is **asymptotically optimal!**

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES KL-UCB

Thompson Sampling

IMED

Sub/Re-sampling strategy
Advanced concentration tools

EXPLOITING STRUCTURE

Thompson Sampling

W. R. Thompson. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples."
Biometrika , 25(3-4):285-294, 1933

Bernoulli bandit model $\nu = (\mathcal{B}(\mu_1), \dots, \mathcal{B}(\mu_A))$

- ▶ **frequentist view:** μ_1, \dots, μ_A are **unknown parameters**
- ⇒ tools: estimators, confidence intervals

Bernoulli bandit model $\nu = (\mathcal{B}(\mu_1), \dots, \mathcal{B}(\mu_A))$

► **Bayesian view:** μ_1, \dots, μ_A are **random variables**

prior distribution : $\mu_a \sim \mathcal{U}([0, 1])$

⇒ tool: **posterior** distribution

$$\pi_a(t) = \text{Law}(\mu_a | X_{a,1}, \dots, X_{a,t})$$

Bernoulli bandit model $\nu = (\mathcal{B}(\mu_1), \dots, \mathcal{B}(\mu_A))$

► Bayesian view: μ_1, \dots, μ_A are **random variables**

prior distribution : $\mu_a \sim \mathcal{U}([0, 1])$

⇒ tool: **posterior** distribution

$$\pi_a(t) = \text{Law}(\mu_a | X_{a,1}, \dots, X_{a,t})$$

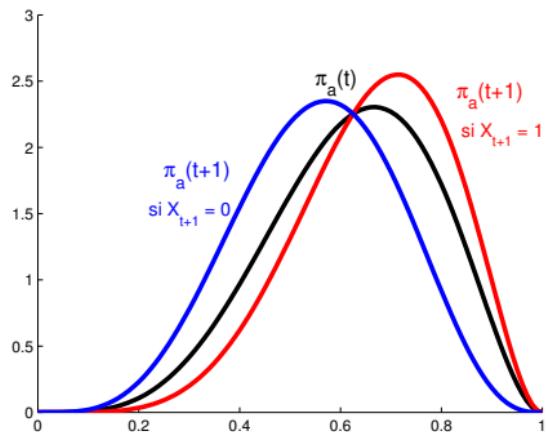
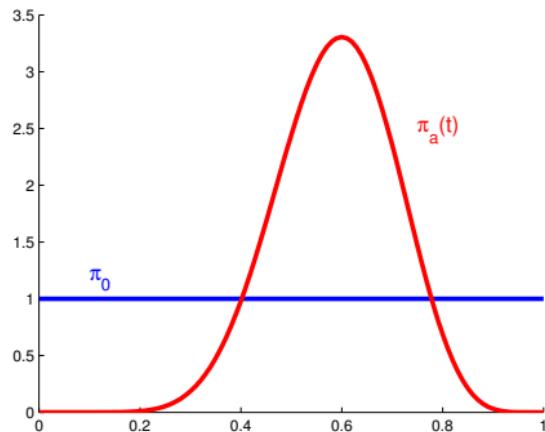
For Bernoulli distribution, and uniform prior, posterior is a Beta:

$$\pi_a(t) = \text{Beta}(S_a(t) + 1, N_a(t) - S_a(t) + 1)$$

$$\text{where } S_a(t) = \sum_{s=1}^t Y_s \mathbb{I}_{(A_s=a)}.$$

Further, for Beta prior, posterior is Beta: **Conjugate prior**.

A **Bayesian bandit algorithm** exploits the posterior distributions of the means to decide which arm to select.



Idea: Use a Bayesian approach to estimate the means $\{\mu_a\}_a$

Idea: Use a Bayesian approach to estimate the means $\{\mu_a\}_a$

Algorithm: Assuming Bernoulli arms and a *Beta* prior on the mean

- ▶ Compute

$$\pi_a(t) = \text{Beta}(S_a(t) + 1, N_a(t) - S_a(t) + 1)$$

- ▶ Draw a mean sample as

$$\tilde{\mu}_a(t) \sim \pi_a(t)$$

- ▶ Pull arm

$$A_t = \arg \max \tilde{\mu}_a(t)$$

- ▶ Update $N_{A_t, t+1} = N_{A_t, t} + 1$. If $X_{A_t, t} = 1$ update $S_{A_t, t+1} = S_{A_t, t} + 1$.

Regret:

$$\lim_{T \rightarrow \infty} \frac{\mathcal{R}_T}{\ln(T)} = \sum_{a \in \mathcal{A}} \frac{\Delta_a}{\text{kl}(\mu_a, \mu^*)}$$

THOMPSON SAMPLING

$$\left\{ \begin{array}{l} \forall a \in \mathcal{A}, \quad \theta_a(t) \sim \pi_a(t) \\ A_{t+1} = \operatorname{argmax}_{a \in \mathcal{A}} \theta_a(t). \end{array} \right.$$

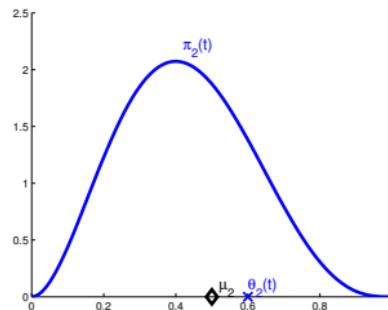
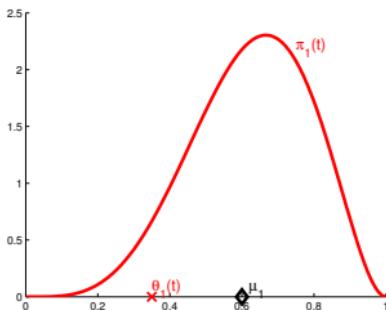


Figure: TS selects arm 2 as $\theta_2(t) \geq \theta_1(t)$

- ⇒ the first bandit algorithm! [Thompson 1933]
 - ⇒ very efficient, beyond Bernoulli bandits
 - ⇒ matches the Lai and Robbins bound for Bernoulli bandits
- K., Korda and Munos, *Thompson Sampling: an Asymptotically Optimal Finite-Time Analysis*, ALT 2012

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

KL-UCB

Thompson Sampling

IMED

Sub/Re-sampling strategy

Advanced concentration tools

EXPLOITING STRUCTURE

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

KL-UCB

Thompson Sampling

IMED

Sub/Re-sampling strategy

Advanced concentration tools

EXPLOITING STRUCTURE

Best Empirical Sub-sampling Average

"Sub-sampling for multi-armed bandits",
Baransi, Maillard, Mannor *ECML*, 2014.

Sub-Sampling Duelling Algorithms

"Sub-sampling for Efficient Non-Parametric Bandit Exploration",
Baudry, Kaufmann, Maillard, *Neurips*, 2020.

Theorem (Burnetas and Katehakis, 1996)

For any strategy π that is consistent (for any bandit, sub-optimal arm a , $\beta > 0$ it holds $\mathbb{E}[N_{T,a}^\pi] = o(T^\beta)$), and $\mathcal{D} \subset \mathcal{P}([0, 1])$

$$\liminf_{T \rightarrow \infty} \frac{R_T}{\ln T} \geq \sum_{a: \Delta_a > 0} \frac{(\mu^* - \mu_a)}{\mathcal{K}_{\inf}(\nu_a, \mu^*)},$$

where $\mathcal{K}_{\inf}(\nu_a, \mu^*) \stackrel{\text{def}}{=} \inf\{KL(\nu_a || \nu), \nu \in \mathcal{D} \text{ has mean } > \mu^*\}$.

Class of optimal algorithms

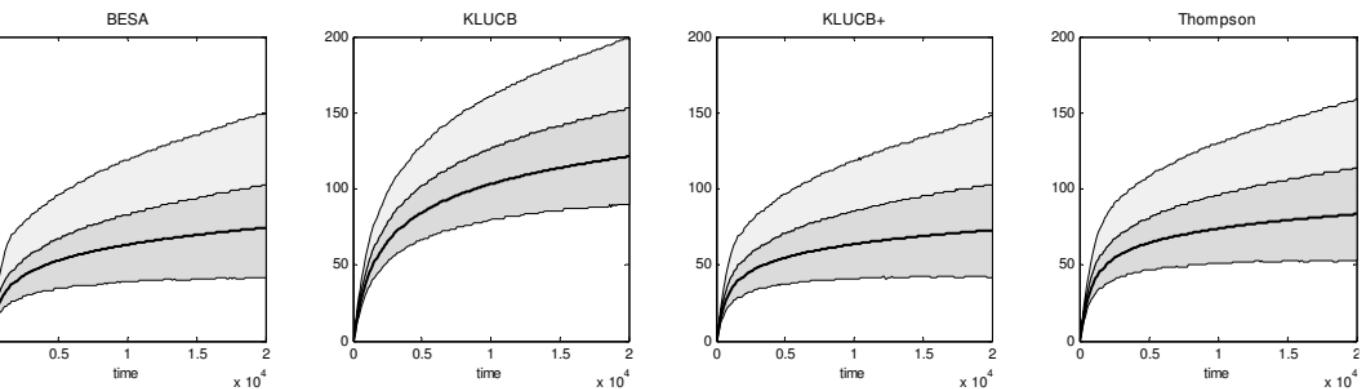
- ▶ Confidence bound: e.g. (Burnetas-Katsehakis, 1996)
- ▶ Bayesian: e.g. Thompson Sampling (Thompson, 1933)
- ▶ Sub-sampling?
- ▶ Provably optimal finite-time regret for **some** \mathcal{D}
- ▶ A **different** algorithm for each \mathcal{D} : TS or KL-UCB for Bernoulli, for Poisson, for Exponential, etc.

Can we get a **uniformly** good algorithm over many \mathcal{D} ?

PUZZLING EXPERIMENTS ($T = 20,000, 50,000$ REPLICATES)

- ▶ 10 Bernoulli($0.1, 3\{0.05\}, 3\{0.02\}, 3\{0.01\}$)

	BESA	kl-UCB	kl-UCB+	TS	Others
Regret	74.4	121.2	72.8	83.4	100-400
Beat BESA	-	1.6%	35.4%	3.1%	
Run Rime	13.9X	2.8X	3.1X	X	



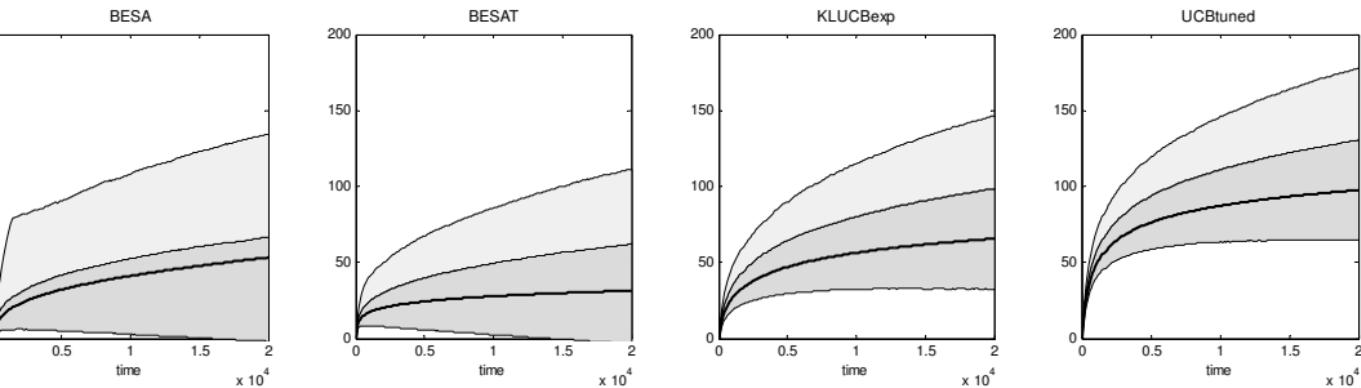
Others: UCB, Moss, UCB-Tunes, DMED, UCB-V.

(Credit: Akram Baransi)

PUZZLING EXPERIMENTS ($T = 20,000, 50,000$ REPLICATES)

► Exponential($\frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1$)

	BESA	KL-UCB-exp	UCB-tuned	FTL 10	Others
Regret at BESA run Rime	53.3 -	65.7 5.7%	97.6 4.3%	306.5 -	60-110,120+
		6X	2.8X	X	



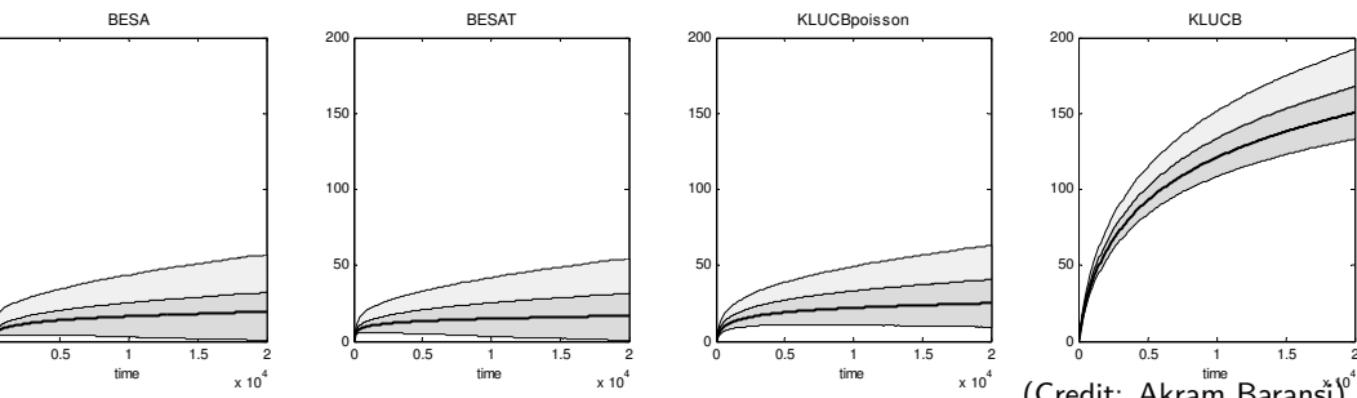
Others: UCB, Moss, kl-UCB, UCB-V.

(Credit: Akram Baransi)

PUZZLING EXPERIMENTS ($T = 20,000, 50,000$ REPLICATES)

► Poisson($\{\frac{1}{2} + \frac{i}{3}\}_{i=1,\dots,6}$)

	BESA	KL-UCB-Poisson	kl-UCB	FTL 10
Regret	19.4	25.1	150.6	144.6
Beat BESA	-	4.1%	0.7%	-
Run Rime	3.5X	1.2X	X	-

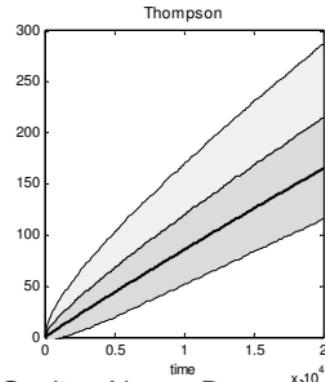
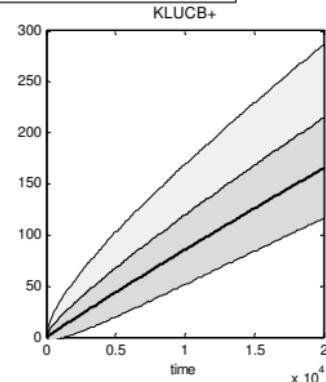
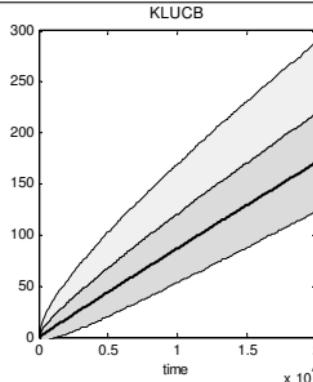
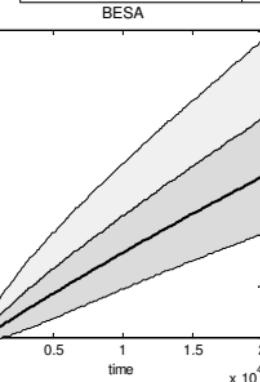


(Credit: Akram Baransi)

PUZZLING EXPERIMENTS ($T = 20,000, 50,000$ REPLICATES)

- **Bernoulli** all half but one 0.51.

	BESA	KL-UCB	KL-UCB+	TS
Regret	156.7	170.8	165.3	165.1
Beat BESA	-	41.4%	41.6%	40.8%
Run Rime	19.6X	2.8X	3X	X



(Credit: Akram Baransi)

BESA

- ▶ Competitive regret against state-of-the-art for various \mathcal{D} .
 - ▶ Same algorithm for all \mathcal{D} .
 - ▶ Not relying on upper confidence bounds, not Bayesian...
 - ▶ ...and extremely simple to implement.
-
- ▶ How? Optimality? For which distributions ?

Proposal: Subsampling Dueling Algorithms (SDA)

A **round-based** approach: at each round we

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).

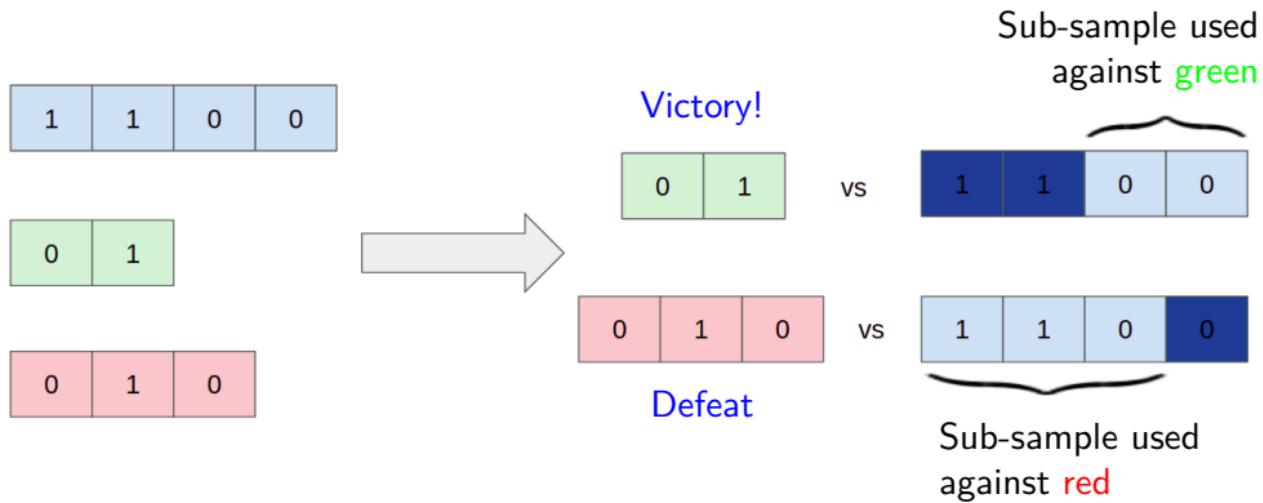
→ possibly several arms drawn per round.

The "sub-sampling step": outcome of a **Duel**

- Challenger → **empirical mean** (full sample size N_k).
- Leader → **mean** of a **subsample** of size N_k chosen from its history.
- Winner: arm with the largest index!

→ We can plug any **independent sampler** into SDA, i.e any sub-sampling algorithm that is independent of the value of the rewards. This includes both *randomized* and *deterministic* samplers.

Example of round

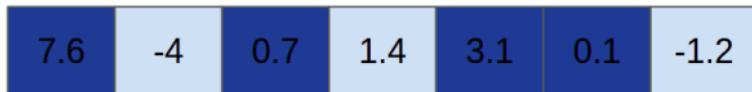


In this example the leader is *blue*: *green* wins against *blue*, *red* loses
 \Rightarrow only *green* is drawn at the end of the round.

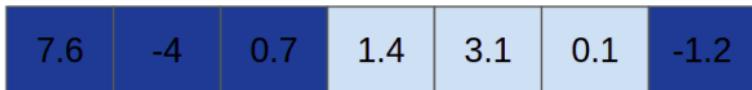
Examples of sub-sampling algorithms

Randomized Samplers

- *Sampling Without Replacement* (SW-SDA): Draw m elements uniformly without replacement.



- *Random Block Sampling* (RB-SDA): draw $n_0 \in [1, N - m + 1]$ and return $\{n_0, n_0 + 1, \dots, n_0 + m - 1\}$.



REGRET BOUND (SLIGHTLY SIMPLIFIED STATEMENT)

Let $\mathcal{A} = \{\star, a\}$ and define

$$\alpha(M, n) = \mathbb{E}_{Z^\star \sim \nu_{\star, n}} \left[\left(\mathbb{P}_{Z \sim \nu_{a, n}}(Z > Z^\star) + \frac{1}{2} \mathbb{P}_{Z \sim \nu_{a, n}}(Z = Z^\star) \right)^M \right],$$

Theorem [Baudry et al., 2020, Regret of the SDA strategy]

Under mild assumption on the α fct (e.g. $\exists \alpha \in (0, 1)$, $c > 0$ such that $\alpha(M, 1) \leq c\alpha^M$), for exponential families of dimension 1 then

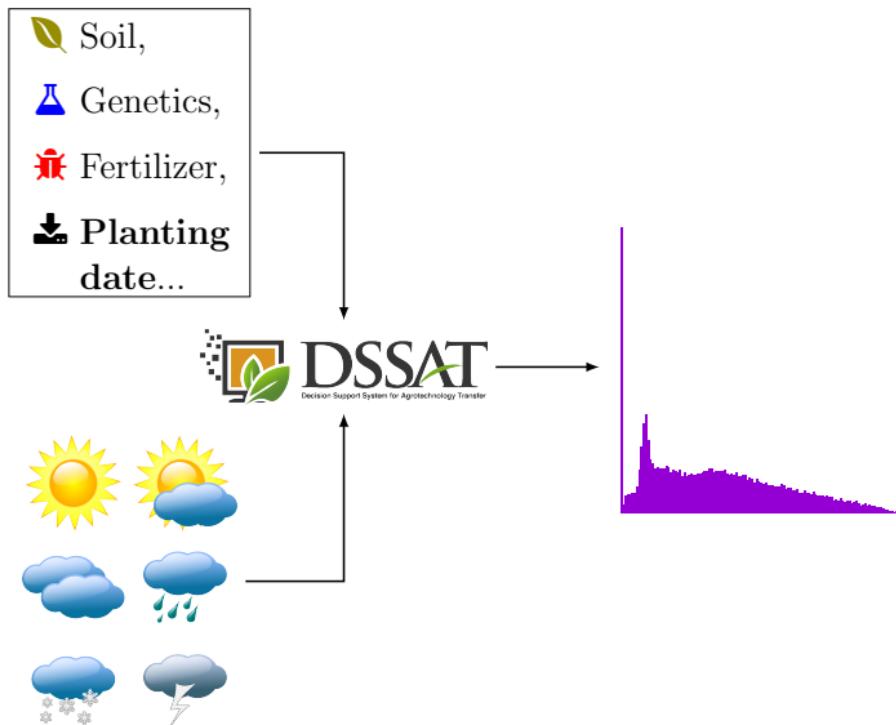
$$\mathbb{E}[N_a(T)] \leq \frac{1 + \varepsilon}{\text{kl}(\mu_a, \mu^\star)} \ln(T) + o_{\nu, \varepsilon}(\ln(T)),$$

Example

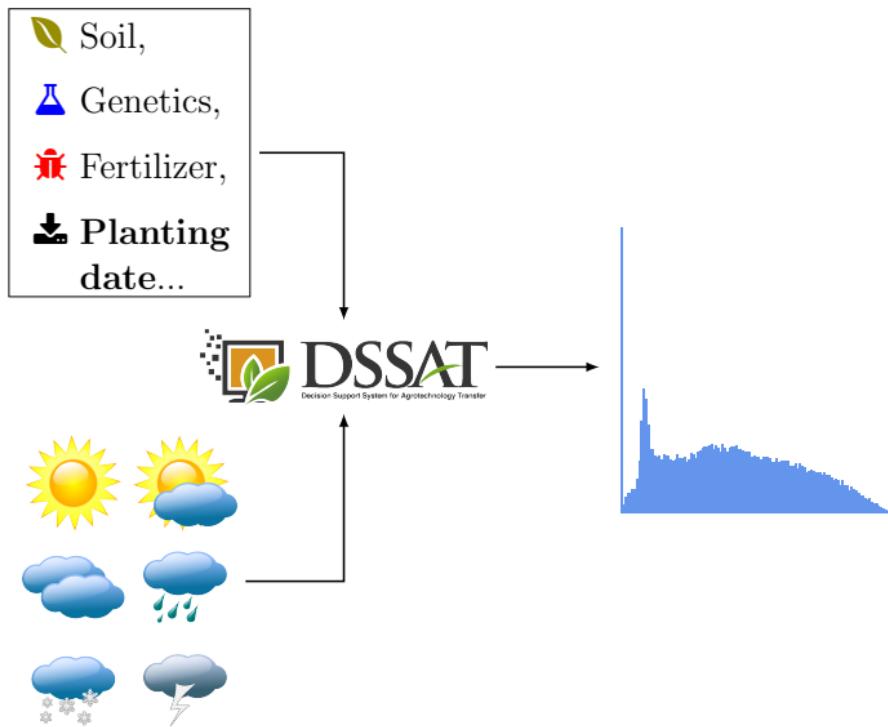
- ▶ Bernoulli μ_a, μ_\star : $\alpha(M, 1) = O\left(\left(\frac{\mu_a \vee (1 - \mu_a)}{2}\right)^M\right)$
- ▶ Interestingly, satisfied for Bernoulli, Gaussian, Poisson, but not Exponential.

