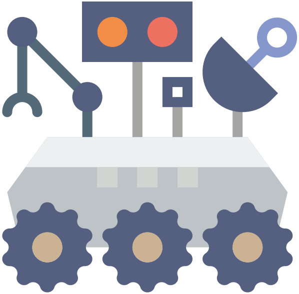
Exploration and Exploitation in Structured Stochastic Bandits



Wouter M. Koolen

CWI

Centrum Wiskunde & Informatica

Collaborators

Rémy Degenne

Han Shao (邵涵)

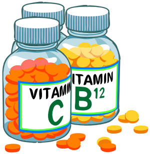
Emilie Kaufmann

Pierre Ménard

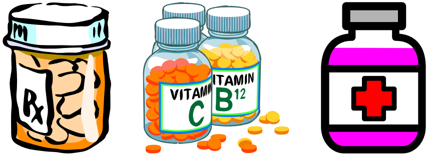
Aurélien Garivier

Outline

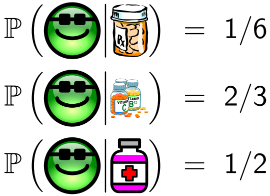
Stochastic Bandit



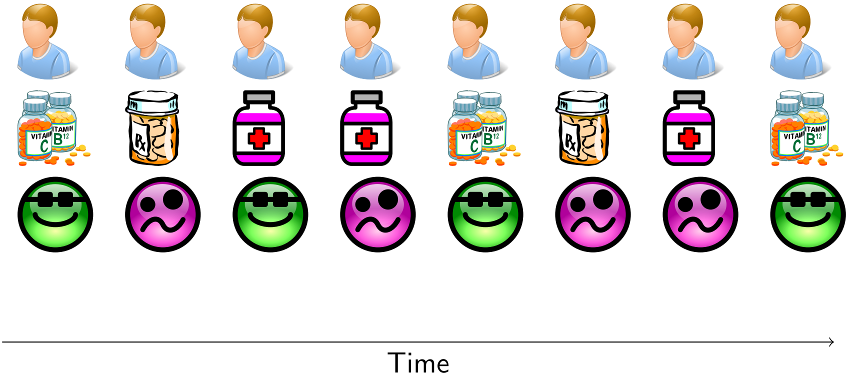
Stochastic Bandit



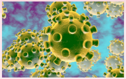
Model (Unknown)



Stochastic Bandit Interaction

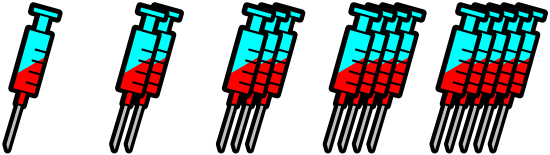


Tasks

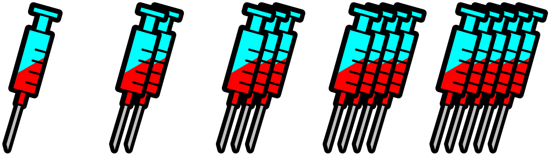


1. Best Arm Identification: use trial to cure **population**
2. Reward Maximisation: cure **patients in trial**

Structured Stochastic Bandit

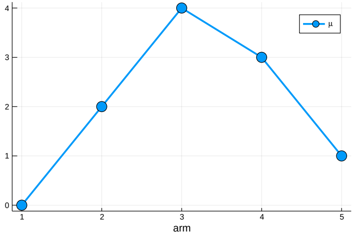


Structured Stochastic Bandit

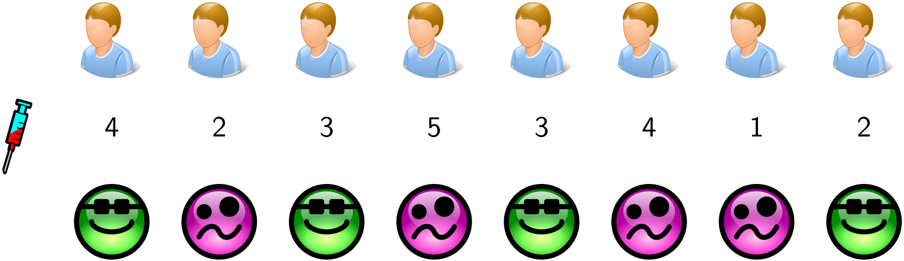


Model (Unknown)