- ▷ We can build that is **simultaneously optimal** for several (exponential) parametric families **without** using its parametrization.
- ▷ Can we go fully **non-parametric**?
- ▷ Application?

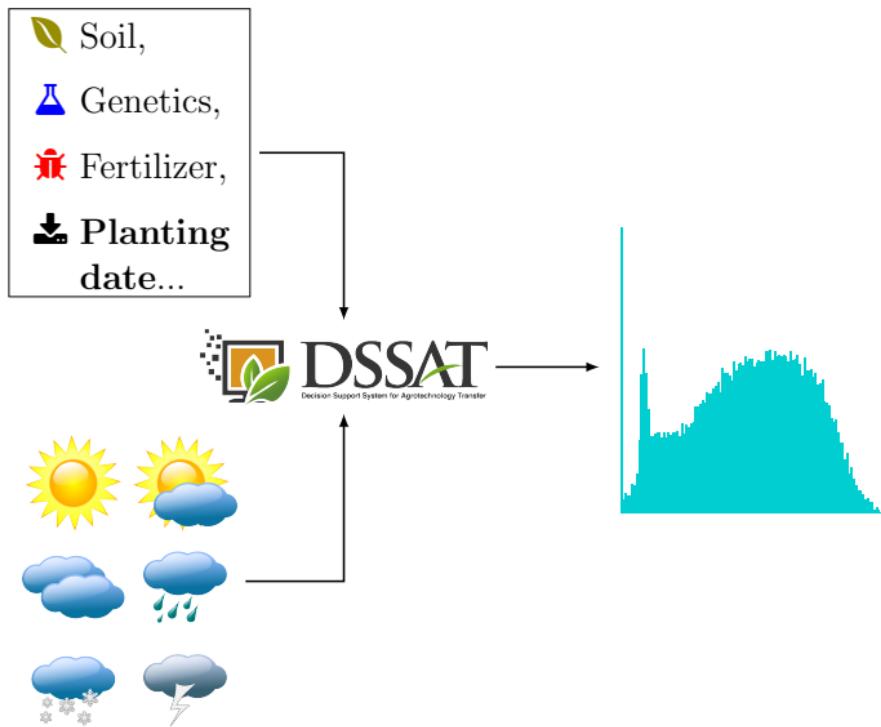
Motivation: recommendations in the real world



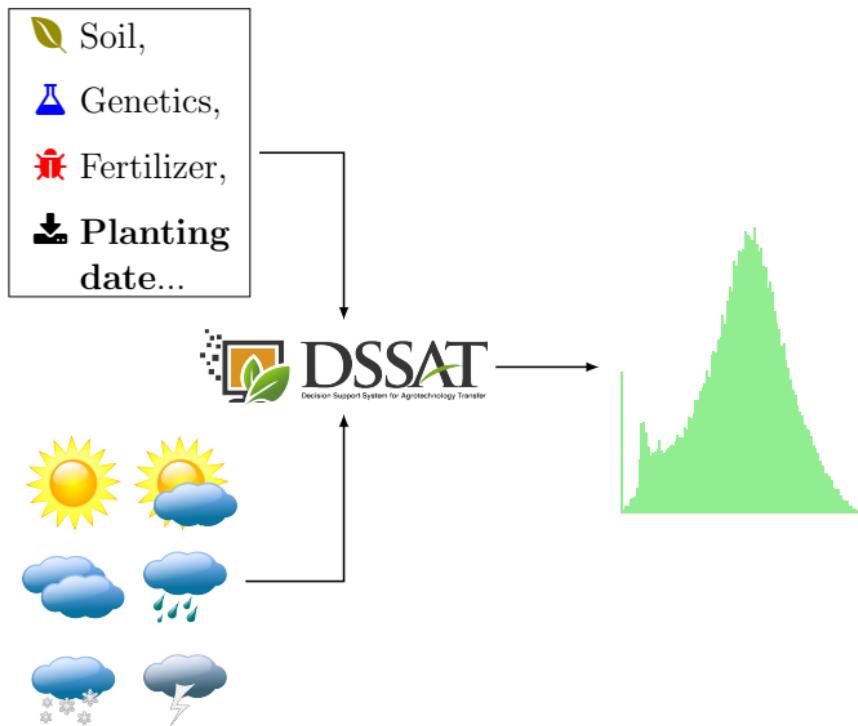
Motivation: recommendations in the real world



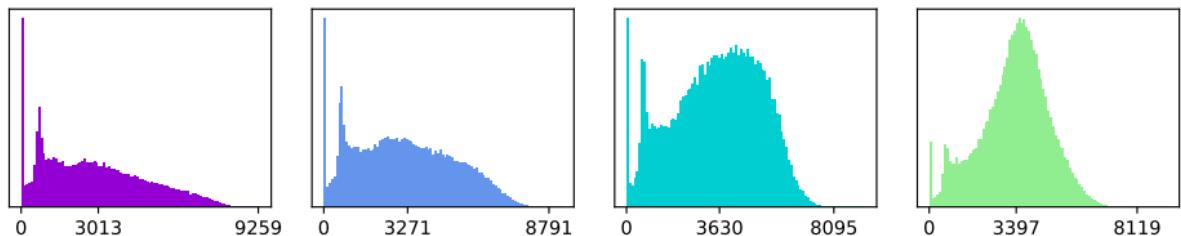
Motivation: recommendations in the real world



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 $\nu \in \mathcal{F}$:

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

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

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		$\text{KL}(\nu_\theta, \nu_{\theta'})$
IMED ²	Single Parameter	$\text{KL}(\nu_\theta, \nu_{\theta'})$
Thompson Sampling ³	Exponential Family (SPEF) $(\nu_\theta)_{\theta \in \Theta}$	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).

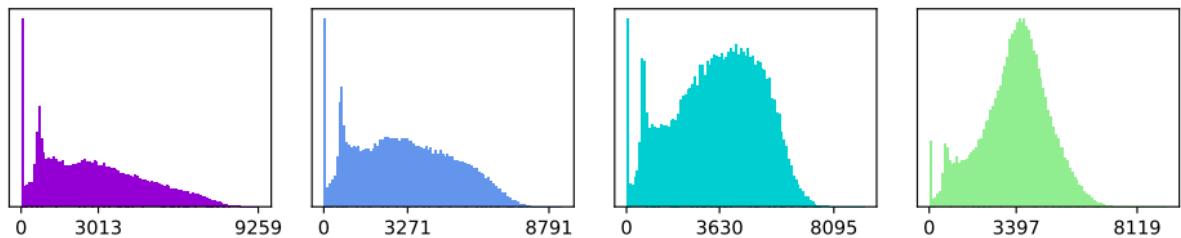
Optimal bandit algorithms: bounded

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

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



$$B_1 \leq B_2 \implies \mathcal{K}_{\inf}^{\mathcal{F}_{B_1}}(\nu_k, \mu^*) \geq \mathcal{K}_{\inf}^{\mathcal{F}_{B_2}}(\nu_k, \mu^*).$$

- Light-tailed ? Reasonable assumption ✓

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

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

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,

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

- $w \sim \mathcal{D}_{n+1}(1, \dots, 1)$ (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

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

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).

→ possibly several arms drawn per round.

Pairwise comparison (**Duel**) step:

- Leader → empirical mean $\hat{\mu}_\ell$.
- Challenger → Dirichlet Sampling, bonus $\mathcal{B}(k, \ell)$.
- 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!

Dirichlet Randomized Exploration

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



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

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 !):

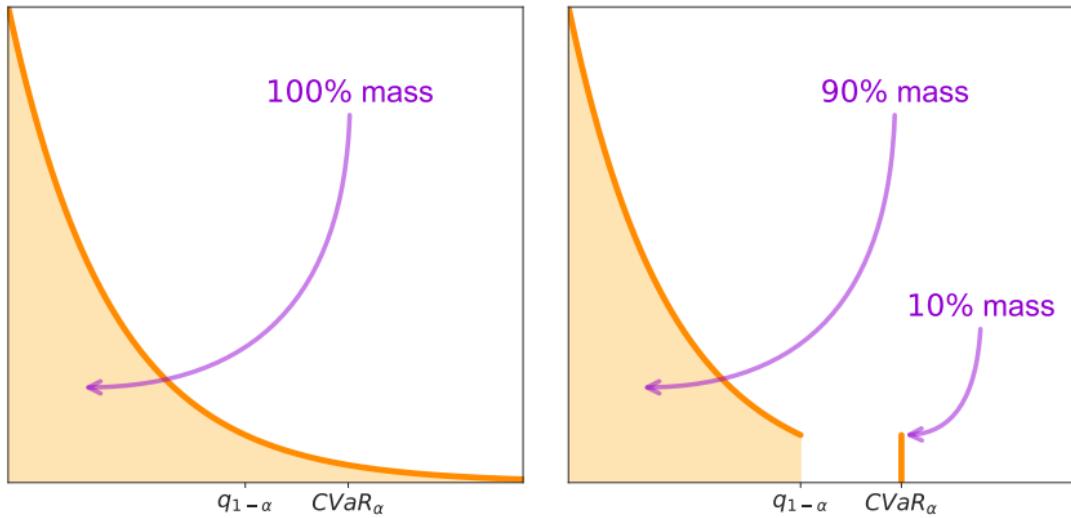
$$\mathfrak{B}(\ell, k) = \max \left(B(\mathcal{X}_k, \hat{\mu}_\ell, \rho), \max_{i=1, \dots, n} X_{k,i} + \gamma \right) .$$

Theorem

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

Algorithm #2: Quantile Dirichlet Sampling (QDS)

? What about unbounded distributions?



✂ ... truncate them!

Algorithm #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:

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

Theorem

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

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

↪ We recommend $\rho_n = \sqrt{\log n}$ as a baseline!

Dirichlet Randomized Exploration

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



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



$$\mathfrak{B}(k, \ell) = B(\mathcal{X}_k, \hat{\mu}_\ell, \rho)$$

$$= \hat{\mu}_\ell + \rho \times \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_\ell - X_{k,i})^+$$

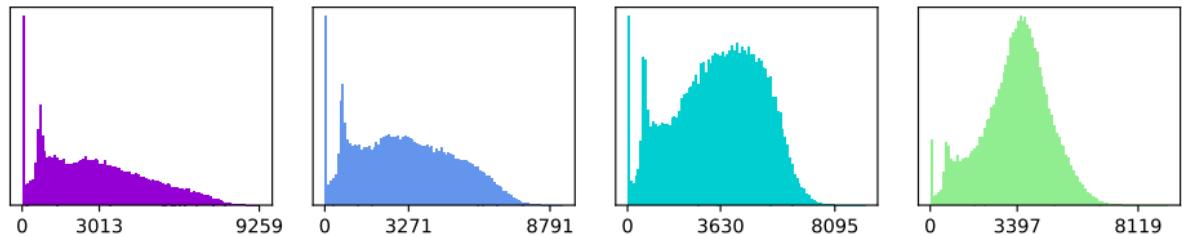


BDS
 $\rho \geq \frac{-1}{\log(1-p)}$

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

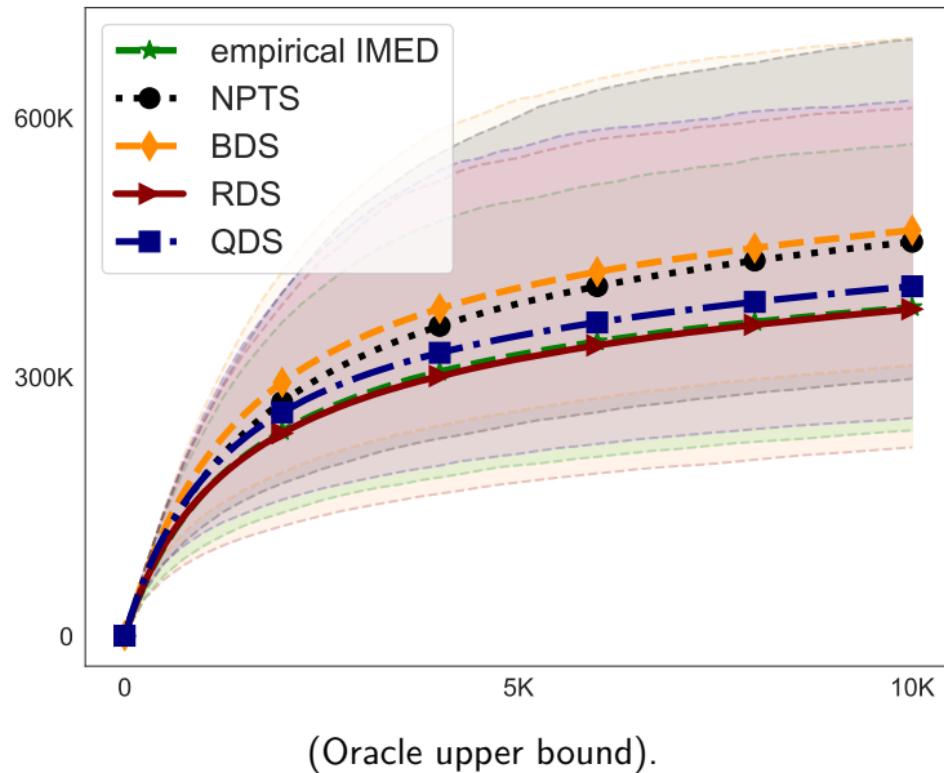
RDS
 $\rho_n = \sqrt{\log(n)}$

Experiments: recommendations in agriculture

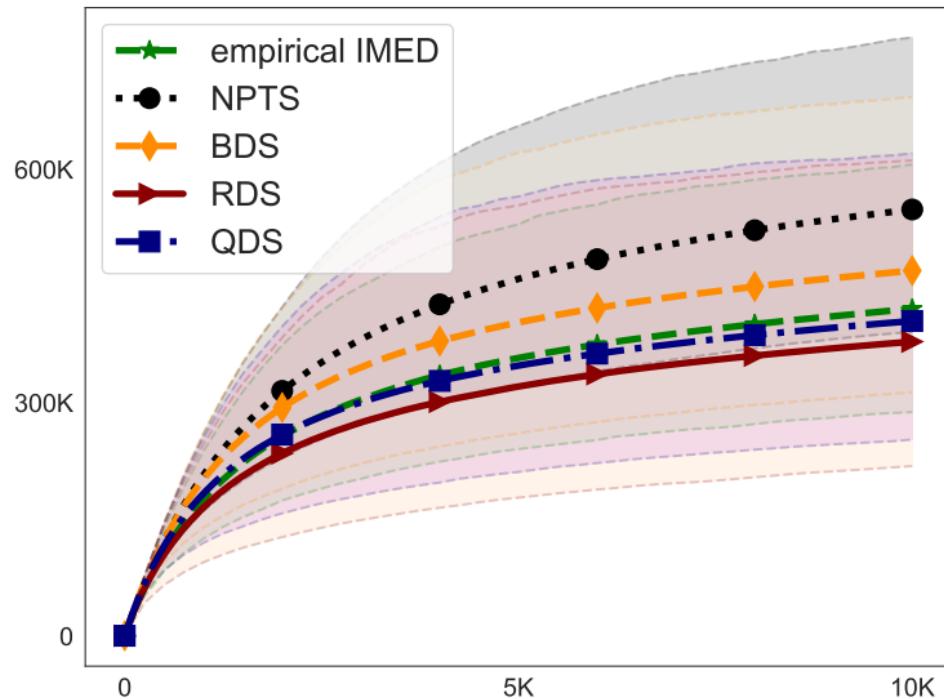


We compare DS algorithms with optimal algorithms considering bounded distributions with known B .

Experiments: recommendations in agriculture



Experiments: recommendations in agriculture



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

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

KL-UCB

Thompson Sampling

IMED

Sub/Re-sampling strategy

Advanced concentration tools

EXPLOITING STRUCTURE

Finite sample guarantee:

$$\mathbb{P}\left[\frac{1}{n} \sum_{t=1}^n X_t - \mathbb{E}[X_1] > (b - a) \sqrt{\frac{\ln 1/\delta}{2n}}\right] \leq \delta$$

Handle **random stopping** times (e.g. $N_T(a)$) carefully

$$\text{(Union bound)} \quad \mathbb{P}\left(\frac{1}{\tau_t} \sum_{i=1}^{\tau_t} (\mu - X_i) \geq \sqrt{\frac{\ln(t/\delta)}{2\tau_t}}\right) \leq \delta.$$

$$\text{(Peeling method)} \quad \mathbb{P}\left(\frac{1}{\tau_t} \sum_{i=1}^{\tau_t} (\mu - X_i) \geq \sqrt{\frac{\alpha}{2\tau_t} \ln \left(\left\lceil \frac{\ln(t)}{\ln(\alpha)} \right\rceil \frac{1}{\delta} \right)}\right) \leq \delta$$

$$\text{(Laplace method)} \quad \mathbb{P}\left(\frac{1}{\tau} \sum_{i=1}^{\tau} (\mu - X_i) \geq \sqrt{\frac{1 + \frac{1}{\tau}}{2\tau} \ln (\sqrt{\tau+1}/\delta)}\right) \leq \delta$$

Provably **reduces regret**, thus mistakes (**saves lives**).

Suggests alternative bounds for UCB strategy.

- ▷ For more details: <https://hal.archives-ouvertes.fr/tel-02077035>

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

OPTIMAL STRUCTURE EXPLOITATION

CONCLUSION, PERSPECTIVE

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

Real-world is structured

Structured actions

Linear structure and regression

Example: Graph-linear bandits

Linear UCB, Linear TS

Infinite dimension

YOUR FAVORITE BANDIT APPLICATION

Eco-sustainable decision making

- ▶ Plant health-care:



$$: \mathcal{A} = \left\{$$



$$\right\}$$

- ▶ Ground health-care:



$$: \mathcal{A} = \left\{$$



$$\right\}$$

Medical decision companion

- ▶ Emergency admission filtering:



$$, \quad$$



$$, \quad$$



$$, \quad$$



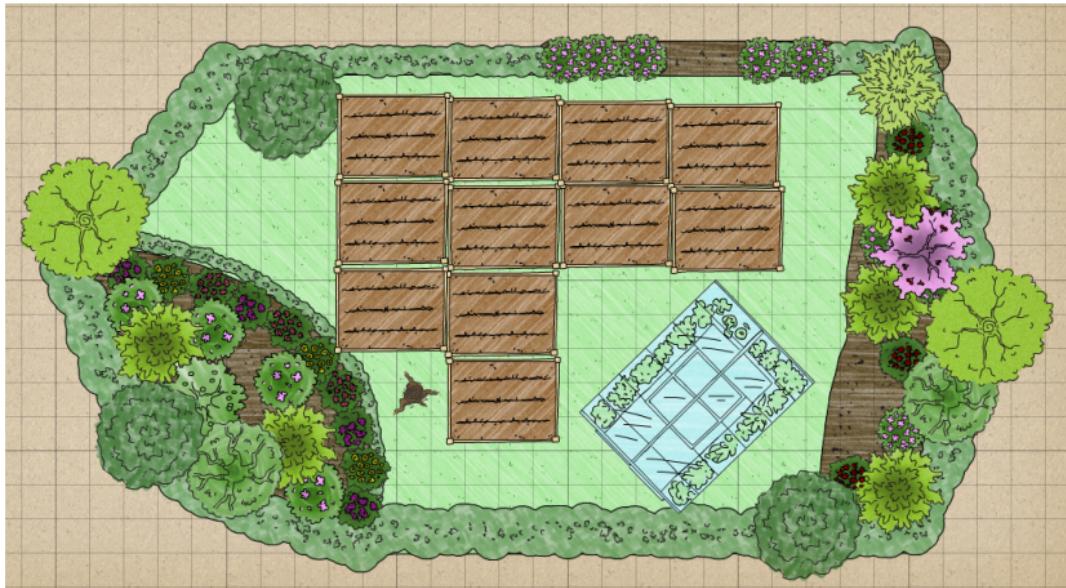
$$\right\}$$

SUSTAINABLE FARMING



- ▶ Recommend good practice between farms/share knowledge.

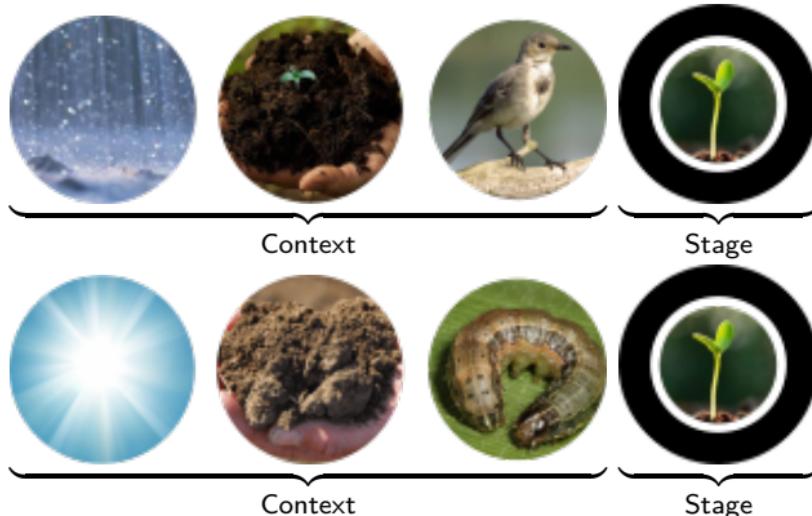
SUSTAINABLE FARMING



- ▶ Recommend good practice between farms/share knowledge.
- ▶ Context, strong correlations, hidden variables, delayed feedback.