$$
\begin{aligned}
\mathbb{P}\left(\;\middle|\; \times 1\right) &= 1/6\\
\mathbb{P}\left(\;\middle|\; \times 2\right) &= 3/6\\
\mathbb{P}\left(\;\middle|\; \times 3\right) &= 5/6\\
\mathbb{P}\left(\;\middle|\; \times 4\right) &= 4/6\\
\mathbb{P}\left(\;\middle|\; \times 5\right) &= 2/6
\end{aligned}
$$



Structured Stochastic Bandit Interaction



This Talk

We will develop **efficient** **structure-adaptive** learning algorithms for **Best Arm Identification** and **Reward Maximisation**.

Information-theoretic lower bounds will tell us that the complexity of each task is characterised by a certain **two-player zero-sum game**.

We will base our learning algorithms on iterative **saddle point solvers** for this game.

Why are we doing this?

Structure interesting in practise

- Unimodal [Combes and Proutiere, 2014]
- Lipschitz [Magureanu, Combes, and Proutière, 2014]
- Rank-1 [Katariya, Kveton, Szepesvári, Vernade, and Wen, 2017]
- Linear [Lattimore and Szepesvári, 2017]
- Sparse [Kwon, Perchet, and Vernade, 2017]
- Categorised [Jedor, Perchet, and Louedec, 2019]
- Combinatorial, duelling, ...

Why are we doing this?

Structure interesting in practise

- Unimodal [Combes and Proutiere, 2014]
- Lipschitz [Magureanu, Combes, and Proutière, 2014]
- Rank-1 [Katariya, Kveton, Szepesvári, Vernade, and Wen, 2017]
- Linear [Lattimore and Szepesvári, 2017]
- Sparse [Kwon, Perchet, and Vernade, 2017]
- Categorised [Jedor, Perchet, and Louedec, 2019]
- Combinatorial, duelling, ...

Sub-modules (training ground) for

- reinforcement learning
- simulator-based planning
- environments with selfish or adversarial agents

Outline

Environments

We fix an 1-d exponential family (Bernoulli, Gaussian, ...) parameterised by the **mean**. KL divergence denoted by $d(\mu, \lambda)$.

Multi-armed bandit model

A K-**armed bandit model** is a tuple $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$.

Environments

We fix an 1-d exponential family (Bernoulli, Gaussian, ...) parameterised by the **mean**. KL divergence denoted by $d(\mu, \lambda)$.

Multi-armed bandit model

A K-**armed bandit model** is a tuple $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$.

Learning Target

The best arm for $\boldsymbol{\mu}$ is

$$i^*(\boldsymbol{\mu}) \ := \ \operatorname*{argmax}_i \mu_i$$

Environments

We fix an 1-d exponential family (Bernoulli, Gaussian, ...) parameterised by the **mean**. KL divergence denoted by $d(\mu, \lambda)$.

Multi-armed bandit model

A K-**armed bandit model** is a tuple $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$.

Learning Target

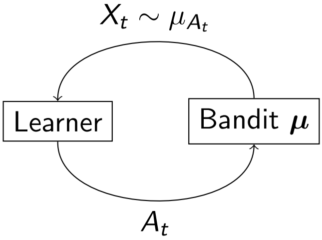
The best arm for $\boldsymbol{\mu}$ is

$$i^*(\boldsymbol{\mu}) \; := \; \operatorname*{argmax}_i \mu_i$$

Structure

Set of possible bandit models $\mathcal{M} \subseteq \mathbb{R}^K$.

Interaction



Best Arm Identification: Strategy for Learner

Strategy

- **Stopping rule** $\tau \in \mathbb{N}$
- In round $t \leq \tau$ **sampling rule** picks $A_t \in [K]$. **See** $X_t \sim \mu_{A_t}$**.**
- **Recommendation rule** $\hat{I} \in [K]$.