CONTEXTUAL INFORMATION

- In practice we observe side information (weather, soil, etc.)



Q: Given a context, what action?

Several lands with different variables:

- ▶ **Geography**: inter-locations, size, surroundings.
- ▶ **Weather**: conditions and exposure (local).
- ▶ **Soil**: chemical, biological, physical (= function(position))
- ▶ **Flora**: spontaneous (presence, quantity).
- ▶ **Fauna**: helpful (bees, etc.), harmful (pests, etc.) or both.
- ▶ **Available actions**: available tools, labor, seeds, etc.
- ▶ **Design/local geography**: facilities inter-locations, pathways, etc.
- ▶ **Farmer target**: mix of survival, cattle, market, aesthetics, etc.
- ▶ **Farmer mind**: risk-aversion, compliance, communication protocol, etc.
- ▶ Other: $X_d, X_{d+1}, X_{d+2}, \dots$

Personalized

Each farm has its own **context**.

- ▶ **Healthy** plants (% disease)
- ▶ **Nutritious** plants (% minerals, etc.)
- ▶ **Healthy** soil (che, phy, bio- properties)
- ▶ **Many** insects, birds, auxiliaries
- ▶ Large **harvest** on one culture cycle (**short**-term), or several (**long**-term).
- ▶ Harvest adapted to **storage** and **distribution** capacity.
- ▶ Low Farmer **Labor** (prevent exhaustion)
- ▶ Few external **resources consumption** (water, fertilizer, fuel, etc.)
- ▶ Low **pest attacks**
- ▶ High **resilience** to pest attacks, draught, etc.
- ▶ Produce **novel/diverse** varieties?
- ▶ Other goals: $R_m, R_{m+1}, R_{m+2}, \dots$

Personalized

Each decision-maker has different prioritization of goals: goal context.



Choice of treatment (action) against disease (observation = health status)



$$: \mathcal{A} = \left\{ \quad , \quad , \quad , \quad , \quad \right\}$$



e.g. design treatments against Tumors, HIV, COVID, etc; insuline dosage (action) for Glycemia regulation (observation);

- ▷ Time series, hidden variables, risk-aversion.
- ▷ Health-status: could be made insanely complicated (e.g. using omics data).



- ▶ Recommend drug dosage w.r.t. genome of individuals.



- ▶ Recommend drug dosage w.r.t. genome of individuals.
- ▶ Huge dimension, Gene interactions.

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

Real-world is structured

Structured actions

Linear structure and regression

Example: Graph-linear bandits

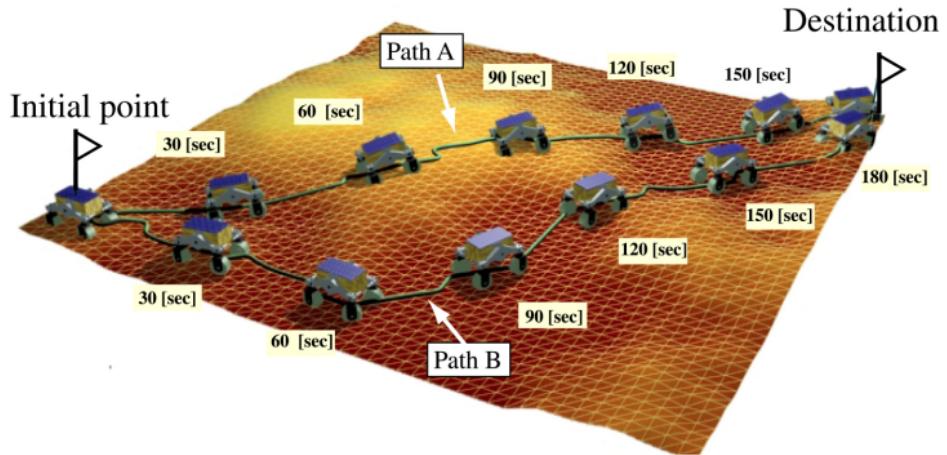
Linear UCB, Linear TS

Infinite dimension

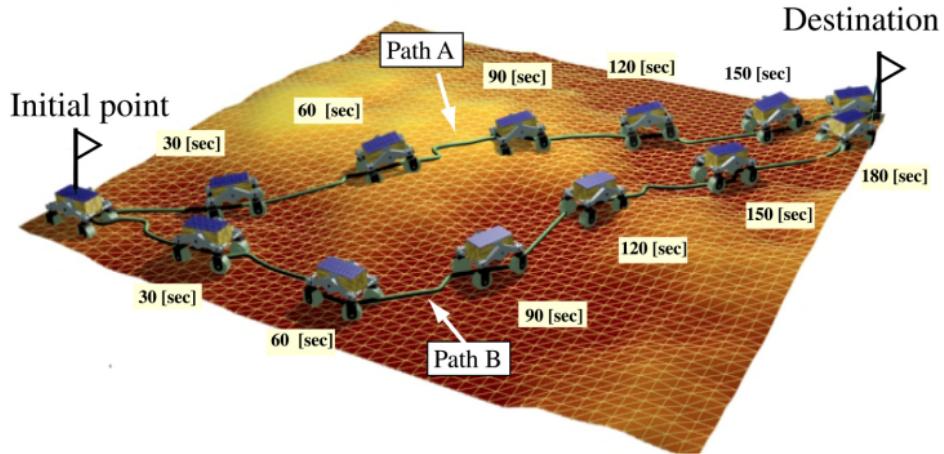
WHAT IS STRUCTURE?

Structure is a **limitation** on the set of possible bandit configurations.
A structure imposes that the distribution of an arm **cannot be changed** arbitrarily without having to change that of **other** arms.

STRUCTURE : PATHS

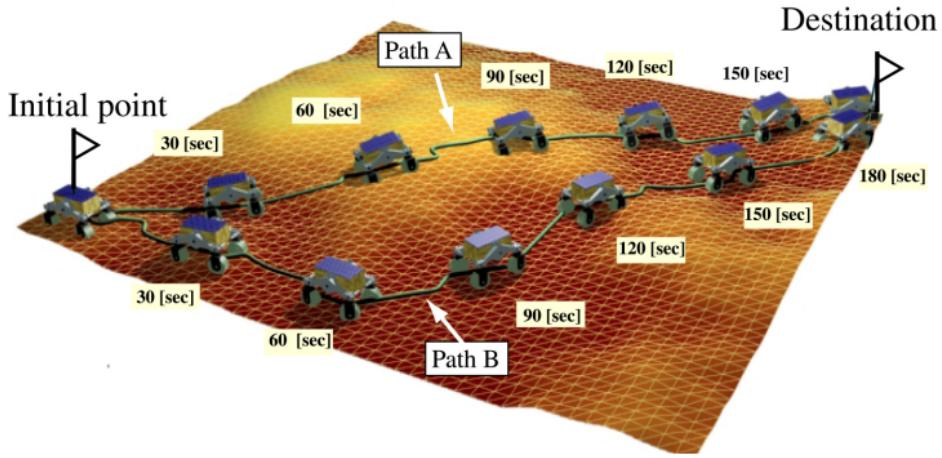


STRUCTURE : PATHS



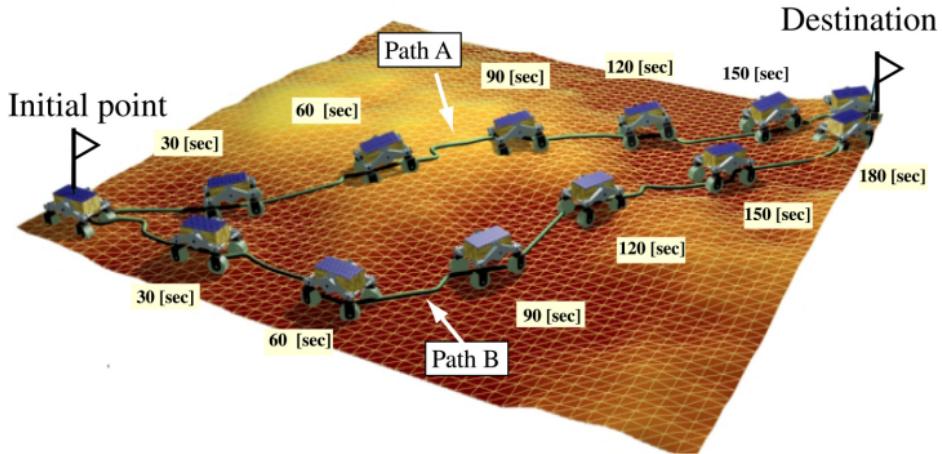
- ▶ Actions: (valued) Paths.

STRUCTURE : PATHS



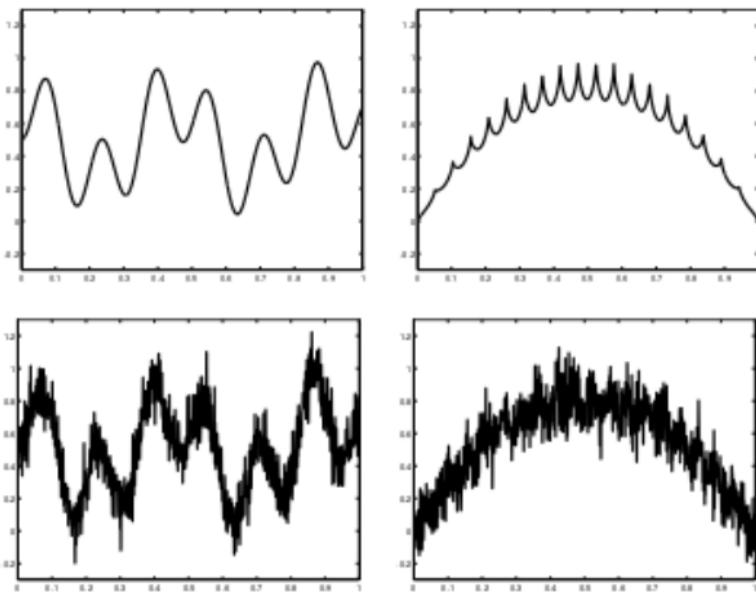
- ▶ Actions: (valued) Paths.
- ▶ Reward/loss: cumulative value on the path.

STRUCTURE : PATHS

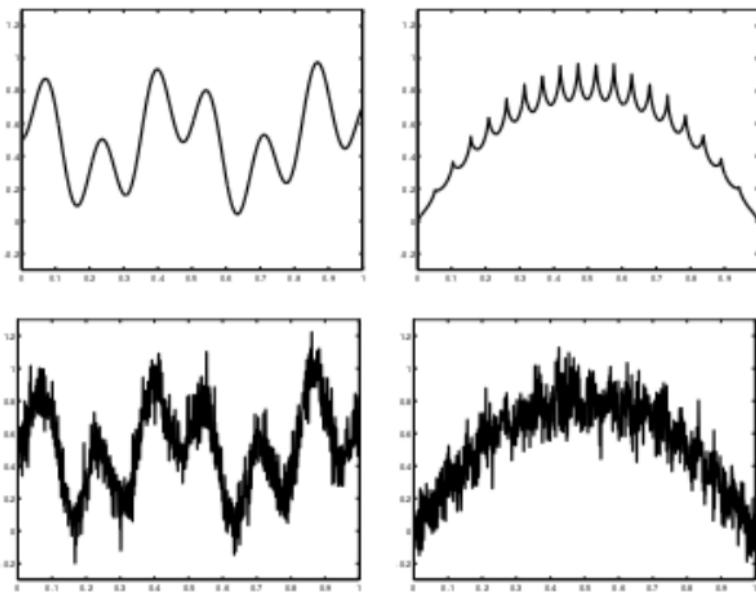


- ▶ Actions: (valued) Paths.
- ▶ Reward/loss: cumulative value on the path.
- ▶ Paths have edges in common.

STRUCTURE: SMOOTH REWARDS

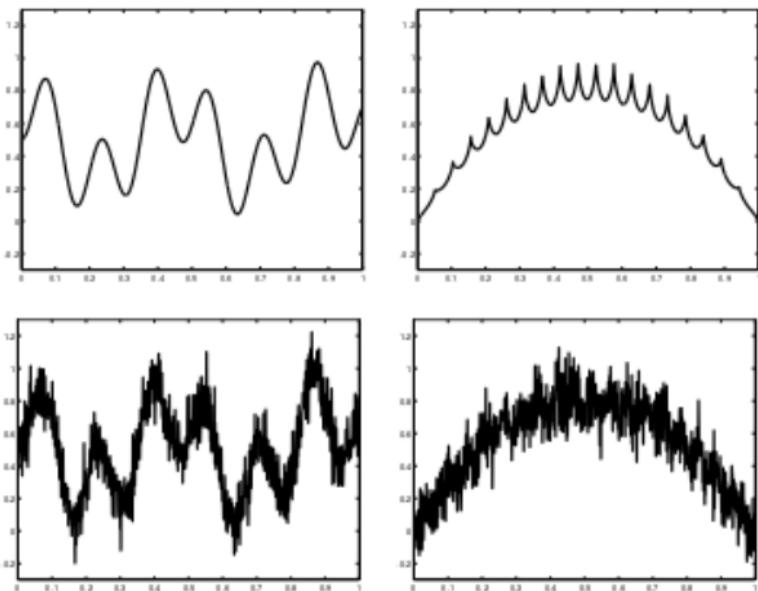


STRUCTURE: SMOOTH REWARDS



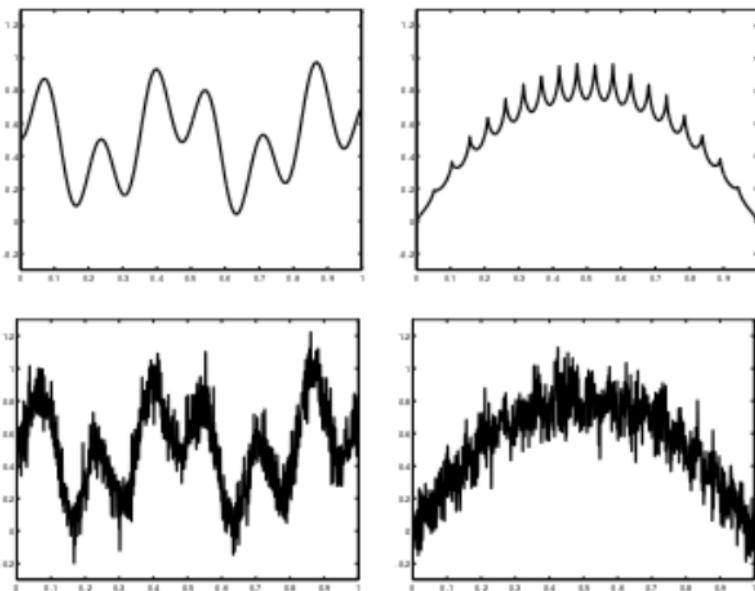
► Actions: $x \in \mathbb{R}$

STRUCTURE: SMOOTH REWARDS

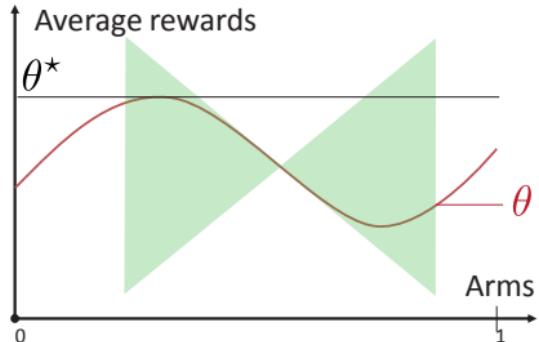


- ▶ Actions: $x \in \mathbb{R}$
- ▶ Reward/loss: $f(x) + \xi$

STRUCTURE: SMOOTH REWARDS



- ▶ Actions: $x \in \mathbb{R}$
- ▶ Reward/loss: $f(x) + \xi$
- ▶ Regularity.



- ▷ Mean parameters: $\Theta_L = \{\theta \in (0, 1)^K : |\theta_i - \theta_j| \leq L|x_i - x_j|, \forall i, j \leq K\}$
- ▷ Exploit this additional information to reduce the achievable regret.

STRUCTURE: LISTS

Google Custom Search Beta

camera UCSD Computer Vision Web Search Search

Camera Calibration Toolbox for Matlab
This is a release of a Camera Calibration Toolbox for Matlab® with a complete documentation. This document may also be used as a tutorial on camera ...
www.vision.caltech.edu/bouguet/calib_doc/ - 14k - [Cached](#)

Omnivis 2003: Omnidirectional Vision and Camera Networks
A complete paper, not longer than six (6) pages including figures and references, should be submitted in camera-ready IEEE 2-column format of single-spaced ...
www.cs.wustl.edu/~pless/omnivis2003/ - 5k - [Cached](#)

Camera Calibration Toolbox for Matlab
A Camera Calibration Toolbox from the Institute of Robotics and Mechatronics, Germany - DLR CalDe and DLR CalLab is a very complete tool for camera ...
www.vision.caltech.edu/bouguet/calib_doc/htmls/links.html - 16k - [Cached](#)

The Page of Omnidirectional Vision
ICCV 2005 Omnidvis'05Sixth Workshop on Omnidirectional Vision, Camera ... Automatic Surveillance Using Omnidirectional and Active Cameras at the PRIP Lab, ...
www.cis.upenn.edu/~kostas/omni.html - 35k - [Cached](#)

Digital Camera Characteristics
It is necessary to know your camera characteristics if you intend to make full use of all of the functions available on your camera ...
www.ncsu.edu/sciencejunction/route/usetech/digitalcamera/ - 10k - [Cached](#)

PDF A Comparison of PMD-Cameras and Stereo-Vision for the Task of ...
File Format: PDF/Adobe Acrobat - [View as HTML](#)
systems and PMD cameras is discussed qualitatively and ... the stereo system as well as the PMD camera will be compared in section 4 based on those ...
vision.middlebury.edu/conferences/bencos2007/pdf/beder.pdf

STRUCTURE: LISTS

The screenshot shows a Google search results page. The search bar at the top contains the word "camera". Below the search bar, there are two tabs: "Custom Search" and "UCSD Computer Vision", with "Custom Search" being the active tab. To the right of the tabs is a "Search" button. The search results are listed below:

- Camera Calibration Toolbox for Matlab**
This is a release of a Camera Calibration Toolbox for Matlab® with a complete documentation. This document may also be used as a tutorial on camera ...
www.vision.caltech.edu/bouguet/calib_doc/ - 14k - [Cached](#)
- Omnivis 2003: Omnidirectional Vision and Camera Networks**
A complete paper, not longer than six (6) pages including figures and references, should be submitted in camera-ready IEEE 2-column format of single-spaced ...
www.cs.wustl.edu/~pless/omnivis2003/ - 5k - [Cached](#)
- Camera Calibration Toolbox for Matlab**
A Camera Calibration Toolbox from the Institute of Robotics and Mechatronics, Germany - DLR CalDe and DLR CalLab is a very complete tool for camera ...
www.vision.caltech.edu/bouguet/calib_doc/htmls/links.html - 16k - [Cached](#)
- The Page of Omnidirectional Vision**
ICCV 2005 Omnidvis'05Sixth Workshop on Omnidirectional Vision, Camera ... Automatic Surveillance Using Omnidirectional and Active Cameras at the PRIP Lab, ...
www.cis.upenn.edu/~kostas/omni.html - 35k - [Cached](#)
- Digital Camera Characteristics**
It is necessary to know your camera characteristics if you intend to make full use of all of the functions available on your camera ...
www.ncsu.edu/sciencejunction/route/usetech/digitalcamera/ - 10k - [Cached](#)
- [PDF] A Comparison of PMD-Cameras and Stereo-Vision for the Task of ...**
File Format: PDF/Adobe Acrobat - [View as HTML](#)
systems and PMD cameras is discussed qualitatively and ... the stereo system as well as the PMD camera will be compared in section 4 based on those ...
vision.middlebury.edu/conferences/bencos2007/pdf/beder.pdf

- ▶ Actions: List of items.

STRUCTURE: LISTS

Google Custom Search BETA

camera UCSD Computer Vision Web Search Search

Camera Calibration Toolbox for Matlab
This is a release of a Camera Calibration Toolbox for Matlab® with a complete documentation. This document may also be used as a tutorial on camera ...
www.vision.caltech.edu/bouguet/calib_doc/ - 14k - [Cached](#)

Omnivis 2003: Omnidirectional Vision and Camera Networks
A complete paper, not longer than six (6) pages including figures and references, should be submitted in camera-ready IEEE 2-column format of single-spaced ...
www.cs.wustl.edu/~pless/omnivis2003/ - 5k - [Cached](#)

Camera Calibration Toolbox for Matlab
A Camera Calibration Toolbox from the Institute of Robotics and Mechatronics, Germany - DLR CalDe and DLR CalLab is a very complete tool for camera ...
www.vision.caltech.edu/bouguet/calib_doc/htmls/links.html - 16k - [Cached](#)

The Page of Omnidirectional Vision
ICCV 2005 Omnidvis'05Sixth Workshop on Omnidirectional Vision, Camera ... Automatic Surveillance Using Omnidirectional and Active Cameras at the PRIP Lab, ...
www.cis.upenn.edu/~kostas/omni.html - 35k - [Cached](#)

Digital Camera Characteristics
It is necessary to know your camera characteristics if you intend to make full use of all of the functions available on your camera ...
www.ncsu.edu/sciencejunction/route/usetech/digitalcamera/ - 10k - [Cached](#)

PDF A Comparison of PMD-Cameras and Stereo-Vision for the Task of ...
File Format: PDF/Adobe Acrobat - [View as HTML](#)
systems and PMD cameras is discussed qualitatively and ... the stereo system as well as the PMD camera will be compared in section 4 based on those ...
vision.middlebury.edu/conferences/bencos2007/pdf/beder.pdf

- ▶ Actions: List of items.
- ▶ Reward/loss: Ranking of preferred item.