Best Arm Identification: Strategy for Learner

Strategy

- **Stopping rule** $\tau \in \mathbb{N}$
- In round $t \leq \tau$ **sampling rule** picks $A_t \in [K]$. **See** $X_t \sim \mu_{A_t}$**.**
- **Recommendation rule** $\hat{I} \in [K]$.

Realisation of interaction: $\mathcal{H} := \left(A_1, X_1, \ldots, A_\tau, X_\tau, \hat{I}\right)$.

Best Arm Identification: Strategy for Learner

Strategy

- **Stopping rule** $\tau \in \mathbb{N}$
- In round $t \leq \tau$ **sampling rule** picks $A_t \in [K]$. **See** $X_t \sim \mu_{A_t}$**.**
- **Recommendation rule** $\hat{I} \in [K]$.

Realisation of interaction: $\mathcal{H} := \left(A_1, X_1, \ldots, A_\tau, X_\tau, \hat{I}\right)$.

Two objectives: **sample efficiency** τ and **correctness** $\hat{I} = i^*(\boldsymbol{\mu})$.

Best Arm Identification Goal: PAC learning

Definition

Fix small confidence $\delta \in (0,1)$. A strategy is δ-**correct** if

$$\mathbb{P}_{\boldsymbol{\mu}}\big(\hat{I} \neq i^*(\boldsymbol{\mu})\big) \;\leq\; \delta \qquad \text{for every bandit model } \boldsymbol{\mu} \in \mathcal{M}.$$

Best Arm Identification Goal: PAC learning

Definition

Fix small confidence $\delta \in (0,1)$. A strategy is δ-**correct** if

$$\mathbb{P}_{\boldsymbol{\mu}}\big(\hat{I} \neq i^*(\boldsymbol{\mu})\big) \;\leq\; \delta \qquad \text{for every bandit model } \boldsymbol{\mu} \in \mathcal{M}.$$

Goal: minimise sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$ over **all δ-correct strategies**.

Best Arm Identification Goal: PAC learning

Definition

Fix small confidence $\delta \in (0, 1)$. A strategy is δ-**correct** if

$$\mathbb{P}_{\boldsymbol{\mu}}\big(\hat{I} \neq i^*(\boldsymbol{\mu})\big) \;\leq\; \delta \qquad \text{for every bandit model } \boldsymbol{\mu} \in \mathcal{M}.$$

Goal: minimise sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$ over **all δ-correct strategies**.

Hope

Efficient δ-correct algorithm with instance-optimal sample complexity

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \;\preceq\; \square_{\boldsymbol{\mu}} \ln \tfrac{1}{\delta} \qquad \text{for all } \boldsymbol{\mu} \in \mathcal{M}.$$

Regret Minimisation: Strategy and Goal

In round $t \leq T$ **sampling rule** picks $A_t \in [K]$, and sees $X_t \sim \mu_{A_t}$.

Regret Minimisation: Strategy and Goal

In round $t \leq T$ **sampling rule** picks $A_t \in [K]$, and sees $X_t \sim \mu_{A_t}$.

Realisation of interaction: $\mathcal{H} := (A_1, X_1, \dots, A_T, X_T)$.

Definition

The objective is

$$\mathrm{R}_T(\boldsymbol{\mu}) \;:=\; \sum_{k=1}^{K} \mathbb{E}[N_T^k] \Delta^k$$

where the sub-optimality gaps are given by $\Delta^k = \mu^* - \mu^k$.

Regret Minimisation: Strategy and Goal

In round $t \leq T$ **sampling rule** picks $A_t \in [K]$, and sees $X_t \sim \mu_{A_t}$.

Realisation of interaction: $\mathcal{H} \coloneqq (A_1, X_1, \ldots, A_T, X_T)$.

Definition

The objective is

$$\mathrm{R}_T(\boldsymbol{\mu}) \;\coloneqq\; \sum_{k=1}^{K} \mathbb{E}[N_T^k] \Delta^k$$

where the sub-optimality gaps are given by $\Delta^k = \mu^* - \mu^k$.