STRUCTURE: LISTS

The screenshot shows a Google search results page for the query "camera". The search bar at the top contains "camera". Below it, there are several search results:

- Camera Calibration Toolbox for Matlab**
This is a release of a Camera Calibration Toolbox for Matlab® with a complete documentation. This document may also be used as a tutorial on camera ...
www.vision.caltech.edu/bouguet/calib_doc/ - 14k - [Cached](#)
- Omnivis 2003: Omnidirectional Vision and Camera Networks**
A complete paper, not longer than six (6) pages including figures and references, should be submitted in camera-ready IEEE 2-column format of single-spaced ...
www.cs.wustl.edu/~pless/omnivis2003/ - 5k - [Cached](#)
- Camera Calibration Toolbox for Matlab**
A Camera Calibration Toolbox from the Institute of Robotics and Mechatronics, Germany - DLR CalDe and DLR CalLab is a very complete tool for camera ...
www.vision.caltech.edu/bouguet/calib_doc/htmls/links.html - 16k - [Cached](#)
- The Page of Omnidirectional Vision**
ICCV 2005 Omnidvis'05Sixth Workshop on Omnidirectional Vision, Camera ... Automatic Surveillance Using Omnidirectional and Active Cameras at the PRIP Lab, ...
www.cis.upenn.edu/~kostas/omni.html - 35k - [Cached](#)
- Digital Camera Characteristics**
It is necessary to know your camera characteristics if you intend to make full use of all of the functions available on your camera ...
www.ncsu.edu/sciencejunction/route/usetech/digitalcamera/ - 10k - [Cached](#)
- PDF A Comparison of PMD-Cameras and Stereo-Vision for the Task of ...**
File Format: PDF/Adobe Acrobat - [View as HTML](#)
systems and PMD cameras is discussed qualitatively and ... the stereo system as well as the PMD camera will be compared in section 4 based on those ...
vision.middlebury.edu/conferences/bencos2007/pdf/beder.pdf

- ▶ Actions: List of items.
- ▶ Reward/loss: Ranking of preferred item.
- ▶ Ordering

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

Real-world is structured

Structured actions

Linear structure and regression

Example: Graph-linear bandits

Linear UCB, Linear TS

Infinite dimension

Sequential optimization game

At each time $t \in \mathbb{N}$, sample at $x_t \in \mathcal{X}$, receive $y_t \in \mathbb{R}$, where

$$y_t = \underbrace{f_\star(x_t)}_{\text{target}} + \underbrace{\xi_t}_{\text{noise}}.$$

Goal: Minimize cumulative regret

$$\mathcal{R}_T \stackrel{\text{def}}{=} \sum_{t=1}^T f_\star(\star) - f_\star(x_t) \text{ where } \star \in \operatorname{Argmax} f_\star(x).$$

Sequential optimization game

At each time $t \in \mathbb{N}$, sample at $x_t \in \mathcal{X}$, receive $y_t \in \mathbb{R}$, where

$$y_t = \underbrace{f_\star(x_t)}_{\text{target}} + \underbrace{\xi_t}_{\text{noise}}.$$

Goal: Minimize cumulative regret

$$\mathcal{R}_T \stackrel{\text{def}}{=} \sum_{t=1}^T f_\star(\star) - f_\star(x_t) \text{ where } \star \in \operatorname{Argmax} f_\star(x).$$

- Actions : $x \in \mathcal{X}$.

Sequential optimization game

At each time $t \in \mathbb{N}$, sample at $x_t \in \mathcal{X}$, receive $y_t \in \mathbb{R}$, where

$$y_t = \underbrace{f_\star(x_t)}_{\text{target}} + \underbrace{\xi_t}_{\text{noise}}.$$

Goal: Minimize cumulative regret

$$\mathcal{R}_T \stackrel{\text{def}}{=} \sum_{t=1}^T f_\star(\star) - f_\star(x_t) \text{ where } \star \in \operatorname{Argmax} f_\star(x).$$

- ▶ Actions : $x \in \mathcal{X}$.
- ▶ Means : $f_\star(x)$. Mean at x and x' not arbitrarily different !

- ▶ Set of arms \mathcal{X}

- ▶ Set of arms \mathcal{X}
- ▶ At time t , pick $X_t \in \mathcal{X}$, receive

$$Y_t = f_*(X_t) + \xi_t$$

where ξ_t is centered and further conditionally sub-Gaussian.

f_* belongs to a linear function space:

$$\mathcal{F}_\Theta = \left\{ f_\theta : x \mapsto \theta^\top \varphi(x), \theta \in \Theta \right\} \text{ where } \Theta \in \mathbb{R}^d, \varphi : \mathcal{X} \rightarrow \mathbb{R}^d.$$

θ : Parameter, φ : Feature function.

- ▶ Set of arms \mathcal{X}
- ▶ At time t , pick $X_t \in \mathcal{X}$, receive

$$Y_t = f_*(X_t) + \xi_t$$

where ξ_t is centered and further conditionally sub-Gaussian.

f_* belongs to a linear function space:

$$\mathcal{F}_\Theta = \left\{ f_\theta : x \mapsto \theta^\top \varphi(x), \theta \in \Theta \right\} \text{ where } \Theta \in \mathbb{R}^d, \varphi : \mathcal{X} \rightarrow \mathbb{R}^d.$$

θ : Parameter, φ : Feature function.

- ▶ Unknown parameter $\theta_* \in \mathbb{R}^d$.

- ▶ Set of arms \mathcal{X}
- ▶ At time t , pick $X_t \in \mathcal{X}$, receive

$$Y_t = f_*(X_t) + \xi_t$$

where ξ_t is centered and further conditionally sub-Gaussian.

f_* belongs to a linear function space:

$$\mathcal{F}_\Theta = \left\{ f_\theta : x \mapsto \theta^\top \varphi(x), \theta \in \Theta \right\} \text{ where } \Theta \in \mathbb{R}^d, \varphi : \mathcal{X} \rightarrow \mathbb{R}^d.$$

θ : Parameter, φ : Feature function.

- ▶ Unknown parameter $\theta_* \in \mathbb{R}^d$.
- ▶ Best arm $x_* = \operatorname{argmax}_{x \in \mathcal{X}} \langle \theta_*, \varphi(x) \rangle$

- ▶ **Polynomials:** $\mathcal{X} = \mathbb{R}$, $\varphi(x) = (1, x, x^2, \dots, x^{d-1})$, $\Theta = \mathcal{B}_{2,d}(0, 1)$ unit Euclidean ball of \mathbb{R}^d .
- ▶ **Bandits:** $\mathcal{X} = \mathcal{A} = \{1, \dots, \mathcal{A}\}$, $\varphi(a) = e_a \in \mathbb{R}^{\mathcal{A}}$, $\Theta = [0, 1]^{\mathcal{A}}$.
- ▶ **Shortest path:** $\mathcal{X} \subset \mathcal{A}^L$ (paths of length L), $\varphi_{(a,\ell)}(x) = \mathbb{I}\{x_\ell = a\}$,
 $\Theta = [0, 1]^{|\mathcal{X}|}$.
 $\mathcal{X} \subset \{0, 1\}^d$, paths in graph with d edges, $\varphi(x) = x$, $\Theta \subset [0, 1]^d$ mean travel time for each edge (Combes et al. 2015).
- ▶ **Contextual bandits:** $\mathcal{X} = \mathcal{C} \times \mathcal{A}$, $\varphi((c, a)) = (1, c, a, ca, \dots)$
- ▶ **Smooth function on graph:** $\mathcal{X} = \text{nodes of a graph with adjacency matrix } G$,
 $\varphi = \text{eigenfunctions of the Graph-Laplacian}$.

Linear regression is virtually everywhere...

- ▶ **Linear space:** $\mathcal{F} = \left\{ f_\theta : f_\theta(x) = \langle \theta, \varphi(x) \rangle, \theta \in \mathbb{R}^d, \theta \in \Theta \right\}$.
Ex: $\varphi(x) = (1, x, x^2)$, $f_\theta(x) = 2 + \frac{1}{2}x - 2x^2$, $\theta = (2, 1/2, -2)$.

- ▶ **Linear space:** $\mathcal{F} = \left\{ f_\theta : f_\theta(x) = \langle \theta, \varphi(x) \rangle, \theta \in \mathbb{R}^d, \theta \in \Theta \right\}$.
Ex: $\varphi(x) = (1, x, x^2)$, $f_\theta(x) = 2 + \frac{1}{2}x - 2x^2$, $\theta = (2, 1/2, -2)$.
- ▶ **Loss :** $\ell(y, y') = \frac{(y-y')^2}{2}$

- ▶ **Linear space:** $\mathcal{F} = \left\{ f_\theta : f_\theta(x) = \langle \theta, \varphi(x) \rangle, \theta \in \mathbb{R}^d, \theta \in \Theta \right\}$.
Ex: $\varphi(x) = (1, x, x^2)$, $f_\theta(x) = 2 + \frac{1}{2}x - 2x^2$, $\theta = (2, 1/2, -2)$.
- ▶ **Loss :** $\ell(y, y') = \frac{(y-y')^2}{2}$
- ▶ **Objective :** from $(x_n, y_n)_{n \leq N}$ optimize

$$\min_{\theta \in \Theta} \sum_{n=1}^N \ell\left(y_n, f_\theta(x_n)\right).$$

$$\min_{\theta \in \Theta} \sum_{n=1}^N \left(y_n - \theta^\top \varphi(x_n) \right)^2. \quad (1)$$

- ▶ Any solution to (1) must satisfy

$$G_N \theta = \sum_{n=1}^N \varphi(x_n) y_n, \text{ where } G_N = \sum_{n=1}^N \varphi(x_n) \varphi(x_n)^\top \text{ (*d* } \times \text{ *d* matrix).}$$

- ▶ Any solution to (1) must satisfy

$$G_N \theta = \sum_{n=1}^N \varphi(x_n) y_n, \text{ where } G_N = \sum_{n=1}^N \varphi(x_n) \varphi(x_n)^\top \text{ (*d* } \times \text{ *d* matrix).}$$

- ▶ **Matrix notations:**

$$\begin{aligned} Y_N &= (y_1, \dots, y_N)^\top \in \mathbb{R}^N, \\ \Phi_N &= (\varphi^\top(x_1), \dots, \varphi^\top(x_N))^\top \text{ (*N* } \times \text{ *d* matrix).} \end{aligned}$$

$$G_N \theta = \Phi_N^\top Y_N, \text{ where } G_N = \Phi_N^\top \Phi_N.$$

- ▶ Specific solution: $\theta_N^\dagger = G_N^\dagger \Phi_N^\top Y_N$ where G_N^\dagger : pseudo-inverse of G_N .

- ▶ Specific solution: $\theta_N^\dagger = G_N^\dagger \Phi_N^\top Y_N$ where G_N^\dagger : pseudo-inverse of G_N .
- ▶ Solutions:

$$\begin{aligned}\Theta_N &= \{\theta \in \Theta : G_N(\theta_N^\dagger - \theta) = 0\} \\ &= \{\theta_N^\dagger + \ker(G_N)\} \cap \Theta.\end{aligned}$$

- ▶ Specific solution: $\theta_N^\dagger = G_N^\dagger \Phi_N^\top Y_N$ where G_N^\dagger : pseudo-inverse of G_N .
- ▶ Solutions:

$$\begin{aligned}\Theta_N &= \{\theta \in \Theta : G_N(\theta_N^\dagger - \theta) = 0\} \\ &= \{\theta_N^\dagger + \ker(G_N)\} \cap \Theta.\end{aligned}$$

- ▶ When $\Theta = \mathbb{R}^d$ and G_N is invertible, $G_N^\dagger = G_N^{-1}$,

$$(\text{ Ordinary Least-squares}) \quad \theta_N = G_N^{-1} \Phi_N^\top Y_N.$$

► **Error control:**

$$\forall x \in \mathcal{X}, \quad |f_\star(x) - f_{\theta_N}(x)| \leq \|\theta_\star - \theta_N\|_A \|\varphi(x)\|_{A^{-1}}. \quad (2)$$

for each invertible matrix A , where $\|x\|_A = \sqrt{x^T A x}$.

► **Error control:**

$$\forall x \in \mathcal{X}, \quad |f_\star(x) - f_{\theta_N}(x)| \leq \|\theta_\star - \theta_N\|_A \|\varphi(x)\|_{A^{-1}}. \quad (2)$$

for each invertible matrix A , where $\|x\|_A = \sqrt{x^T A x}$.

- Matrix $A = G_N$ has natural interpretation: for $\theta \in \Theta_N$ (solution),

$$\sum_{n=1}^N (f_\star(x_n) - f_\theta(x_n))^2 = \sum_{n=1}^N (\theta^\star - \theta)^\top \varphi(x_n) \varphi(x_n)^\top (\theta^\star - \theta) = \|\theta^\star - \theta\|_{G_N}^2.$$

(Over-fitting is $\forall \theta \in \Theta_N$, $\|\theta^\star - \theta\|_{G_N} = 0$).

Study $\ \theta_\star - \theta_N\ _{G_N}$

REGULARIZED LEAST-SQUARES

When G_N is not invertible, introduce regularization parameter $\lambda \in \mathbb{R}_\star^+$.

When G_N is not invertible, introduce regularization parameter $\lambda \in \mathbb{R}_\star^+$.

► **Regularized** solution

$$\theta_{N,\lambda} = G_{N,\lambda}^{-1} \Phi_N^\top Y_N \text{ where } G_{N,\lambda} = \Phi_N^\top \Phi_N + \lambda I_d.$$

REGULARIZED LEAST-SQUARES

When G_N is not invertible, introduce regularization parameter $\lambda \in \mathbb{R}_\star^+$.

- **Regularized** solution

$$\theta_{N,\lambda} = G_{N,\lambda}^{-1} \Phi_N^\top Y_N \text{ where } G_{N,\lambda} = \Phi_N^\top \Phi_N + \lambda I_d.$$

- Bayesian interpretation:

For **Prior** $\theta \sim \mathcal{N}(0, \Sigma)$, i.i.d. setup, Gaussian noise ($\xi_n \sim \mathcal{N}(0, \sigma^2)$),

Posterior: $\hat{f}_N(x) | x, x_1, y_1, \dots, x_N, y_N \sim \mathcal{N}(\mu_N(x), \sigma_N^2(x))$ where

$$\mu_N(x) = \varphi(x)^\top (\Phi_N^\top \Phi_N + \sigma^2 \Sigma^{-1})^{-1} \Phi_N^\top Y_N$$

$$\sigma_N^2(x) = \sigma^2 \varphi(x)^\top (\Phi_N^\top \Phi_N + \sigma^2 \Sigma^{-1})^{-1} \varphi(x).$$

REGULARIZED LEAST-SQUARES

When G_N is not invertible, introduce regularization parameter $\lambda \in \mathbb{R}_\star^+$.

- **Regularized** solution

$$\theta_{N,\lambda} = G_{N,\lambda}^{-1} \Phi_N^\top Y_N \text{ where } G_{N,\lambda} = \Phi_N^\top \Phi_N + \lambda I_d.$$

- Bayesian interpretation:

For **Prior** $\theta \sim \mathcal{N}(0, \Sigma)$, i.i.d. setup, Gaussian noise ($\xi_n \sim \mathcal{N}(0, \sigma^2)$),

Posterior: $\hat{f}_N(x) | x, x_1, y_1, \dots, x_N, y_N \sim \mathcal{N}(\mu_N(x), \sigma_N^2(x))$ where

$$\mu_N(x) = \varphi(x)^\top (\Phi_N^\top \Phi_N + \sigma^2 \Sigma^{-1})^{-1} \Phi_N^\top Y_N$$

$$\sigma_N^2(x) = \sigma^2 \varphi(x)^\top (\Phi_N^\top \Phi_N + \sigma^2 \Sigma^{-1})^{-1} \varphi(x).$$

- Prior $\Sigma = \frac{\sigma^2}{\lambda} I_d$ gives **regularized least-squares** $\mu_N(x) = \varphi(x)^\top \theta_{N,\lambda}$.

REGULARIZED LEAST-SQUARES

When G_N is not invertible, introduce regularization parameter $\lambda \in \mathbb{R}_\star^+$.

- **Regularized** solution

$$\theta_{N,\lambda} = G_{N,\lambda}^{-1} \Phi_N^\top Y_N \text{ where } G_{N,\lambda} = \Phi_N^\top \Phi_N + \lambda I_d.$$

- Bayesian interpretation:

For **Prior** $\theta \sim \mathcal{N}(0, \Sigma)$, i.i.d. setup, Gaussian noise ($\xi_n \sim \mathcal{N}(0, \sigma^2)$),

Posterior: $\hat{f}_N(x) | x, x_1, y_1, \dots, x_N, y_N \sim \mathcal{N}(\mu_N(x), \sigma_N^2(x))$ where

$$\begin{aligned}\mu_N(x) &= \varphi(x)^\top (\Phi_N^\top \Phi_N + \sigma^2 \Sigma^{-1})^{-1} \Phi_N^\top Y_N \\ \sigma_N^2(x) &= \sigma^2 \varphi(x)^\top (\Phi_N^\top \Phi_N + \sigma^2 \Sigma^{-1})^{-1} \varphi(x).\end{aligned}$$

- Prior $\Sigma = \frac{\sigma^2}{\lambda} I_d$ gives **regularized least-squares** $\mu_N(x) = \varphi(x)^\top \theta_{N,\lambda}$.
- Interpret λ as prior value on variance.

Study $\|\theta_\star - \theta_{N,\lambda}\|_{G_{N,\lambda}}$

Standard regression noise assumptions

- **iid samples** $(x_t)_t$ are i.i.d., $(\xi_t)_t$ are i.i.d., independent from $(x_t)_t$.

Standard regression noise assumptions

- **iid samples** $(x_t)_t$ are i.i.d., $(\xi_t)_t$ are i.i.d., independent from $(x_t)_t$.

Standard regression noise assumptions

- ▶ **iid samples** $(x_t)_t$ are i.i.d., $(\xi_t)_t$ are i.i.d., independent from $(x_t)_t$.
- ▶ **sub-Gaussian** noise: For some $\sigma^2 > 0$,

$$\forall t \in \mathbb{N}, \forall \gamma \in \mathbb{R}, \quad \ln \mathbb{E}[\exp(\gamma \xi_t)] \leq \frac{\gamma^2 \sigma^2}{2}.$$

Standard regression noise assumptions

- ▶ **iid samples** $(x_t)_t$ are i.i.d., $(\xi_t)_t$ are i.i.d., independent from $(x_t)_t$.
- ▶ **sub-Gaussian** noise: For some $\sigma^2 > 0$,

$$\forall t \in \mathbb{N}, \forall \gamma \in \mathbb{R}, \quad \ln \mathbb{E}[\exp(\gamma \xi_t)] \leq \frac{\gamma^2 \sigma^2}{2}.$$

- ▶ = for $\mathcal{N}(0, \sigma^2)$ [Exercise]

Sequential regression noise assumption

- ▶ **Predictable sequence** (not iid): x_t is \mathcal{H}_{t-1} -measurable and y_t is \mathcal{H}_t -measurable. \mathcal{H}_t : history.

Standard regression noise assumptions

- ▶ **iid samples** $(x_t)_t$ are i.i.d., $(\xi_t)_t$ are i.i.d., independent from $(x_t)_t$.
- ▶ **sub-Gaussian** noise: For some $\sigma^2 > 0$,

$$\forall t \in \mathbb{N}, \forall \gamma \in \mathbb{R}, \quad \ln \mathbb{E} \left[\exp(\gamma \xi_t) \right] \leq \frac{\gamma^2 \sigma^2}{2}.$$

- ▶ = for $\mathcal{N}(0, \sigma^2)$ [Exercise]

Sequential regression noise assumption

- ▶ **Predictable sequence** (not iid): x_t is \mathcal{H}_{t-1} -measurable and y_t is \mathcal{H}_t -measurable. \mathcal{H}_t : history.
- ▶ **Conditionally** sub-Gaussian noise: For some $\sigma^2 > 0$,

$$\forall t \in \mathbb{N}, \forall \gamma \in \mathbb{R}, \quad \ln \mathbb{E} \left[\exp(\gamma \xi_t) \middle| \mathcal{H}_{t-1} \right] \leq \frac{\gamma^2 \sigma^2}{2}.$$

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

Real-world is structured

Structured actions

Linear structure and regression

Example: Graph-linear bandits

Linear UCB, Linear TS

Infinite dimension

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$ graph with set of notes $\mathcal{V} = \{1, \dots, N\}$, and edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$.

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$ graph with set of notes $\mathcal{V} = \{1, \dots, N\}$, and edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$.

- ▶ $W = (w_{i,j})_{i,j}$ Weight matrix (non-negative weights)

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$ graph with set of notes $\mathcal{V} = \{1, \dots, N\}$, and edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$.

- ▶ $W = (w_{i,j})_{i,j}$ Weight matrix (non-negative weights)
- ▶ $D = \text{Diag}((\sum_j w_{i,j})_i)$ Degree matrix

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$ graph with set of notes $\mathcal{V} = \{1, \dots, N\}$, and edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$.

- ▶ $W = (w_{i,j})_{i,j}$ Weight matrix (non-negative weights)
- ▶ $D = \text{Diag}((\sum_j w_{i,j})_i)$ Degree matrix
- ▶ $L = D - W$ graph Laplacian matrix

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$ graph with set of notes $\mathcal{V} = \{1, \dots, N\}$, and edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$.

- ▶ $W = (w_{i,j})_{i,j}$ Weight matrix (non-negative weights)
- ▶ $D = \text{Diag}((\sum_j w_{i,j})_i)$ Degree matrix
- ▶ $L = D - W$ graph Laplacian matrix

Properties:

- ▶ L is symmetric, positive, semi-definite.
- ▶ Eigenvalues : $0 = \lambda_1 \leqslant \lambda_2 \leqslant \dots \leqslant \lambda_N$
Let $L = Q^\top \Lambda Q$ where
 - ▶ Λ : $N \times N$ diagonal matrix with eigenvalues of L
 - ▶ Q : $N \times N$ matrix whose columns are eigenvectors of L .

Any graph-function f decomposes as $f = Q\alpha$ for some α , that is

- ▶ $f(i) = \sum_{j \in \mathcal{V}} \alpha_j Q_{i,j} = \langle \alpha, q(i) \rangle$ where $q(i) = (Q_{i,j})_j$ is i^{th} eigenvector.
- ▶ Then, $f^\top L f = \frac{1}{2} \sum_{i,j \leq N} w_{i,j} (f_i - f_j)^2 = \sum_{i \in \mathcal{V}} \lambda_i \alpha_i^2 = \|\alpha\|_\Lambda \stackrel{\text{def}}{=} \|f\|_{\mathcal{G}}$

GRAPH SMOOTHNESS

Any graph-function f decomposes as $f = Q\alpha$ for some α , that is

- ▶ $f(i) = \sum_{j \in \mathcal{V}} \alpha_j Q_{i,j} = \langle \alpha, q(i) \rangle$ where $q(i) = (Q_{i,j})_j$ is i^{th} eigenvector.
 - ▶ Then, $f^\top L f = \frac{1}{2} \sum_{i,j \leq N} w_{i,j} (f_i - f_j)^2 = \sum_{i \in \mathcal{V}} \lambda_i \alpha_i^2 = \|\alpha\|_\Lambda \stackrel{\text{def}}{=} \|f\|_{\mathcal{G}}$
- ⇒ **Linear space** induced by the Graph:

$$\mathcal{F}_{\mathcal{G}} = \{f : f(x) = \langle \alpha, q(x) \rangle, \|\alpha\|_\Lambda \leq 1\}$$

Low-norm $\|f\|_{\mathcal{G}}$ means:

- ▶ $(f_i - f_j)^2$ is small if $w_{i,j}$ is large
- ▶ **similar value** between **neighbor nodes**.