Hope

Efficient algorithm with instance-optimal regret

$$\mathrm{R}_T(\boldsymbol{\mu}) \;\preceq\; \square_{\boldsymbol{\mu}} \ln T \qquad \text{for all } \boldsymbol{\mu} \in \mathcal{M}.$$

Outline

Instance-Dependent Sample Complexity Lower Bound

Intuition (going back at least to Lai and Robbins [1985]): if observations are likely under both $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$, yet $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$, then learner cannot stop and be correct in both.

Instance-Dependent Sample Complexity Lower Bound

Intuition (going back at least to Lai and Robbins [1985]): if observations are likely under both $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$, yet $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$, then learner cannot stop and be correct in both.

Define the **alternative** to $\boldsymbol{\mu}$ by $\mathrm{Alt}(\boldsymbol{\mu}) \coloneqq \{\boldsymbol{\lambda} \in \mathcal{M} | i^*(\boldsymbol{\lambda}) \neq i^*(\boldsymbol{\mu})\}$.

Instance-Dependent Sample Complexity Lower Bound

Intuition (going back at least to Lai and Robbins [1985]): if observations are likely under both $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$, yet $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$, then learner cannot stop and be correct in both.

Define the **alternative** to $\boldsymbol{\mu}$ by $\text{Alt}(\boldsymbol{\mu}) \coloneqq \{\boldsymbol{\lambda} \in \mathcal{M} | i^*(\boldsymbol{\lambda}) \neq i^*(\boldsymbol{\mu})\}$.

Theorem (Castro 2014, Garivier and Kaufmann 2016)

Fix a δ-correct strategy. Then for every bandit model $\boldsymbol{\mu} \in \mathcal{M}$

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \;\geq\; T^*(\boldsymbol{\mu}) \ln \frac{1}{\delta}$$

where the characteristic time $T^*(\boldsymbol{\mu})$ *is given by*

$$\frac{1}{T^*(\boldsymbol{\mu})} \;=\; \boxed{\max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \text{Alt}(\boldsymbol{\mu})} \sum_{i=1}^K w_i d(\mu_i, \lambda_i)}$$

$w^k \propto N^k$ pulls

Example

$K = 5$ Bernoulli arms, $\boldsymbol{\mu} = (0.4, 0.3, 0.2, 0.1, 0.0)$.

$$T^*(\boldsymbol{\mu}) \;=\; 200.4 \qquad \boldsymbol{w}^*(\boldsymbol{\mu}) \;=\; (0.45, 0.46, 0.06, 0.02, 0.01)$$

At confidence $\delta = 0.05$ we have $\ln \frac{1}{\delta} = 3.0$ and hence $\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq 601.2$.

Instance-Dependent Regret Lower Bound

Theorem (Graves and Lai 1997)

Any asymptotically consistent algorithm for structure $\mathcal{M}$ must incur on each $\boldsymbol{\mu} \in \mathcal{M}$ regret at least

$$\mathrm{R}_T(\boldsymbol{\mu}) \succeq V(\boldsymbol{\mu}) \ln T$$

where the characteristic regret rate *is given by*

$$\frac{1}{V(\boldsymbol{\mu})} = \boxed{\max_{\tilde{\boldsymbol{w}} \in \triangle} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_k \tilde{w}^k \frac{d(\mu^k, \lambda^k)}{\Delta^k}}$$

$\tilde{w}^k \propto N^k \Delta^k$
regret

Outline

Lower Bounds Inspire Strategies

Recall sample complexity/regret lower bound governed by

$$\boxed{\max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \text{Alt}(\boldsymbol{\mu})} \sum_{i=1}^K w_i d(\mu_i, \lambda_i)} \quad \text{or} \quad \boxed{\max_{\tilde{\boldsymbol{w}} \in \triangle} \inf_{\boldsymbol{\lambda} \in \text{Alt}(\boldsymbol{\mu})} \sum_k \tilde{w}^k \frac{d(\mu^k, \lambda^k)}{\Delta^k}}$$

Lower Bounds Inspire Strategies

Recall sample complexity/regret lower bound governed by

$$\boxed{\max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \text{Alt}(\boldsymbol{\mu})} \sum_{i=1}^K w_i d(\mu_i, \lambda_i)} \quad \text{or} \quad \boxed{\max_{\tilde{\boldsymbol{w}} \in \triangle} \inf_{\boldsymbol{\lambda} \in \text{Alt}(\boldsymbol{\mu})} \sum_k \tilde{w}^k \frac{d(\mu^k, \lambda^k)}{\Delta^k}}$$

Matching algorithms **must** sample with **argmax** (**oracle**) proportions.