GRAPH SMOOTHNESS

Any graph-function f decomposes as $f = Q\alpha$ for some α , that is

- ▶ $f(i) = \sum_{j \in \mathcal{V}} \alpha_j Q_{i,j} = \langle \alpha, q(i) \rangle$ where $q(i) = (Q_{i,j})_j$ is i^{th} eigenvector.
 - ▶ Then, $f^\top L f = \frac{1}{2} \sum_{i,j \leq N} w_{i,j} (f_i - f_j)^2 = \sum_{i \in \mathcal{V}} \lambda_i \alpha_i^2 = \|\alpha\|_\Lambda \stackrel{\text{def}}{=} \|f\|_{\mathcal{G}}$
- ⇒ **Linear space** induced by the Graph:

$$\mathcal{F}_{\mathcal{G}} = \{f : f(x) = \langle \alpha, q(x) \rangle, \|\alpha\|_\Lambda \leq 1\}$$

Low-norm $\|f\|_{\mathcal{G}}$ means:

- ▶ $(f_i - f_j)^2$ is small if $w_{i,j}$ is large
- ▶ **similar value** between **neighbor nodes**.

Further references for bandits on graphs:

- ▶ Michal Valko, Rémi Munos, Branislav Kveton, Tomás Kocák: *Spectral Bandits for Smooth Graph Functions*, in International Conference on Machine Learning (ICML 2014).
- ▶ Alexandra Carpentier, Michal Valko: *Revealing graph bandits for maximizing local influence*, in International Conference on Artificial Intelligence and Statistics (AISTATS 2016).

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

Real-world is structured

Structured actions

Linear structure and regression

Example: Graph-linear bandits

Linear UCB, Linear TS

Infinite dimension

- ▷ Action $x \in \mathcal{X}$, Reward: $y = r(x) = \langle \theta, \varphi(x) \rangle + \text{noise}$.
- ▶ **Least-squares (regularized) estimate** of θ_\star :

$$\theta_{t,\lambda} = \underbrace{[\Phi_t^\top \Phi_t + \lambda I_d]^{-1}}_{G_{t,\lambda}} \Phi_t^\top Y_t.$$

- ▶ Choose $X_{t+1} = \operatorname{argmax}_{x \in \mathcal{X}} \langle \theta_{t,\lambda}, \varphi(x) \rangle$.
(cf. API, AVI, LSPI, etc.)

- ▷ Action $x \in \mathcal{X}$, Reward: $y = r(x) = \langle \theta, \varphi(x) \rangle + \text{noise}$.
- ▶ **Least-squares (regularized) estimate** of θ_\star :

$$\theta_{t,\lambda} = \underbrace{[\Phi_t^\top \Phi_t + \lambda I_d]^{-1}}_{G_{t,\lambda}} \Phi_t^\top Y_t.$$

- ▶ Choose $X_{t+1} = \operatorname{argmax}_{x \in \mathcal{X}} \langle \theta_{t,\lambda}, \varphi(x) \rangle$.
(cf. API, AVI, LSPI, etc.)
- ⇒ Exploitation only !

Optimism in Face of Uncertainty - Linear

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári "Improved Algorithms for
Linear Stochastic Bandits"
NIPS, 2011.

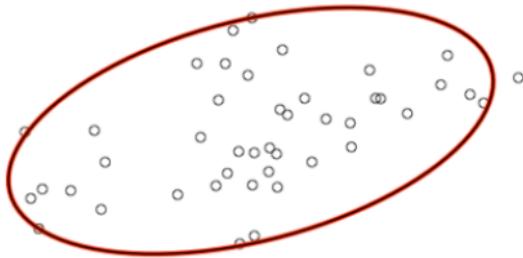
$$X_{t+1} = \operatorname{argmax}_{x \in \mathcal{X}} \max \left\{ f_\theta(x) : \theta \text{ is plausible} \right\}$$

$$X_{t+1} = \operatorname{argmax}_{x \in \mathcal{X}} \max \left\{ f_\theta(x) : \theta \text{ is plausible} \right\}$$

- ▶ Plausible: $C_t(\delta) = \left\{ \theta : \|\theta - \theta_{t,\lambda}\|_{G_{t,\lambda}} \leq B_t(\delta) \right\}$

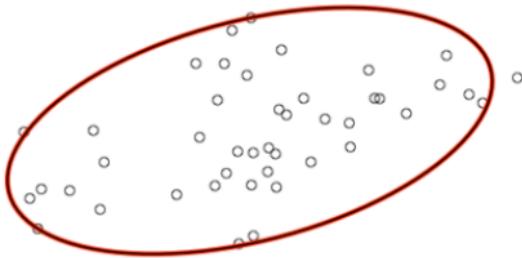
$$X_{t+1} = \operatorname{argmax}_{x \in \mathcal{X}} \max \left\{ f_\theta(x) : \theta \text{ is plausible} \right\}$$

- ▶ Plausible: $C_t(\delta) = \left\{ \theta : \|\theta - \theta_{t,\lambda}\|_{G_{t,\lambda}} \leq B_t(\delta) \right\}$
- ▶ Confidence ellipsoid such that $\mathbb{P}(\theta_* \in C_t(\delta)) \geq 1 - \delta$.



$$X_{t+1} = \operatorname{argmax}_{x \in \mathcal{X}} \max \left\{ f_\theta(x) : \theta \text{ is plausible} \right\}$$

- ▶ Plausible: $C_t(\delta) = \left\{ \theta : \|\theta - \theta_{t,\lambda}\|_{G_{t,\lambda}} \leq B_t(\delta) \right\}$
- ▶ Confidence ellipsoid such that $\mathbb{P}(\theta_* \in C_t(\delta)) \geq 1 - \delta$.



- ▶ Explicit solution

$$X_{t+1} = \operatorname{argmax}_{x \in \mathcal{X}} \langle \theta_{t,\lambda}, \varphi(x) \rangle + B_t(\delta) \|\varphi(x)\|_{G_{t,\lambda}^{-1}}.$$

⇒ UCB-style exploitation and exploration trade-off!

How to build $B_t(\delta)$?

How to build $B_t(\delta)$?

- ▶ OFUL (Abbasi et al, 2011)

$$B_t(\delta) = \sqrt{\lambda} \|\theta^*\|_2 + \sqrt{2 \ln \left(\frac{\det(G_N + \lambda I)^{1/2}}{\delta \lambda^{d/2}} \right)}$$

$$|f_{\theta^*}(x) - f_{\theta_{N,\lambda}}(x)| \leq \|\theta^* - \theta_{N,\lambda}\|_{G_{N,\lambda}} \|\varphi(x)\|_{G_{N,\lambda}^{-1}}$$

Decomposition lemma

$$\|\theta^* - \theta_{N,\lambda}\|_{G_{N,\lambda}} \leq \sqrt{\lambda} \|\theta^*\|_2 + \|\Phi_N^\top E_N\|_{G_{N,\lambda}^{-1}}$$

where $E_N = (\xi_1, \dots, \xi_N)^\top \in \mathbb{R}^N$.

Key observation: sum of **conditionally centered vector** variables

$$\Phi_N^\top E_N = \sum_{n=1}^N \varphi(x_n) \xi_n \in \mathbb{R}^d.$$

⇒ **Concentration inequality for vectors !**

Make use of **self-normalized** concentration inequalities.

$$\begin{aligned}
\theta^* - \theta_{N,\lambda} &= \theta^* - G_{N,\lambda}^{-1} \Phi_N^\top Y_N \\
&= \theta^* - G_{N,\lambda}^{-1} \Phi_N^\top (\Phi_N \theta^* + E_N) \\
&= (I - G_{N,\lambda}^{-1} G_N) \theta^* - G_{N,\lambda}^{-1} \Phi_N^\top E_N \\
&= G_{N,\lambda}^{-1} (G_{N,\lambda} - G_N) \theta^* - G_{N,\lambda}^{-1} \Phi_N^\top E_N . \\
&= \underbrace{\lambda G_{N,\lambda}^{-1} \theta^*}_{(1)} - \underbrace{G_{N,\lambda}^{-1} \Phi_N^\top E_N}_{(2)} .
\end{aligned}$$

$$\begin{aligned}
(1) \quad \|\lambda G_{N,\lambda}^{-1} \theta^*\|_{G_{N,\lambda}} &= \lambda \sqrt{\theta^{*\top} G_{N,\lambda}^{-1} G_{N,\lambda} G_{N,\lambda}^{-1} \theta^*} \\
&\leq \frac{\lambda}{\sqrt{\text{eig}_{\min}(G_{N,\lambda})}} \|\theta^*\|_2 \leq \sqrt{\lambda} \|\theta^*\|_2
\end{aligned}$$

$$(2) \quad \|G_{N,\lambda}^{-1} \Phi_N^\top E_N\|_{G_{N,\lambda}} = \|\Phi_N^\top E_N\|_{G_{N,\lambda}^{-1}} .$$

What it means to be **self-normalized** ?

In dimension $D = 1$, $\lambda = 0$, $G_N = \sum_{n=1}^N \varphi(x_n)^2$

$$\|\Phi_N^\top E_N\|_{G_{N,\lambda}^{-1}} = \frac{|\sum_{n=1}^N \varphi(x_n)\xi_n|}{\sqrt{\sum_{n=1}^N \varphi(x_n)^2}} = \frac{|\sum_{n=1}^N Z_n|}{\sqrt{\sum_{n=1}^N \sigma_n^2}}$$

Basic self-normalized (Gaussian) concentration inequality

For fixed t , Z_1, \dots, Z_t , independent, $Z_n \sim \mathcal{N}(0, \sigma_n^2)$, $\delta \in (0, 1]$

$$\mathbb{P}\left(\left|\frac{\sum_{n=1}^t Z_n}{\sqrt{\sum_{n=1}^t \sigma_n^2}}\right| \geqslant \sqrt{2 \ln(2/\delta)}\right) \leqslant \delta$$

Basic (Gaussian) concentration inequality For fixed t , Z_1, \dots, Z_t i.i.d. $\mathcal{N}(0, \sigma^2)$, $\delta \in (0, 1]$

$$\mathbb{P}\left(\frac{1}{t} \sum_{n=1}^t Z_n \geqslant \sqrt{\frac{2\sigma^2 \ln(1/\delta)}{t}}\right) \leqslant \delta$$

Likewise, using the Chernoff-method, we can show for fixed t , Z_1, \dots, Z_t , independent, $Z_n \sim \mathcal{N}(0, \sigma_n^2)$, $\delta \in (0, 1]$

$$\mathbb{P}\left(\sum_{n=1}^t Z_n \geqslant \sqrt{2 \sum_{n=1}^t \sigma_n^2 \ln(1/\delta)}\right) \leqslant \delta$$

Thus

$$\mathbb{P}\left(\frac{\sum_{n=1}^t Z_n}{\sqrt{\sum_{n=1}^t \sigma_n^2}} \geqslant \sqrt{2 \ln(1/\delta)}\right) \leqslant \delta$$

Extension to dimension d by the **Laplace method** (De la Peña et al., 2004).

Let $Z \in \mathbb{R}^d$ random **vector**, B a $d \times d$ random **matrix** such that

$$(\text{ Sub-Gaussian}) \quad \forall \gamma \in \mathbb{R}^d, \quad \ln \mathbb{E}[\exp(\gamma^\top Z - \frac{1}{2}\gamma^\top B\gamma)] \leq 0.$$

Then for any deterministic $d \times d$ matrix C , w.p. $\geq 1 - \delta$,

$$\|Z\|_{(B+C)^{-1}} \leq \sqrt{2 \ln \left(\frac{\det(B+C)^{1/2}}{\delta \det(C)^{1/2}} \right)}.$$

Extension to dimension d by the **Laplace method** (De la Peña et al., 2004).

Let $Z \in \mathbb{R}^d$ random **vector**, B a $d \times d$ random **matrix** such that

$$(\text{Sub-Gaussian}) \quad \forall \gamma \in \mathbb{R}^d, \quad \ln \mathbb{E}[\exp(\gamma^\top Z - \frac{1}{2}\gamma^\top B\gamma)] \leq 0.$$

Then for any deterministic $d \times d$ matrix C , w.p. $\geq 1 - \delta$,

$$\|Z\|_{(B+C)^{-1}} \leq \sqrt{2 \ln \left(\frac{\det(B+C)^{1/2}}{\delta \det(C)^{1/2}} \right)}.$$

- ▶ Application: $Z = \sum_{n=1}^N \varphi(x_n) \xi_n$, $B = G_{N,0}$ $C = \lambda I_d$.

1) Quantity

$$M_t^\gamma = \exp \left(\langle \gamma, Z \rangle - \frac{1}{2} \|\lambda\|_B^2 \right)$$

is a super martingale such that for all t , $\mathbb{E}[M_t^\gamma] \leq 1$.

1) Quantity

$$M_t^\gamma = \exp \left(\langle \gamma, Z \rangle - \frac{1}{2} \|\lambda\|_B^2 \right)$$

is a super martingale such that for all t , $\mathbb{E}[M_t^\gamma] \leq 1$.

2) Choice of γ ? Replace optimization with integration (Laplace) !

Introduce distribution $\Lambda \sim \mathcal{N}(0, C^{-1})$, and M_t^Λ .

1) Quantity

$$M_t^\gamma = \exp \left(\langle \gamma, Z \rangle - \frac{1}{2} \|\lambda\|_B^2 \right)$$

is a super martingale such that for all t , $\mathbb{E}[M_t^\gamma] \leq 1$.

2) Choice of γ ? Replace optimization with integration (Laplace) !

Introduce distribution $\Lambda \sim \mathcal{N}(0, C^{-1})$, and M_t^Λ .

- a) $\mathbb{E}[M_t^\Lambda] \leq 1$
- b) $\mathbb{E}[M_t^\Lambda] = \mathbb{E}[\mathbb{E}[M_t^\Lambda | \mathcal{F}_\infty]]$ and

$$\mathbb{E}[M_t^\Lambda | \mathcal{F}_\infty] = \int_{\mathbb{R}^d} \exp \left(\langle \gamma, Z \rangle - \frac{1}{2} \|\lambda\|_B^2 \right) f(\lambda) d\lambda$$

where f denotes the pdf of $\Lambda \sim \mathcal{N}(0, C^{-1})$.

3) Direct calculations show that

$$\mathbb{E}[M_t^\Lambda | \mathcal{F}_\infty] = \left(\frac{\det(C)}{\det(B+C)} \right)^{1/2} \exp \left(\frac{1}{2} \|Z\|_{(B+C)^{-1}}^2 \right)$$

Then $\mathbb{E} \left[\left(\frac{\det(C)}{\det(B+C)} \right)^{1/2} \exp \left(\frac{1}{2} \|Z\|_{(B+C)^{-1}}^2 \right) \right] \leq 1$

4) Markov inequality yields:

$$\begin{aligned} & \mathbb{P} \left(\|Z\|_{(B+C)^{-1}}^2 > 2 \ln \left(\frac{\det(B+C)^{1/2}}{\delta \det(B)^{1/2}} \right) \right) \\ &= \mathbb{P} \left(\exp \left(\frac{1}{2} \|Z\|_{(B+C)^{-1}}^2 \right) > \frac{\det(B+C)^{1/2}}{\delta \det(B)^{1/2}} \right) \leq \delta. \end{aligned}$$

► Application: $Z = \sum_{n=1}^N \varphi(x_n) \xi_n$, $B = G_{N,0}$ $C = \lambda I_d$.

$$\mathbb{P}\left(\|\Phi_N^\top E_N\|_{G_{N,\lambda}^{-1}} \geq 2 \ln\left(\frac{\det(G_{N,\lambda})^{1/2}}{\delta \lambda^{d/2}}\right)\right) \leq \delta.$$

- ▶ Application: $Z = \sum_{n=1}^N \varphi(x_n) \xi_n$, $B = G_{N,0}$ $C = \lambda I_d$.
$$\mathbb{P}\left(\|\Phi_N^\top E_N\|_{G_{N,\lambda}^{-1}} \geq 2 \ln\left(\frac{\det(G_{N,\lambda})^{1/2}}{\delta \lambda^{d/2}}\right)\right) \leq \delta.$$
- ▶ Time-uniform bound ($\forall N$): handles random stopping time N .

- Application: $Z = \sum_{n=1}^N \varphi(x_n) \xi_n$, $B = G_{N,0}$ $C = \lambda I_d$.

$$\mathbb{P}\left(\|\Phi_N^\top E_N\|_{G_{N,\lambda}^{-1}} \geq 2 \ln\left(\frac{\det(G_{N,\lambda})^{1/2}}{\delta \lambda^{d/2}}\right)\right) \leq \delta.$$

- Time-uniform bound ($\forall N$): handles random stopping time N .
- Property:

$$\mathbb{E}[M_N^\Delta] = \mathbb{E}[\liminf_{m \rightarrow \infty} M_{\min(N,m)}^\Delta] \leq \liminf_{m \rightarrow \infty} \mathbb{E}[M_{\min(N,m)}^\Delta] \leq 1.$$

⇒ **Confidence ellipsoid** on θ_\star :

$$C_t(\delta) = \left\{ \theta : \|\theta - \theta_{t,\lambda}\|_{G_{t,\lambda}} \leq \sqrt{\lambda} \|\theta^*\|_2 + \sqrt{2 \ln\left(\frac{\det(G_t + \lambda I)^{1/2}}{\delta \lambda^{d/2}}\right)} \right\},$$

Information gain γ_T

Log-determinant Lemma

$$\gamma_T = \ln \left(\frac{\det(G_{T,\lambda})}{\det(\lambda I_d)} \right) = \sum_{t=1}^T \ln (1 + \|\varphi(x_t)\|_{G_{t-1,\lambda}}^2)$$

where $G_{t,\lambda} = G_t + \lambda I$.

Information gain γ_T

Log-determinant Lemma

$$\gamma_T = \ln \left(\frac{\det(G_{T,\lambda})}{\det(\lambda I_d)} \right) = \sum_{t=1}^T \ln (1 + \|\varphi(x_t)\|_{G_{t-1,\lambda}}^2)$$

where $G_{t,\lambda} = G_t + \lambda I$.

- ▶ $\det(\lambda I_d)$: volume before observing data; $\det(G_{T,\lambda})$: volume after observing x_1, \dots, x_T .

Information gain γ_T

Log-determinant Lemma

$$\gamma_T = \ln \left(\frac{\det(G_{T,\lambda})}{\det(\lambda I_d)} \right) = \sum_{t=1}^T \ln (1 + \|\varphi(x_t)\|_{G_{t-1,\lambda}}^2)$$

where $G_{t,\lambda} = G_t + \lambda I$.

- ▶ $\det(\lambda I_d)$: volume before observing data; $\det(G_{T,\lambda})$: volume after observing x_1, \dots, x_t .
- ▶ Captures how much the "**volume of information**" is modified by samples x_1, \dots, x_t .

Information gain γ_T

Log-determinant Lemma

$$\gamma_T = \ln \left(\frac{\det(G_{T,\lambda})}{\det(\lambda I_d)} \right) = \sum_{t=1}^T \ln \left(1 + \|\varphi(x_t)\|_{G_{t-1,\lambda}}^2 \right)$$

where $G_{t,\lambda} = G_t + \lambda I$.

- ▶ $\det(\lambda I_d)$: volume before observing data; $\det(G_{T,\lambda})$: volume after observing x_1, \dots, x_T .
- ▶ Captures how much the "**volume of information**" is modified by samples x_1, \dots, x_T .
- ▶ $\gamma_T = O(d \ln(T))$ for d -dimensional linear space.

$$\begin{aligned}
\det(G_{n,\lambda}) &= \det(G_{n-1,\lambda} + \varphi(x_n)\varphi(x_n)^\top) \\
&= \det(G_{n-1,\lambda}) \det(I + G_{n-1,\lambda}^{-1/2}\varphi(x_n)(G_{n-1,\lambda}^{-1/2}\varphi(x_n))^\top) \\
&= \det(G_{n-1,\lambda})(1 + \|\varphi(x_n)\|_{G_{n-1,\lambda}^{-1}}^2) \\
&= \det(\lambda I) \prod_{t=1}^n (1 + \|\varphi(x_t)\|_{G_{t-1,\lambda}^{-1}}^2)
\end{aligned}$$

Thus,

$$\ln \left(\frac{\det(G_{n,\lambda})}{\lambda^d} \right) = \sum_{t=1}^n \ln (1 + \|\varphi(x_t)\|_{G_{t-1,\lambda}^{-1}}^2)$$

- ▷ We have good **confidence bounds**: let us exploit them!
- ▷ Simplest **optimistic** approach:

$$\begin{aligned} X_{t+1} &= \operatorname{argmax}_{x \in \mathcal{X}} \max\{\langle \theta, \varphi(x) \rangle : \theta \in \mathcal{C}_t(\delta)\}. \\ &= \operatorname{argmax}_{x \in \mathcal{X}} f_t^+(x) \end{aligned}$$

Regret

If $f_\star(x) \in [-1, 1]$ for all x , then w.p. higher than $1 - \delta$,

$$\mathcal{R}_T = O\left(\sqrt{T\gamma_T}\left(\|\theta_\star\|_2 + \sigma\sqrt{2\ln(1/\delta) + 2\gamma_T}\right)\right)$$

- ▷ We have good **confidence bounds**: let us exploit them!
- ▷ Simplest **optimistic** approach:

$$\begin{aligned} X_{t+1} &= \operatorname{argmax}_{x \in \mathcal{X}} \max\{\langle \theta, \varphi(x) \rangle : \theta \in \mathcal{C}_t(\delta)\}. \\ &= \operatorname{argmax}_{x \in \mathcal{X}} f_t^+(x) \end{aligned}$$

Regret

If $f_\star(x) \in [-1, 1]$ for all x , then w.p. higher than $1 - \delta$,

$$\mathcal{R}_T = O\left(\sqrt{T\gamma_T}\left(\|\theta_\star\|_2 + \sigma\sqrt{2\ln(1/\delta) + 2\gamma_T}\right)\right)$$

- ▷ Q: Is this optimal way of exploiting linear structure (cf. Bayesian, Likelihood-based, subsampling methods)?

Instantaneous regret r_t (note: $r_t \leq 2$)

$$\begin{aligned} r_t &= f_\star(x_\star) - f_\star(x_t) \\ &\leq f_{t-1}^+(x_t) - f_\star(x_t) \text{ with high probability} \\ &\leq |f_{t-1}^+(x_t) - f_{\lambda,t-1}(x_t)| + |f_{\lambda,t-1}(x_t) - f_\star(x_t)| \\ &\leq 2\|\varphi(x_t)\|_{G_{t,\lambda}^{-1}} B_{t-1}(\delta). \end{aligned}$$

Thus, we deduce that with probability higher than $1 - \delta$:

$$\begin{aligned} \mathfrak{R}_T &= \sum_{t=1}^T r_t \leq \sum_{t=1}^T 2 \min\{\|\varphi(x_t)\|_{G_{t,\lambda}^{-1}} B_{t-1}(\delta), 1\} \\ &\leq 2B_T(\delta) \sum_{t=1}^T \min\{\|\varphi(x_t)\|_{G_{t,\lambda}^{-1}}, 1\} \\ &\leq 2B_T(\delta) \sqrt{T \sum_{t=1}^T \min\{\|\varphi(x_t)\|_{G_{t,\lambda}^{-1}}^2, 1\}}. \end{aligned}$$

We conclude remarking that $\min\{A, 1\} \leq \frac{\ln(1+A)}{\ln(2)}$ for all $A \geq 0$.

Thompson Sampling for Linear - Bandits

Shipra Agrawal, Navin Goyal "Thompson Sampling for Contextual Bandits with Linear Payoffs"
arXiv:1209.3352, 2014.

► Bayesian model:

$$y_t = x_t^T \theta + \varepsilon_t, \quad \theta \sim \mathcal{N}(0, \kappa^2 I_d), \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2).$$

Explicit posterior: $p(\theta|x_1, y_1, \dots, x_t, y_t) = \mathcal{N}(\hat{\theta}(t), \Sigma_t)$.

► Bayesian model:

$$y_t = x_t^T \theta + \varepsilon_t, \quad \theta \sim \mathcal{N}(0, \kappa^2 I_d), \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2).$$

Explicit posterior: $p(\theta|x_1, y_1, \dots, x_t, y_t) = \mathcal{N}(\hat{\theta}(t), \Sigma_t)$.

► Thompson Sampling

$$\begin{aligned}\tilde{\theta}(t) &\sim \mathcal{N}(\hat{\theta}(t), \Sigma_t), \\ x_{t+1} &= \underset{x \in \mathcal{D}_{t+1}}{\operatorname{argmax}} x^T \tilde{\theta}(t).\end{aligned}$$

[Li et al. 12],[Agrawal & Goyal 13]

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

Real-world is structured

Structured actions

Linear structure and regression

Example: Graph-linear bandits

Linear UCB, Linear TS

Infinite dimension

- ▷ Linear function space with finite dimension $\theta \in \mathbb{R}^d$, finitely many feature functions is **limited**.
- ▷ What about considering all polynomials of arbitrary degree ? all fourier basis ? all wavelet basis ?
- ▷ We need to extend theory from finite dimension to infinite dimension.
- ▷ Offers great flexibility, can work with virtually any reasonable space (RKHS).

Let k be a kernel function (continuous, symmetric positive definite) on a compact \mathcal{X} with positive finite Borel measure μ .

There exists an at most **countable** sequence $(\sigma_i, \psi_i)_{i \in \mathbb{N}^*}$ where $\sigma_i \geq 0$, $\lim_{i \rightarrow \infty} \sigma_i = 0$ and $\{\psi_i\}$ form an orthonormal basis of $L_{2,\mu}(\mathcal{X})$, such that

$$k(x, y) = \sum_{j=1}^{\infty} \sigma_j \psi_j(x) \psi_j(y) \quad \text{and} \quad \|f\|_{\mathcal{K}}^2 = \sum_{j=1}^{\infty} \frac{\langle f, \psi_j \rangle_{L_{2,\mu}}^2}{\sigma_j}$$

Let $\varphi_i = \sqrt{\sigma_i} \psi_i$ (hence $\|\varphi_i\|_{L_2} = \sqrt{\sigma_i}$, $\|\varphi_i\|_{\mathcal{K}} = 1$.)

If $f = \sum_i \theta_i \varphi_i$, then $\|f\|_{\mathcal{K}}^2 = \sum_i \theta_i^2$.

Similar to parametric regression except with infinite parameter.

Let k be a kernel function.

In the parametric case, we built $\theta_{\lambda,t}$, then $f_{\lambda,t}(x) = \langle \theta_{\lambda,t}, \varphi(x) \rangle$.

After observing $Y_t = (y_1, \dots, y_t)^\top \in \mathbb{R}^t$, we now build directly:

$$\text{(Kernel estimate)} \quad f_{\lambda,t}(x) = k_t(x)^\top (K_t + \lambda I_t)^{-1} Y_t,$$

where

Let k be a kernel function.

In the parametric case, we built $\theta_{\lambda,t}$, then $f_{\lambda,t}(x) = \langle \theta_{\lambda,t}, \varphi(x) \rangle$.

After observing $Y_t = (y_1, \dots, y_t)^\top \in \mathbb{R}^t$, we now build directly:

$$\text{(Kernel estimate)} \quad f_{\lambda,t}(x) = k_t(x)^\top (K_t + \lambda I_t)^{-1} Y_t,$$

where

- ▶ $k_t(x) = (k(x, x_{t'}))_{t' \leq t} \in \mathbb{R}^t$,

Let k be a kernel function.

In the parametric case, we built $\theta_{\lambda,t}$, then $f_{\lambda,t}(x) = \langle \theta_{\lambda,t}, \varphi(x) \rangle$.

After observing $Y_t = (y_1, \dots, y_t)^\top \in \mathbb{R}^t$, we now build directly:

$$\text{(Kernel estimate)} \quad f_{\lambda,t}(x) = k_t(x)^\top (K_t + \lambda I_t)^{-1} Y_t,$$

where

- ▶ $k_t(x) = (k(x, x_{t'}))_{t' \leq t} \in \mathbb{R}^t$,

Let k be a kernel function.

In the parametric case, we built $\theta_{\lambda,t}$, then $f_{\lambda,t}(x) = \langle \theta_{\lambda,t}, \varphi(x) \rangle$.

After observing $Y_t = (y_1, \dots, y_t)^\top \in \mathbb{R}^t$, we now build directly:

$$\text{(Kernel estimate)} \quad f_{\lambda,t}(x) = k_t(x)^\top (K_t + \lambda I_t)^{-1} Y_t,$$

where

- ▶ $k_t(x) = (k(x, x_{t'}))_{t' \leq t} \in \mathbb{R}^t$,
- ▶ $K_t = (k(x_s, x_{s'}))_{s, s' \leq t} \in \mathbb{R}^{t \times t}$,

for a parameter $\lambda \in \mathbb{R}$.

Theorem (Durand & M. 2017, Kernel estimation error)

$\forall \delta \in [0, 1]$, with probability higher than $1 - \delta$, it holds simultaneously over all $x \in \mathcal{X}$ and $t \geq 0$,

$$|f_*(x) - f_{\lambda,t}(x)| \leq \sqrt{k_{\lambda,t}(x, x)} \left[\|f_*\|_k + \frac{\sigma}{\sqrt{\lambda}} \sqrt{2 \ln(1/\delta) + 2\gamma_t(\lambda)} \right],$$

where

Theorem (Durand & M. 2017, Kernel estimation error)

$\forall \delta \in [0, 1]$, with probability higher than $1 - \delta$, it holds simultaneously over all $x \in \mathcal{X}$ and $t \geq 0$,

$$|f_*(x) - f_{\lambda,t}(x)| \leq \sqrt{k_{\lambda,t}(x, x)} \left[\|f_*\|_k + \frac{\sigma}{\sqrt{\lambda}} \sqrt{2 \ln(1/\delta) + 2\gamma_t(\lambda)} \right],$$

where

- ▶ $k_{\lambda,t}(x, x) = k(x, x) - k_t(x)^\top (K_t + \lambda I_t)^{-1} k_t(x)$: **posterior variance**.

Theorem (Durand & M. 2017, Kernel estimation error)

$\forall \delta \in [0, 1]$, with probability higher than $1 - \delta$, it holds simultaneously over all $x \in \mathcal{X}$ and $t \geq 0$,

$$|f_*(x) - f_{\lambda,t}(x)| \leq \sqrt{k_{\lambda,t}(x, x)} \left[\|f_*\|_k + \frac{\sigma}{\sqrt{\lambda}} \sqrt{2 \ln(1/\delta) + 2\gamma_t(\lambda)} \right],$$

where

- ▶ $k_{\lambda,t}(x, x) = k(x, x) - k_t(x)^\top (K_t + \lambda I_t)^{-1} k_t(x)$: **posterior variance**.
- ▶ $\gamma_t(\lambda) = \frac{1}{2} \sum_{t'=1}^t \ln \left(1 + \frac{1}{\lambda} k_{\lambda,t'-1}(x_{t'}, x_{t'}) \right)$: **information gain**.

Theorem (Durand & M. 2017, Kernel estimation error)

$\forall \delta \in [0, 1]$, with probability higher than $1 - \delta$, it holds simultaneously over all $x \in \mathcal{X}$ and $t \geq 0$,

$$|f_*(x) - f_{\lambda,t}(x)| \leq \sqrt{k_{\lambda,t}(x,x)} B_{\lambda,t-1}(\delta),$$

where

- ▶ $k_{\lambda,t}(x,x) = k(x,x) - k_t(x)^\top (K_t + \lambda I_t)^{-1} k_t(x)$: **posterior variance**.
- ▶ $\gamma_t(\lambda) = \frac{1}{2} \sum_{t'=1}^t \ln \left(1 + \frac{1}{\lambda} k_{\lambda,t'-1}(x_{t'}, x_{t'}) \right)$: **information gain**.
- ▶ $\|f_*\|_k$: Reproducing Kernel Hilbert Space norm.

$k(x, x')$	Captures	γ_T
$\langle x, x' \rangle$	"Linear functions"	$O(d \ln(T))$
$\exp(-\frac{\ x-x'\ ^2}{2\ell^2})$	"Smooth functions"	$O(\ln(T)^{d+1})$
...

Many kernels, for different properties of the signal
(graph-smoothness, periodic, change points, etc.)

Minimize the regret: $\mathcal{R}_T = \sum_{t=1}^T f_\star(\star) - f_\star(x_t)$.

Kernel-UCB

$$x_t \in \operatorname{argmax}_{x \in \mathcal{X}} f_t^+(x) \quad \text{where } f_t^+(x) = f_{\lambda, t-1}(x) + \sqrt{k_{\lambda, t-1}(x, x)} B_{\lambda, t-1}(\delta).$$

Kernel-TS (on discrete set $\mathbb{X} \subset \mathcal{X}$)

$$x_t \in \operatorname{argmax}_{x \in \mathbb{X}} \tilde{f}_t(x) \quad \text{where } \tilde{f}_t \sim \mathcal{N}(\hat{f}_{t-1}, \hat{\Sigma}_{t-1}) \quad \text{with}$$

$$\hat{f}_{t-1} = (f_{\lambda, t-1}(x))_{x \in \mathbb{X}}, \quad \hat{\Sigma}_{t-1} = (k_{\lambda, t-1}(x, x') B_{\lambda, t-1}(\delta)^2)_{x, x' \in \mathbb{X}}.$$

More info in (Durand et al., 2018, JMLR)

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

OPTIMAL STRUCTURE EXPLOITATION

CONCLUSION, PERSPECTIVE

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

OPTIMAL STRUCTURE EXPLOITATION

Structured lower bounds

Lipschitz bandits

Structure-exploiting strategies

Metric-graph of bandits

Equivariant bandits



REGRET LOWER BOUNDS

Set of optimal arms for $\nu = (\nu_a)_{a \in \mathcal{A}}$: $\mathcal{A}_*(\nu) = \text{Argmax}_{a \in \mathcal{A}} \mu_a(\nu)$.

Definition (Uniformly Good strategies)

A bandit strategy is **uniformly-good** on \mathcal{D} if

$$\forall \nu = (\nu_a)_{a \in \mathcal{A}} \in \mathcal{D}, \forall a \notin \mathcal{A}_*(\nu), \quad \mathbb{E}[N_T(a)] = o(T^\alpha) \quad \text{for all } \alpha \in (0, 1].$$

Theorem ((Lai, Robbins 85) “Price for being uniformly-good”)

Any uniformly good strategy on $\mathcal{D} = \text{Bern}^{\mathcal{A}}$ must satisfy

$$\forall a \notin \mathcal{A}_*(\nu) \quad \liminf_{T \rightarrow \infty} \frac{\mathbb{E}_\nu[N_T(a)]}{\ln(T)} \geq \frac{1}{\text{kl}(\mu_a(\nu), \mu_*(\nu))}.$$

REGRET LOWER BOUNDS

Set of optimal arms for $\nu = (\nu_a)_{a \in \mathcal{A}}$: $\mathcal{A}_*(\nu) = \text{Argmax}_{a \in \mathcal{A}} \mu_a(\nu)$.

Definition (**Uniformly Good strategies**)

A bandit strategy is **uniformly-good** on \mathcal{D} if

$$\forall \nu = (\nu_a)_{a \in \mathcal{A}} \in \mathcal{D}, \forall a \notin \mathcal{A}_*(\nu), \quad \mathbb{E}[N_T(a)] = o(T^\alpha) \quad \text{for all } \alpha \in (0, 1].$$

Theorem ((Lai, Robbins 85) “Price for being uniformly-good”)

Any uniformly good strategy on $\mathcal{D} = \text{Bern}^{\mathcal{A}}$ must satisfy

$$\forall a \notin \mathcal{A}_*(\nu) \quad \liminf_{T \rightarrow \infty} \frac{\mathbb{E}_\nu[N_T(a)]}{\ln(T)} \geq \frac{1}{\text{kl}(\mu_a(\nu), \mu_*(\nu))}.$$

Main tool: **Change of measure**

(Probability) $\forall \Omega, \forall c \in \mathbb{R}, \quad \mathbb{P}_\nu\left(\Omega \cap \left\{ \ln\left(\frac{d\nu}{d\tilde{\nu}}(X)\right) \leq c \right\} \right) \leq \exp(c) \mathbb{P}_{\tilde{\nu}}(\Omega).$

(Expectation) $\mathbb{E}_\nu\left[\ln\left(\frac{d\nu}{d\tilde{\nu}}(X)\right)\right] \geq \sup_{g: \mathcal{X} \rightarrow [0,1]} \text{kl}\left(\mathbb{E}_\nu[g(X)], \mathbb{E}_{\tilde{\nu}}[g(X)]\right).$

FROM KL TO REGRET LOWER BOUND

- From fundamental Weyl result on KL:

$$\sum_{a \in \mathcal{A}} \mathbb{E}_\theta[N_T(a)] \text{KL}(\theta_a, \theta'_a) \geq \text{kl}(\mathbb{P}_\theta[\Omega], \mathbb{P}_{\theta'}[\Omega])$$

Hence For all suboptimal arm $a \neq \star_\theta$,

$$\mathbb{E}_\theta[N_T(a)] \geq \sup_{\Omega, \theta'} \frac{\text{kl}(\mathbb{P}_\theta[\Omega], \mathbb{P}_{\tilde{\theta}}[\Omega]) - \sum_{a' \neq a} \text{KL}(\theta_{a'}, \theta'_{a'}) \mathbb{E}_\theta[N_T(a')]}{\text{KL}(\theta_a, \theta'_a)}.$$

Choose θ' such that a is optimal. Let $\Omega = \{N_T(a) > T^\alpha\}$.

- $\mathbb{P}_\theta[\Omega] \leq \mathbb{E}_\theta[N_T(a)] T^{-\alpha} = o(1)$ (**Consistency**)
- $\sum_{a' \in \mathcal{A}} N_T(a') = T$ (**Construction**)

Thus $\text{kl}(\mathbb{P}_\theta[\Omega], \mathbb{P}_{\tilde{\theta}}[\Omega]) \simeq \ln\left(\frac{1}{\mathbb{P}_{\tilde{\theta}}(N_T(a) \leq T^\alpha)}\right) \geq \ln\left(\frac{T - T^\alpha}{\sum_{a' \neq a} \mathbb{E}_{\tilde{\theta}}[N_T(a')]} \right) \simeq \ln(T)$.

- No constraint** on $\theta'_{a'}$ for $a' \neq a$: $\theta'_{a'} = \theta_{a'}$ kills the blue terms.

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}_\theta[N_T(a)]}{\ln(T)} \geq \frac{1 - 0}{\inf_{\tilde{\theta}_a} \{\text{KL}(\theta_a, \theta'_{a'}) : \mu'_a > \mu_{\star_\theta}\}}$$

Following the same proof as for the **fundamental Lemma** one can obtain the following generalization:

Lemma (\mathcal{D} -constrained regret lower bound)

Let \mathcal{D} be any set of bandit configurations and $\nu \in \mathcal{D}$. Then any uniformly-good strategy on \mathcal{D} must incur a regret

$$\liminf_{T \rightarrow \infty} \frac{\mathfrak{R}_{T,\nu}}{\ln(T)} \geq \inf \left\{ \sum_{a \in \mathcal{A}} c_a (\mu_*(\nu) - \mu_a(\nu)) : \right.$$

$$\left. \forall a \in \mathcal{A}, c_a \geq 0, \inf_{\nu' \in \tilde{\mathcal{D}}(\nu)} \sum_{a \in \mathcal{A}} c_a \text{KL}(\nu_a, \nu'_a) \geq 1 \right\}.$$

where we introduced the set of maximally confusing distributions

$$\tilde{\mathcal{D}}(\nu) = \left\{ \nu' \in \mathcal{D} : \mathcal{A}^*(\nu') \cap \mathcal{A}^*(\nu) = \emptyset, \forall a \in \mathcal{A}^*(\nu), \text{KL}(\nu_a, \nu'_a) = 0 \right\}.$$

- ▶ Solution to an **optimization** problem!
- ▶ Specialization to the multi-armed bandit setup of an even more general result from Graves&Lai, 97 (extending Agrawal 89).

Using similar steps as for unstructured lower bounds, we get

$$\forall a \notin \mathcal{A}^*(\nu), \forall \nu' \in \mathcal{D} \text{ s.t. } \mathcal{A}^*(\nu') = \{a\}$$

$$\liminf_T \frac{\sum_{a' \in \mathcal{A}} \mathbb{E}[N_T(a')] \text{KL}(\nu_{a'}, \nu'_{a'})}{\ln(T)} \geq \liminf_T \frac{\ln(T - T^\alpha)}{\ln(T)} - \frac{\ln\left(\sum_{a' \neq a} \mathbb{E}_{\nu'}[N_T(a')]\right)}{\ln(T)},$$

Using similar steps as for unstructured lower bounds, we get

$$\forall a \notin \mathcal{A}^*(\nu), \forall \nu' \in \mathcal{D} \text{ s.t. } \mathcal{A}^*(\nu') = \{a\}$$

$$\liminf_T \frac{\sum_{a' \in \mathcal{A}} \mathbb{E}[N_T(a')] \text{KL}(\nu_{a'}, \nu'_{a'})}{\ln(T)} \geq \liminf_T \frac{\ln(T - T^\alpha)}{\ln(T)} - \overbrace{\frac{\ln\left(\sum_{a' \neq a} \mathbb{E}_{\nu'}[N_T(a')]\right)}{\ln(T)}}^B,$$

By uniformly-good assumption, it must be that $B = 0$, hence

$$\liminf_T \sum_{a' \in \mathcal{A}} \frac{\mathbb{E}[N_T(a')]}{\ln(T)} \text{KL}(\nu_{a'}, \nu'_{a'}) = \sum_{a' \in \mathcal{A}} \left(\liminf_T \frac{\mathbb{E}[N_T(a')]}{\ln(T)} \right) \text{KL}(\nu_{a'}, \nu'_{a'}) \geq 1.$$

This holds in particular choosing ν' such that $\forall a' \in \mathcal{A}^*(\nu), \text{KL}(\nu_{a'}, \nu'_{a'}) = 0$. We conclude by remarking that

$$\liminf_{T \rightarrow \infty} \frac{\mathfrak{R}_T}{\ln(T)} = \sum_{a \in \mathcal{A}} \underbrace{\left(\liminf_{T \rightarrow \infty} \frac{\mathbb{E}[N_T(a)]}{\ln(T)} \right)}_{c_a} (\mu_\star(\nu) - \mu_a(\nu)).$$

What is the number of times a sub-optimal arm needs to be pulled?

The fundamental change of measure argument plus a simple reordering gives

$$\mathbb{E}_\nu[N_T(a)] \geq \sup_{\nu' \in \mathcal{D}} \frac{\sup_{\Omega} \text{KL}\left(\mathbb{P}_{\tilde{\nu}}[\Omega], \mathbb{P}_\nu[\Omega]\right) - \sum_{a' \in \mathcal{A} \setminus \{a\}} \mathbb{E}_\nu[N_T(a')] \text{KL}(\nu_{a'}, \nu'_{a'})}{\text{KL}(\nu_a, \nu'_a)}.$$

This motivates the following definition:

Definition (Asymptotic price for uniformly-good strategies)

For $\nu \in \mathcal{D}$, $a \notin \mathcal{A}_*(\nu)$, the asymptotic **price** to pay on arm a for **being uniformly-good** on \mathcal{D} is

$$n_T(a, \nu, \mathcal{D}) = \sup_{\nu' \in \mathcal{D}: a \in \mathcal{A}_*(\nu)} \frac{\ln(T) - \sum_{a' \in \mathcal{A} \setminus \{a\}} \mathbb{E}_\nu[N_T(a')] \text{KL}(\nu_{a'}, \nu'_{a'})}{\text{KL}(\nu_a, \nu'_a)}.$$

- ▷ **No structure** (**most confusing** obtained without changing other arms):

$$\begin{aligned}\mathbb{E}_\nu[N_T(a)] &\geq \sup_{\tilde{\nu} \in \mathcal{D}: \mathcal{A}_*(\tilde{\nu}) = \{a\}} \left\{ \frac{\ln(T)}{\text{KL}(\nu_a, \tilde{\nu}_a)} : \tilde{\nu} = (\nu_1, \dots, \tilde{\nu}_a, \dots, \nu_A) \right\} \\ &= \frac{\ln(T)}{\mathcal{K}_{\mathcal{D}}(\nu_a, \mu^*(\nu))}.\end{aligned}$$

- ▷ **No structure** (**most confusing** obtained without changing other arms):

$$\begin{aligned}\mathbb{E}_\nu[N_T(a)] &\geq \sup_{\tilde{\nu} \in \mathcal{D}: \mathcal{A}_*(\tilde{\nu})=\{a\}} \left\{ \frac{\ln(T)}{\text{KL}(\nu_a, \tilde{\nu}_a)} : \tilde{\nu} = (\nu_1, \dots, \tilde{\nu}_a, \dots, \nu_A) \right\} \\ &= \frac{\ln(T)}{\mathcal{K}_{\mathcal{D}}(\nu_a, \mu^*(\nu))}.\end{aligned}$$

- ▷ **Structure** (**most confusing** instance requires changing other arms):

$$\mathbb{E}_\nu[N_T(a)] \geq \sup_{\tilde{\nu} \in \mathcal{D}: \mathcal{A}_*(\tilde{\nu})=\{a\}} \left\{ \frac{\ln(T) - \sum_{a' \in \mathcal{A} \setminus \{a\}} \mathbb{E}_\nu[N_T(a')] \text{KL}(\nu_{a'}, \tilde{\nu}_{a'})}{\text{KL}(\nu_a, \tilde{\nu}_a)} \right\}.$$

How to adapt bandit strategy to handle such structure (ongoing research)?

- (**Collections**) $(\mathcal{A}, (\Theta_a)_{a \in \mathcal{A}}, (\mathcal{Y}_a)_{a \in \mathcal{A}}, (\nu_a)_{a \in \mathcal{A}}, (\mu_a)_{a \in \mathcal{A}})$
- (**Structure**) $\Theta \subset \prod_{a \in \mathcal{A}} \Theta_a$
- (**Parameter**) $\theta \in \Theta$

Finite set \mathcal{A} . For each $a \in \mathcal{A}$:

- (**Collections**) $(\mathcal{A}, (\Theta_a)_{a \in \mathcal{A}}, (\mathcal{Y}_a)_{a \in \mathcal{A}}, (\nu_a)_{a \in \mathcal{A}}, (\mu_a)_{a \in \mathcal{A}})$
- (**Structure**) $\Theta \subset \prod_{a \in \mathcal{A}} \Theta_a$
- (**Parameter**) $\theta \in \Theta$

Finite set \mathcal{A} . For each $a \in \mathcal{A}$:

- ▶ Parameter space Θ_a .

- (**Collections**) $(\mathcal{A}, (\Theta_a)_{a \in \mathcal{A}}, (\mathcal{Y}_a)_{a \in \mathcal{A}}, (\nu_a)_{a \in \mathcal{A}}, (\mu_a)_{a \in \mathcal{A}})$
- (**Structure**) $\Theta \subset \prod_{a \in \mathcal{A}} \Theta_a$
- (**Parameter**) $\theta \in \Theta$

Finite set \mathcal{A} . For each $a \in \mathcal{A}$:

- ▶ Parameter space Θ_a .
- ▶ Observation space \mathcal{Y}_a .

- (**Collections**) $(\mathcal{A}, (\Theta_a)_{a \in \mathcal{A}}, (\mathcal{Y}_a)_{a \in \mathcal{A}}, (\nu_a)_{a \in \mathcal{A}}, (\mu_a)_{a \in \mathcal{A}})$
- (**Structure**) $\Theta \subset \prod_{a \in \mathcal{A}} \Theta_a$
- (**Parameter**) $\theta \in \Theta$

Finite set \mathcal{A} . For each $a \in \mathcal{A}$:

- ▶ Parameter space Θ_a .
- ▶ Observation space \mathcal{Y}_a .
- ▶ Distribution of observations $\nu_a : \Theta_a \rightarrow \mathcal{P}(\mathcal{Y}_a)$

- (**Collections**) $(\mathcal{A}, (\Theta_a)_{a \in \mathcal{A}}, (\mathcal{Y}_a)_{a \in \mathcal{A}}, (\nu_a)_{a \in \mathcal{A}}, (\mu_a)_{a \in \mathcal{A}})$
- (**Structure**) $\Theta \subset \prod_{a \in \mathcal{A}} \Theta_a$
- (**Parameter**) $\theta \in \Theta$

Finite set \mathcal{A} . For each $a \in \mathcal{A}$:

- ▶ Parameter space Θ_a .
- ▶ Observation space \mathcal{Y}_a .
- ▶ Distribution of observations $\nu_a : \Theta_a \rightarrow \mathcal{P}(\mathcal{Y}_a)$
- ▶ Reward: $\mu_a : \Theta \rightarrow \mathbb{R}$ (**Θ and not Θ_a !**)

- ▶ **Classical Bernoulli MAB:** $\mathcal{A} = \{1, \dots, A\}$, $\Theta_a = [0, 1]$, $\mathcal{Y}_a = \{0, 1\}$, $\nu_a(\theta_a) = \text{Bern}(\theta_a)$, $\Theta = [0, 1]^{\mathcal{A}}$ (unstructured) and $\mu_a(\theta) = \theta_a$.

- ▶ **Classical Bernoulli MAB:** $\mathcal{A} = \{1, \dots, A\}$, $\Theta_a = [0, 1]$, $\mathcal{Y}_a = \{0, 1\}$, $\nu_a(\theta_a) = \text{Bern}(\theta_a)$, $\Theta = [0, 1]^{\mathcal{A}}$ (unstructured) and $\mu_a(\theta) = \theta_a$.
- ▶ **Linear bandits:** $\mathcal{A} \subset \mathbb{R}^d$, $\Theta_a = \{\langle \alpha, a \rangle : \alpha \in \mathbb{R}^d\}$, $\mathcal{Y}_a = \mathbb{R}$, $\nu_a(\theta_a) = \mathcal{N}(\theta_a, 1)$, $\Theta = \{\theta = (\langle \alpha, a \rangle)_{a \in \mathcal{A}}, \alpha \in \mathbb{R}^d\}$, $\mu_a(\theta) = \theta_a$.

- ▶ **Classical Bernoulli MAB:** $\mathcal{A} = \{1, \dots, A\}$, $\Theta_a = [0, 1]$, $\mathcal{Y}_a = \{0, 1\}$, $\nu_a(\theta_a) = \text{Bern}(\theta_a)$, $\Theta = [0, 1]^A$ (unstructured) and $\mu_a(\theta) = \theta_a$.
- ▶ **Linear bandits:** $\mathcal{A} \subset \mathbb{R}^d$, $\Theta_a = \{\langle \alpha, a \rangle : \alpha \in \mathbb{R}^d\}$, $\mathcal{Y}_a = \mathbb{R}$, $\nu_a(\theta_a) = \mathcal{N}(\theta_a, 1)$, $\Theta = \{\theta = (\langle \alpha, a \rangle)_{a \in \mathcal{A}}, \alpha \in \mathbb{R}^d\}$, $\mu_a(\theta) = \theta_a$.
- ▶ **Lipschitz bandits:** $\mathcal{A} \subset \mathcal{X}$, $\Theta_a \subset \mathbb{R}$, $\mathcal{Y}_a = \mathbb{R}$, $\nu_a(\theta_a) = \mathcal{N}(\theta_a, 1)$, $\Theta = \{\theta : \max_{a, a' \in \mathcal{X}} \frac{|\theta_a - \theta_{a'}|}{\ell(a, a')} \leq 1\}$, $\mu_a(\theta) = \theta_a$.

- ▶ **Classical Bernoulli MAB:** $\mathcal{A} = \{1, \dots, A\}$, $\Theta_a = [0, 1]$, $\mathcal{Y}_a = \{0, 1\}$, $\nu_a(\theta_a) = \text{Bern}(\theta_a)$, $\Theta = [0, 1]^A$ (unstructured) and $\mu_a(\theta) = \theta_a$.
- ▶ **Linear bandits:** $\mathcal{A} \subset \mathbb{R}^d$, $\Theta_a = \{\langle \alpha, a \rangle : \alpha \in \mathbb{R}^d\}$, $\mathcal{Y}_a = \mathbb{R}$, $\nu_a(\theta_a) = \mathcal{N}(\theta_a, 1)$, $\Theta = \{\theta = (\langle \alpha, a \rangle)_{a \in \mathcal{A}}, \alpha \in \mathbb{R}^d\}$, $\mu_a(\theta) = \theta_a$.
- ▶ **Lipschitz bandits:** $\mathcal{A} \subset \mathcal{X}$, $\Theta_a \subset \mathbb{R}$, $\mathcal{Y}_a = \mathbb{R}$, $\nu_a(\theta_a) = \mathcal{N}(\theta_a, 1)$, $\Theta = \{\theta : \max_{a, a' \in \mathcal{X}} \frac{|\theta_a - \theta_{a'}|}{\ell(a, a')} \leq 1\}$, $\mu_a(\theta) = \theta_a$.
- ▶ **Combinatorial semi-bandit:** $\mathcal{A} \subset \{0, 1\}^d$, $\Theta_a \subset \mathbb{R}^d$, $\mathcal{Y}_a = \mathbb{R}$, $\nu_a(\theta_a) = \mathcal{N}(\theta_a, I_d)$, $\Theta = \{\theta : \theta_a = (\alpha_1 a_1, \dots, \alpha_d a_d), \alpha \in \mathbb{R}^d\}$, $\mu_a(\theta) = \langle \theta_a, 1 \rangle$.

Theorem (Agrawal 1989)

Assume Θ is discrete, $\star(\theta) = \text{Argmax}_{a \in \mathcal{A}} \mu_a(\theta)$ is unique. Then for any uniformly good strategy,

$$\liminf_{T \rightarrow \infty} \frac{R_T(\theta)}{\ln(T)} \geq \mathbf{C}(\theta) \quad \text{where}$$

$$\mathbf{C}(\theta) = \min \left\{ \frac{\sum_{a \in \mathcal{A} \setminus \star(\theta)} \eta_a (\mu_{\star}(\theta) - \mu_a(\theta))}{\inf_{\lambda \in \Lambda(\theta)} \sum_{a \in \mathcal{A} \setminus \star(\theta)} \eta_a \text{KL}(\nu_a(\theta_a), \nu_a(\lambda_a))} : \eta \in \mathcal{P}(\mathcal{A} \setminus \star(\theta)) \right\}$$

$$\text{with } \Lambda(\theta) = \left\{ \lambda \in \Theta : \star(\theta) \neq \star(\lambda), \text{ and } \text{KL}(\nu_a(\theta_a), \nu_a(\lambda_a)) = 0 \text{ for } a = \star(\theta) \right\}.$$

Theorem (Agrawal 1989)

Assume Θ is discrete, $\star(\theta) = \text{Argmax}_{a \in \mathcal{A}} \mu_a(\theta)$ is unique. Then for any uniformly good strategy,

$$\liminf_{T \rightarrow \infty} \frac{R_T(\theta)}{\ln(T)} \geq C(\theta) \quad \text{where}$$

$$C(\theta) = \min \left\{ \frac{\sum_{a \in \mathcal{A} \setminus \star(\theta)} \eta_a (\mu_{\star}(\theta) - \mu_a(\theta))}{\inf_{\lambda \in \Lambda(\theta)} \sum_{a \in \mathcal{A} \setminus \star(\theta)} \eta_a \text{KL}(\nu_a(\theta_a), \nu_a(\lambda_a))} : \eta \in \mathcal{P}(\mathcal{A} \setminus \star(\theta)) \right\}$$

$$\text{with } \Lambda(\theta) = \left\{ \lambda \in \Theta : \star(\theta) \neq \star(\lambda), \text{ and } \text{KL}(\nu_a(\theta_a), \nu_a(\lambda_a)) = 0 \text{ for } a = \star(\theta) \right\}.$$

- ▶ Confusing parameters **statistically indistinguishable** from θ when playing only $\star(\theta)$.

Theorem (Graves, Lai 1997)

Assume $\star(\theta) = \text{Argmax}_{a \in \mathcal{A}} \mu_a(\theta)$ is unique. Then for any uniformly good strategy,

$$\liminf_{T \rightarrow \infty} \frac{R_T(\theta)}{\ln(T)} \geq C(\theta) \quad \text{where}$$

$$\begin{aligned} C(\theta) &= \min \left\{ \sum_{a \in \mathcal{A}} n_a (\mu_{\star}(\theta) - \mu_a(\theta)) : \forall a, n_a \geq 0 \right. \\ &\quad \left. \text{and } \inf_{\lambda \in \Lambda(\theta)} \sum_{a \in \mathcal{A}} n_a \text{KL}(\nu_a(\theta_a), \nu_a(\lambda_a)) \geq 1 \right\} \end{aligned}$$

with $\Lambda(\theta) = \left\{ \lambda \in \Theta : \star(\theta) \neq \star(\lambda), \text{ and } \text{KL}(\nu_a(\theta_a), \nu_a(\lambda_a)) = 0 \text{ for } a = \star(\theta) \right\}$.

Theorem (Graves, Lai 1997)

Assume $\star(\theta) = \text{Argmax}_{a \in \mathcal{A}} \mu_a(\theta)$ is unique. Then for any uniformly good strategy,

$$\liminf_{T \rightarrow \infty} \frac{R_T(\theta)}{\ln(T)} \geq C(\theta) \quad \text{where}$$

$$\begin{aligned} C(\theta) &= \min \left\{ \sum_{a \in \mathcal{A}} n_a (\mu_{\star}(\theta) - \mu_a(\theta)) : \forall a, n_a \geq 0 \right. \\ &\quad \left. \text{and } \inf_{\lambda \in \Lambda(\theta)} \sum_{a \in \mathcal{A}} n_a \text{KL}(\nu_a(\theta_a), \nu_a(\lambda_a)) \geq 1 \right\} \end{aligned}$$

with $\Lambda(\theta) = \left\{ \lambda \in \Theta : \star(\theta) \neq \star(\lambda), \text{ and } \text{KL}(\nu_a(\theta_a), \nu_a(\lambda_a)) = 0 \text{ for } a = \star(\theta) \right\}$.

- ▶ Confusing parameters **statistically indistinguishable** from θ when playing only $\star(\theta)$.

- ▷ Pick your favorite **structured bandit problem**
- ▷ Study the problem-dependent **lower bound** : Most confusing instance ?
- ▷ Each arm should be pulled some **minimum number** of times.
- ▷ Suggests an algorithm (sometimes optimal) !

Is OFUL strategy making **optimal use** of linear structure?

- ▷ We use it to build tight confidence sets.
- ▷ But not quite enough (Article "The end of optimism", Lattimore et al.)

Open question

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

OPTIMAL STRUCTURE EXPLOITATION

Structured lower bounds

Lipschitz bandits

Structure-exploiting strategies

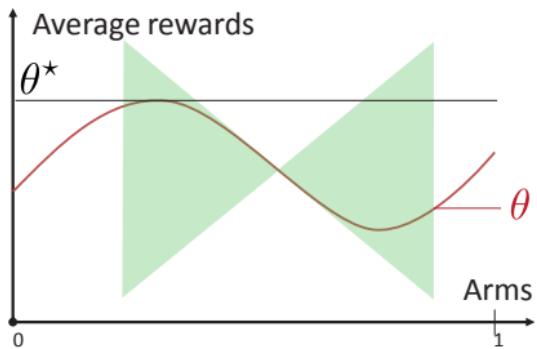
Metric-graph of bandits

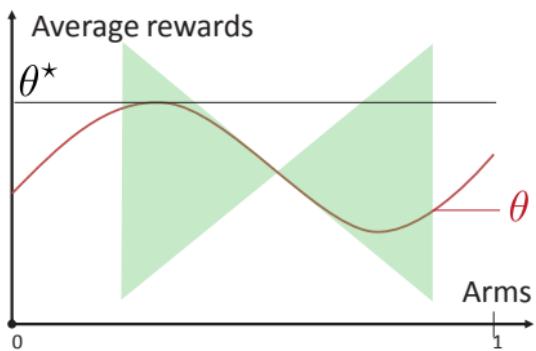
Equivariant bandits



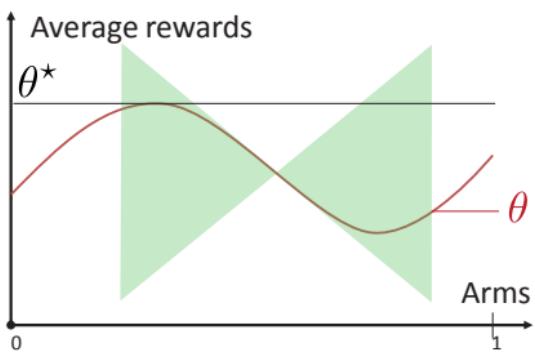
Lipschitz Bandits: Regret Lower Bounds and Optimal Algorithms

Stefan Magureanu, Richard Combes and Alexandre Proutiere, COLT 2014.

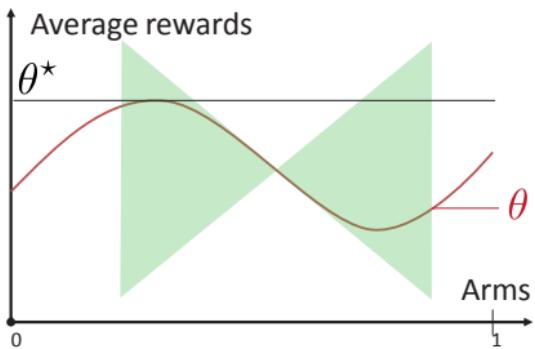




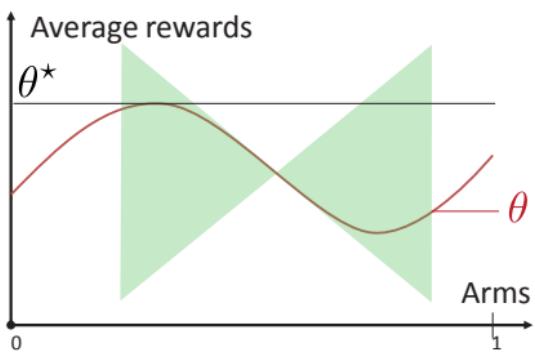
- ▶ The decision maker is given a **constant L**



- ▶ The decision maker is given a **constant L**
- ▶ Each $k \in \mathcal{K}$, is assigned a fixed and known coordinate $x_k \in (0, 1)$

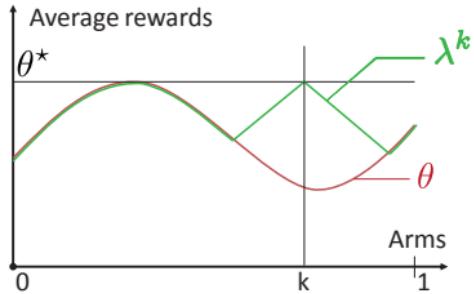


- ▶ The decision maker is given a **constant L**
- ▶ Each $k \in \mathcal{K}$, is assigned a fixed and known coordinate $x_k \in (0, 1)$
- ▶ Then : $\Theta_L = \{\theta \in (0, 1)^K : |\theta_i - \theta_j| \leq L|x_i - x_j|, \forall i, j \leq K\}$



- ▶ The decision maker is given a **constant L**
- ▶ Each $k \in \mathcal{K}$, is assigned a fixed and known coordinate $x_k \in (0, 1)$
- ▶ Then : $\Theta_L = \{\theta \in (0, 1)^K : |\theta_i - \theta_j| \leq L|x_i - x_j|, \forall i, j \leq K\}$
- ▶ Our goal is to exploit this additional information in order to reduce the achievable regret, relative to that of the classic setting.

LIPSCHITZ BANDITS - REGRET LOWER BOUNDS (PRELIMINARIES)



Let us define the most confusing *bad* parameter λ^k of an arm k :

$$\lambda_j^k = \max(\theta_j, \theta^* - L|x_j - x_k|), \forall j \in \mathcal{K}$$

Theorem (Lower bound)

For all $\theta \in \Theta_L$ and uniformly good algorithms π , we have:

$$\liminf_{T \rightarrow \infty} \frac{\mathcal{R}^\pi(T)}{\ln(T)} \geq C(\theta)$$

where $C(\theta)$ is the minimal value of the following optimization problem:

$$\min_{c_k > 0; k \in \mathcal{K}^-} \sum_{k \in \mathcal{K}^-} c_k (\theta^* - \theta_k) \text{ s.t. } \sum_{k' \in \mathcal{K}^-} c_{k'} \text{KL}(\theta_{k'}, \lambda_{\theta^*, k'}^k) \geq 1, \forall k \in \mathcal{K}^-$$

Theorem (Lower bound)

For all $\theta \in \Theta_L$ and uniformly good algorithms π , we have:

$$\liminf_{T \rightarrow \infty} \frac{\mathcal{R}^\pi(T)}{\ln(T)} \geq C(\theta)$$

where $C(\theta)$ is the minimal value of the following optimization problem:

$$\min_{c_k > 0; k \in \mathcal{K}^-} \sum_{k \in \mathcal{K}^-} c_k (\theta^* - \theta_k) \text{ s.t. } \sum_{k' \in \mathcal{K}^-} c_{k'} \text{KL}(\theta_{k'}, \lambda_{\theta^*, k'}^k) \geq 1, \forall k \in \mathcal{K}^-$$

- ▶ Follows result by Graves, Todd L., and Tze Leung Lai. "Asymptotically efficient adaptive choice of control laws in controlled markov chains." SIAM journal on control and optimization 35.3 (1997): 715-743

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

OPTIMAL STRUCTURE EXPLOITATION

Structured lower bounds

Lipschitz bandits

Structure-exploiting strategies

Metric-graph of bandits

Equivariant bandits



Indexed in Minimum of Empirical - Divergence

Junya Honda and Akimichi Takemura. "Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards"
Machine Learning, 16:3721–3756, 2015.

- ▷ Compute **Information** index:

$$I_a(t) = N_a(t) \mathcal{K}_{\mathcal{D}}(\hat{\nu}_a(t); \hat{\mu}^*(t)) + \ln(N_a(t))$$

where $\mathcal{K}_{\mathcal{D}}(\nu; \mu) = \inf_{\nu' \in \mathcal{D}} \{ \text{kl}(\nu, \nu') : \mu_{\nu'} > \nu \}$

- ▷ Play $\operatorname{argmin}_{a \in \mathcal{A}} I_a(t)$
- ▷ Directly **inspired from lower bound** (maximally confusing instance associated with empirical distributions)
- ▷ Uses plug-in empirical estimates up to **$\ln(N_a(t))$ bonus**.
- ▷ Provably **asymptotically optimal**

- ▷ $\Theta \subset [-1, 1]^{\mathcal{A}^2}$ is **known** to the learner.
- ▷ Structured bandit configurations:

$$\mathcal{D} = \left\{ \nu \in \mathcal{B} : \exists \theta \in \Theta, \forall a, a' \in \mathcal{A}, \quad \mu_a - \mu_{a'} \leq \theta_{a,a'} \right\}, \quad (3)$$

where \mathcal{B} is the set of Bernoulli distributions with means in $(0, 1)$.

- ▷ Assume pseudo-metric assumption on the $\theta_{a,a'}$ (definite, triangular ineq.)

Examples:

- ▷ **Lipschitz** bandits: $\theta_{a,a'} = k|a - a'|$
- ▷ **Unimodal** bandits: $\theta_{a,a'} = 1 - \mathbb{I}\{a < a' < *$ or $* < a' < a\}$
- ▷ **Linear** bandits: $\theta_{a,a'} = \langle \varphi_a - \varphi_{a'}, v \rangle$

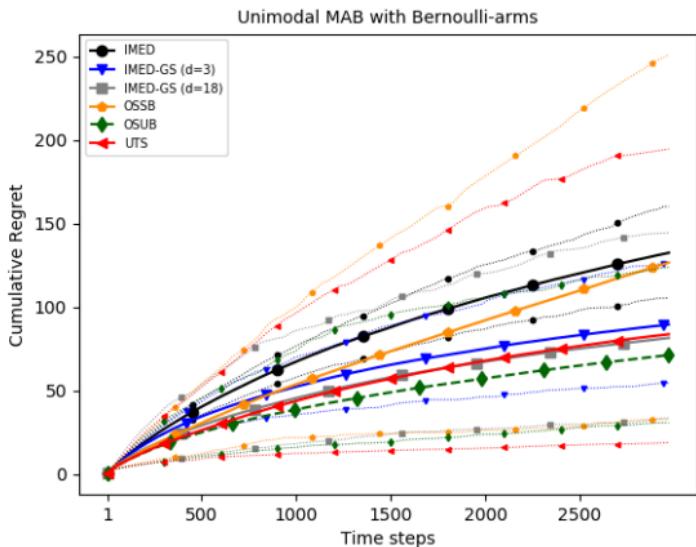
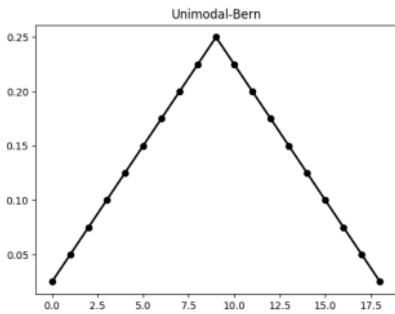
- ▷ Identify a set of **empirical best arm** $\widehat{\mathcal{A}}^*(t)$, with maximal empirical mean $\widehat{\mu}^*(t)$.
- ▷ Compute an **information** index

$$\mathcal{I}_a(t) = \begin{cases} \ln(N_a(t)) & \text{if } a \in \widehat{\mathcal{A}}^*(t) \\ \min_{\theta \in \Theta_a} \sum_{a' \in \widehat{\mathcal{A}}_a(\theta, t)} N_{a'}(t) \text{kl}(\widehat{\mu}_{a'}(t), \widehat{\mu}^*(t) - \theta_{a,a'}) + \ln(N_{\widehat{\mathcal{A}}_a(\theta, t)}(t)) & \text{otherwise} \end{cases}$$

where

- ▶ $\Theta_a = \{\theta : \theta_{a,*} > 0 \text{ and } \forall a' \neq *, \theta_{a,a'} \geq 0\}$
- ▶ $\widehat{\mathcal{A}}_a(\theta, t) = \{a' : \widehat{\mu}_{a'}(t) \leq \widehat{\mu}^*(t) - \theta_{a,a'}\}$
- ▷ Play $a_t = \underset{a \in \mathcal{A}}{\operatorname{argmin}} \mathcal{I}_a(t)$ (Actual strategy is slightly more elaborate)
- ▷ Provably achieves optimal regret (with correct constants).

► Results on **Unimodal** bandits



IMED-BASED STRATEGY: RESULTS

► Results on **Lipschitz** bandits

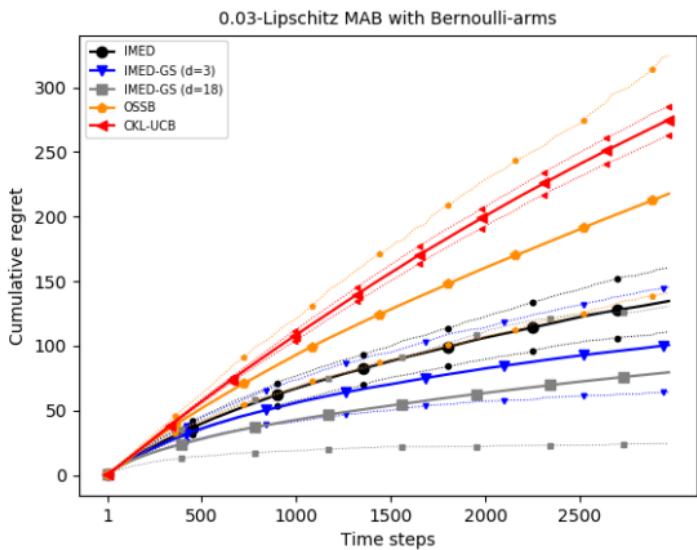
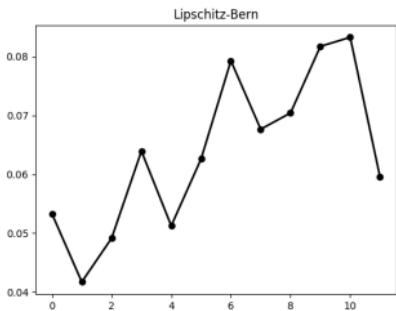


TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

OPTIMAL STRUCTURE EXPLOITATION

Structured lower bounds

Lipschitz bandits

Structure-exploiting strategies

Metric-graph of bandits

Equivariant bandits



- ▶ Bandit **configurations**: $\nu = (\nu_{a,b})_{a \in \mathcal{A}, b \in \mathcal{B}}$ with means $(\mu_{a,b})_{a \in \mathcal{A}, b \in \mathcal{B}}$
- ▶ \mathcal{A} : arms, \mathcal{B} : users.
- ▶ **Active contextual** bandit: At time t , learner chooses $b_t \in \mathcal{B}$, then $a_t \in \mathcal{A}$.
- ▶ **Regret**:

$$\mathcal{R}(\nu, T) = \mathbb{E}_\nu \left[\sum_{t=1}^T \max_{a \in \mathcal{A}} \mu_{a,b_t} - X_t \right] = \sum_{a,b \in \mathcal{C}_\nu^-} \Delta_{a,b} \mathbb{E}_\nu [N_{a,b}(T)].$$

where $\mathcal{C}_\nu^- = \left\{ (a, b) \in \mathcal{A} \times \mathcal{B} : \mu_{a,b} < \mu_b^\star \right\}$.

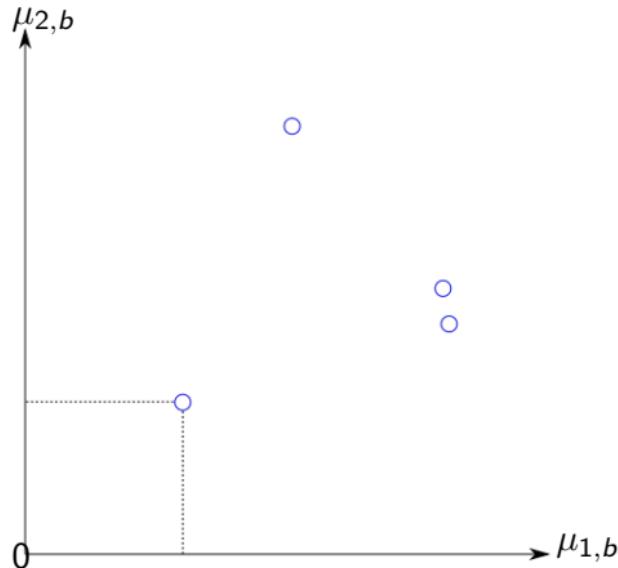
Definition(Uniformly spread strategy)

There exists $\gamma_1 > 0$ and a random variable Γ_2 with $\mathbb{E}_\nu[\Gamma_2] < 0$, such that

$$\forall b \in \mathcal{B}, \forall t \in \mathbb{N}, \quad N_b(t) \geq \gamma_1 \cdot t - \Gamma_2.$$

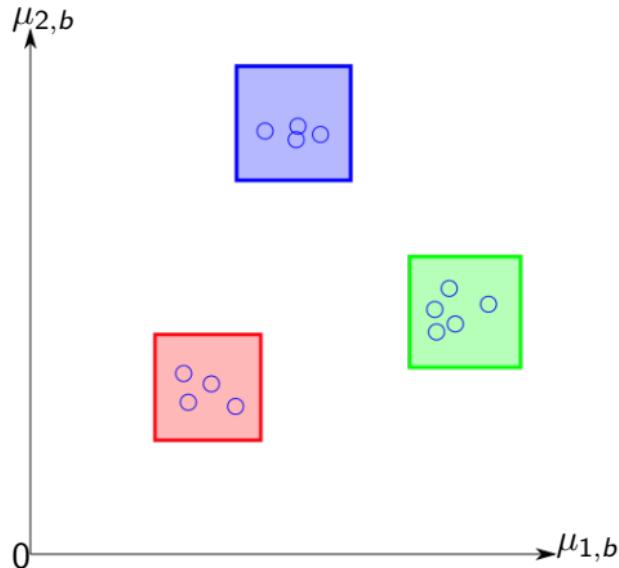
METRIC-GRAPH OF BANDITS

- ▶ Contextual bandits configuration means: $(\mu_{a,b})_{a \in \mathcal{A}, b \in \mathcal{B}}$
- ▶ Set of allowed 2-arm bandits ($\mathcal{A} = \{1, 2\}$):



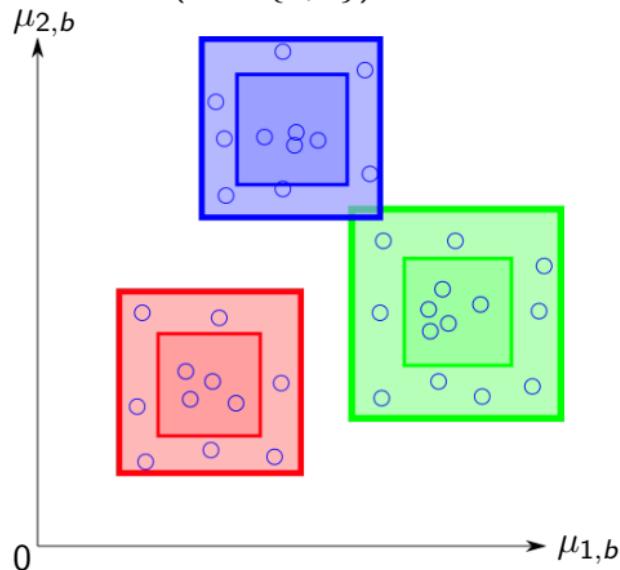
METRIC-GRAPH OF BANDITS

- ▶ Contextual bandits configuration means: $(\mu_{a,b})_{a \in \mathcal{A}, b \in \mathcal{B}}$
- ▶ Set of allowed 2-arm bandits ($\mathcal{A} = \{1, 2\}$):



METRIC-GRAPH OF BANDITS

- ▶ Contextual bandits configuration means: $(\mu_{a,b})_{a \in \mathcal{A}, b \in \mathcal{B}}$
- ▶ Set of allowed 2-arm bandits ($\mathcal{A} = \{1, 2\}$):



Bandit configurations ($\nu \in \mathcal{P}([0, 1])^{\mathcal{A} \times \mathcal{B}}$ with mean $\mu \in [0, 1]^{\mathcal{A} \times \mathcal{B}}$):

$$\mathcal{D}_\omega = \left\{ \nu : \forall b, b' \in \mathcal{B} \quad \max_{a \in \mathcal{A}} |\mu_{a,b} - \mu_{a,b'}| \leq \omega_{b,b'} \right\},$$

for a known weight matrix $\omega = (\omega_{b,b'})_{b,b' \in \mathcal{B}}$, symmetric, null-diagonal, with positive entries, and satisfying $\omega_{b,b'} \leq \omega_{b,b''} + \omega_{b'',b'}$.

Large values: not structured. Low value: highly structured.

Definition (Consistent strategy)

$$\forall \nu \in \mathcal{D}_\omega, \forall (a, b) \in \mathcal{C}_\nu^-, \forall \alpha \in (0, 1) \quad \lim_{T \rightarrow \infty} \mathbb{E}_\nu \left[\frac{N_{a,b}(T)^\alpha}{N_b(T)} \right] = 0.$$

Proposition (Regret lower bound)

Any uniformly spread and consistent strategy must satisfy

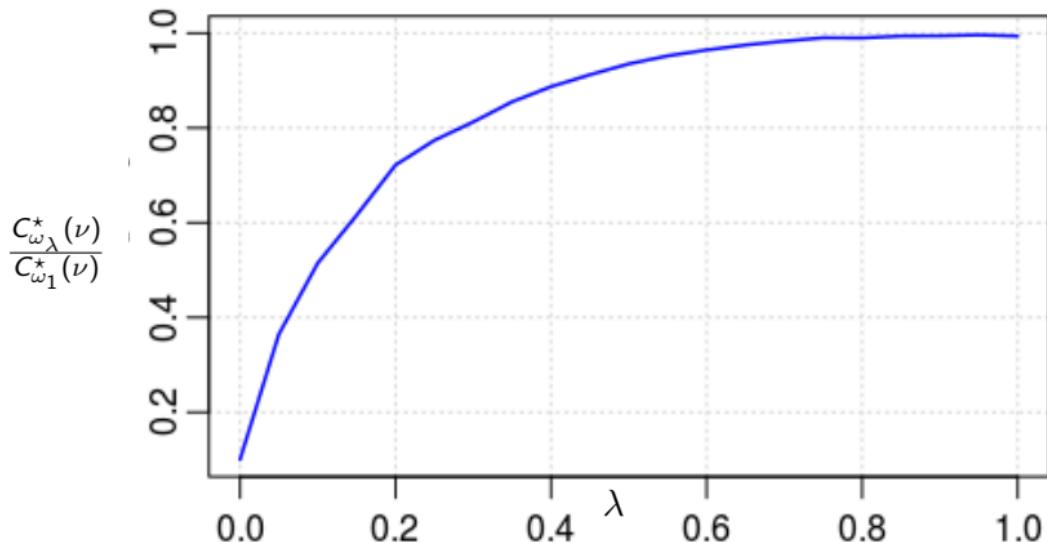
$$\liminf_{T \rightarrow \infty} \frac{\mathcal{R}(\nu, T)}{\ln(T)} \geq C_\omega^*(\nu)$$

where $C_\omega^*(\nu) = \min_{n \in \mathbb{R}_+^{|\mathcal{C}^-|}} \sum_{a, b \in \mathcal{C}^-} n_{a,b} \Delta_{a,b}$ s.t.

$$\forall (a, b) \in \mathcal{C}^-, \sum_{b' \in \mathcal{B}: (a, b') \in \mathcal{C}^-} \text{kl}^+(\mu_{a,b'} | \mu_b^* - \omega_{b,b'}) n_{a,b'} \geq 1.$$

SPECIAL CASES

- ▶ Let ω_λ be a matrix where all the weights are equal to $\lambda \in [0, 1]$ except for the zero diagonal.
- ▶ $\lambda = 1$: **no-structure**, $\lambda = 0$: one unique cluster.
- ▶ We recover that $C_{\omega_1}^*(\nu) = \sum_{a,b \in \mathcal{C}^-} \frac{\Delta_{a,b}}{\text{k1}(\mu_{a,b} | \mu_b^*)}$ (unstructured lower bound)
- ▶ More generally:



- ▶ Explicit lower bound spanning unstructured to highly structured pbs.
- ▶ See (Saber et al., submitted) for an algorithm:
 - ▶ Provably asymptotically optimal.
 - ▶ Computationally cheap
 - ▶ Without explicit forced exploration (still some implicit forcing).

TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

OPTIMAL STRUCTURE EXPLOITATION

Structured lower bounds

Lipschitz bandits

Structure-exploiting strategies

Metric-graph of bandits

Equivariant bandits

- **Structure \mathcal{B}_q :** For each $a \in \mathcal{A}$, there are **at least q** arms with same mean as a .

$$\forall a \in \mathcal{A}, \quad |\{a' \in \mathcal{A} : \mu_{a'} = \mu_a\}| \geq q$$

- This defines **equivalence classes** with minimal size q .
- **Lower bound** makes appear **combinatorial** terms. For any sub-optimal class $c \subset \mathcal{A}$:

$$\liminf_{T \rightarrow \infty} \frac{\min_{c_q \subseteq c} \sum_{a \in c_q} \mathbb{E}_{\nu}(N_a(T)) \mathcal{K}_{\text{inf}}(\nu_a \| \mu_*) + \inf_{\mu' \in \mathcal{B}_q(c_q, \mu_*)} \sum_{a \notin c_q} \mathbb{E}_{\nu}(N_a(T)) \mathcal{K}_{\text{eq}}(\nu_a \| \mu'_a)}{\log T} \geq 1, \quad (4)$$

where

- c_q is any subset of c having q distincts arms within it.
- $\mathcal{B}_q(c_q, \mu_*)$: distributions in \mathcal{B}_q s.t. arms in c_q have mean $\geq \mu_*$

IMED FOR EQUIVALENCE CLASSES

- ▷ First define unstructured IMED index for each $a \in \mathcal{A}$:

$$I_a(t) = N_a(t) \mathcal{K}_{inf}(\hat{\mu}_a(t) \| \hat{\mu}^*(t)) + \log N_a(t),$$

- ▷ **Key observation:** no need to solve a combinatorial problem !

$$I(t) = \min_{\substack{\mathcal{A}' \subset \mathcal{A} \\ |\mathcal{A}'| = q}} \sum_{a' \in \mathcal{A}'} I_{a'}(t) = \sum_{a' \in \mathcal{A}_q(t)} I_{a'}(t).$$

Pull each arm once

for $t = |\mathcal{A}| \dots T - 1$ **do**

if $\min_{a \in \mathcal{A}_*(t)} I_a(t) \leq I(t)$ **then**

 Pull $a_{t+1} \in \arg \min_{a \in \mathcal{A}_*(t)} N_a(t)$ (chosen arbitrarily)

else

 Pull $a_{t+1} \in \arg \min_{a \notin \mathcal{A}_*(t)} I_a(t)$ (chosen arbitrarily)

end if

end for

Theorem (Upper bound on the number of pulls)

Under the IMED-EC algorithms for all suboptimal arm a it holds

$$\mathbb{E}_\nu[N_a(T)] \leq \frac{\log T}{\mathbf{q}\mathcal{K}_{inf}(\nu_a\|\mu_*)}(1 + \alpha(\varepsilon)) + f(\varepsilon), \quad (5)$$

where $0 < \varepsilon < \frac{1}{3} \min_{a \in \mathcal{A} \setminus \mathcal{A}_*} (\mu_* - \mu_a)$, f is function that depends on concentration properties on \mathcal{F} , and α tends to 0 as ε tends to 0.

- ▷ We can show this is **optimal**, up to a constant factor never larger than 2.

EMPIRICAL REGRET

- ▷ Results with 7 classes each of size $q = 8$:

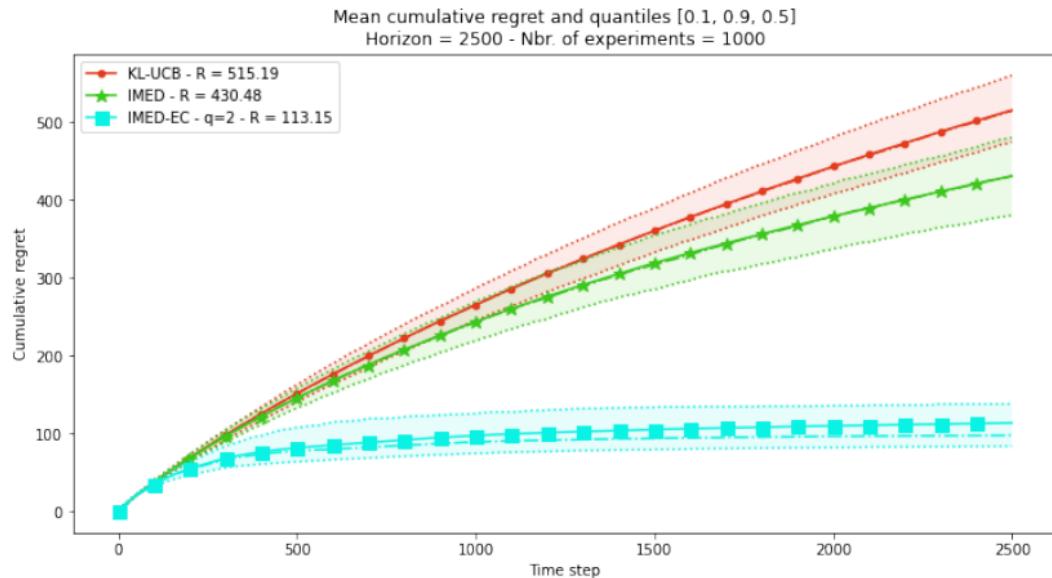


TABLE OF CONTENTS

WHY BANDITS?

VANILLA STOCHASTIC BANDITS

OPTIMAL BANDIT STRATEGIES

EXPLOITING STRUCTURE

OPTIMAL STRUCTURE EXPLOITATION

CONCLUSION, PERSPECTIVE

To minimize **regret**, **DO NOT** play

$$\hat{\star}_t = \operatorname{argmax}_{a \in \mathcal{A}} \hat{\mu}_a(t) \quad \text{where } \hat{\mu}_a(t) = \frac{1}{N_t(a)} \sum_{t=1}^T Y_t \mathbb{I}\{A_t = a\}$$

(yet all MC-based methods Q-learning, (LS)TD, DQN, etc. do so)

Four type of perturbed strategies instead

- ▷ UCB: $\operatorname{argmax}_{a \in \mathcal{A}} \hat{\mu}_a(t) + B_a(t)$ where **Optimistic** $= \hat{\mu}_a(t)$ with high probability.
- ▷ TS/DS: $\operatorname{argmax}_{a \in \mathcal{A}} \tilde{\mu}_a(t)$ where $\tilde{\mu}_a \sim \text{Beta}(\hat{\mu}_a(t), N_t(a))$ /Randomly reweighted **mean**.
- ▷ IMED: $\operatorname{argmin}_{a \in \mathcal{A}} N_t(a) \mathbf{D}(\hat{\mu}_a(t), \max_a \hat{\mu}_a(t)) + \ln(N_t(a))$ with divergence D .
- ▷ SDA: All $\{a : \mu_a^\dagger(t) \geq \max_a \hat{\mu}_a(t)\}$ where $\mu_a^\dagger(t)$ mean of $N_t(\hat{\star}_t)$ -many **randomly chosen** observations from a . **Sub-sampling**

- ▷ Pick your favorite **structured bandit problem**
- ▷ Study the problem-dependent **lower bound** : **Most confusing instance** ?
- ▷ Each arm should be pulled some **minimum number** of times.
- ▷ Suggests an algorithm (often optimal) !

- ▷ **No structure** (**most confusing** obtained without changing other arms):

$$\begin{aligned}\mathbb{E}_\nu[N_T(a)] &\geq \sup_{\tilde{\nu} \in \mathcal{D}: \mathcal{A}_*(\tilde{\nu}) = \{a\}} \left\{ \frac{\ln(T)}{\text{KL}(\nu_a, \tilde{\nu}_a)} : \tilde{\nu} = (\nu_1, \dots, \tilde{\nu}_a, \dots, \nu_A) \right\} \\ &= \frac{\ln(T)}{\mathcal{K}_{\mathcal{D}}(\nu_a, \mu^*(\nu))}.\end{aligned}$$

- ▷ **No structure** (**most confusing** obtained without changing other arms):

$$\begin{aligned}\mathbb{E}_\nu[N_T(a)] &\geq \sup_{\tilde{\nu} \in \mathcal{D}: \mathcal{A}_*(\tilde{\nu})=\{a\}} \left\{ \frac{\ln(T)}{\text{KL}(\nu_a, \tilde{\nu}_a)} : \tilde{\nu} = (\nu_1, \dots, \tilde{\nu}_a, \dots, \nu_A) \right\} \\ &= \frac{\ln(T)}{\mathcal{K}_{\mathcal{D}}(\nu_a, \mu^*(\nu))}.\end{aligned}$$

- ▷ **Structure** (**most confusing** instance requires changing other arms):

$$\mathbb{E}_\nu[N_T(a)] \geq \sup_{\tilde{\nu} \in \mathcal{D}: \mathcal{A}_*(\tilde{\nu})=\{a\}} \left\{ \frac{\ln(T) - \sum_{a' \in \mathcal{A} \setminus \{a\}} \mathbb{E}_\nu[N_T(a')] \text{KL}(\nu_{a'}, \tilde{\nu}_{a'})}{\text{KL}(\nu_a, \tilde{\nu}_a)} \right\}.$$

How to adapt bandit strategy to handle such structure (ongoing research)?

- ▷ **Generic** structure:
 - ▶ Generic algorithm ? IMED-structure is very promising.
 - ▶ Structure identification (instead of being given) ?
- ▷ **Non-parametric** and structured : IMED + SDA?
- ▷ From bandits to **contextual** bandits, to **MDPs**:
Q-learning, DQN, PPO, etc + IMED/SDA/DS?
- ▷ Beyond structure? **No** stochastic model?

“The more applied you go, the stronger theory you need”

MERCI

odalricambrym.maillard@inria.fr

odalricambrymmaillard.wordpress.com