Lower Bounds Inspire Strategies

Earlier work [Combes et al., 2017, Garivier and Kaufmann, 2016]
At each time step

- compute plug-in **oracle solution** $\boldsymbol{w}^*(\hat{\boldsymbol{\mu}}_t)$ or $\tilde{\boldsymbol{w}}^*(\hat{\boldsymbol{\mu}}_t)$.
- sample arm A_t to track that solution
- **force exploration** to ensure $\hat{\boldsymbol{\mu}}_t \to \boldsymbol{\mu}$.

Lower Bounds Inspire Strategies

Earlier work [Combes et al., 2017, Garivier and Kaufmann, 2016]
At each time step

- compute plug-in **oracle solution** $\boldsymbol{w}^*(\hat{\boldsymbol{\mu}}_t)$ or $\tilde{\boldsymbol{w}}^*(\hat{\boldsymbol{\mu}}_t)$.
- sample arm A_t to track that solution
- **force exploration** to ensure $\hat{\boldsymbol{\mu}}_t \to \boldsymbol{\mu}$.

Coming up

- **Iteratively** solve lower bounds by full information online learning.
- Use iterates to drive sampling rule.
- Add optimism to induce exploration.
- Cap gap estimates $\hat{\boldsymbol{\Delta}}_t$ from below to reduce estimation variance
- **Compose** regret bound from saddle-point regret + estimation regret

Outline

Interleaved Iterative Solution

Standard technique: can approximately solve saddle point problems like

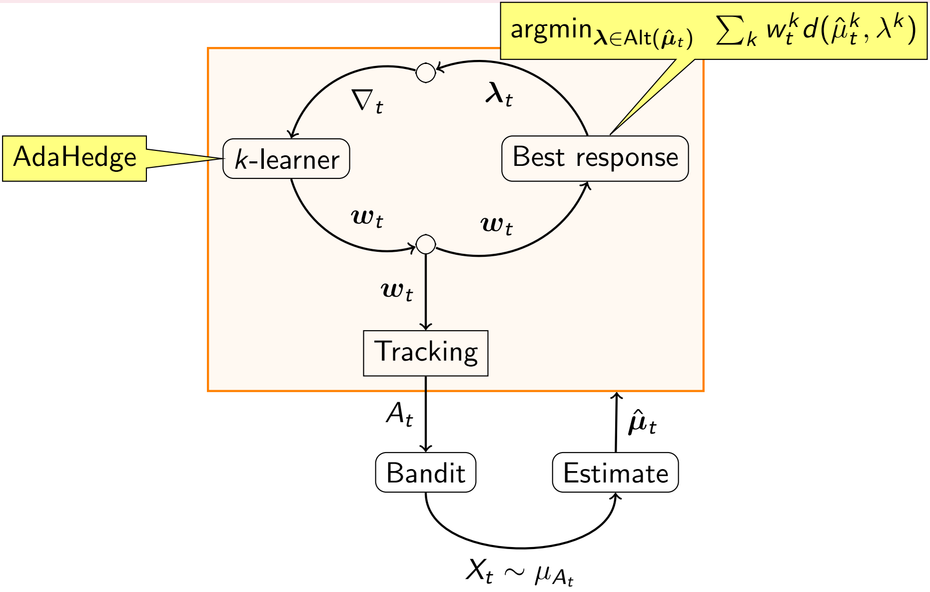
$$\boxed{\max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{i=1}^K w_i d(\mu_i, \lambda_i)} \quad \text{or} \quad \boxed{\max_{\tilde{\boldsymbol{w}} \in \triangle} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_k \tilde{w}^k \frac{d(\mu^k, \lambda^k)}{\Delta^k}}$$

iteratively using two online learners.

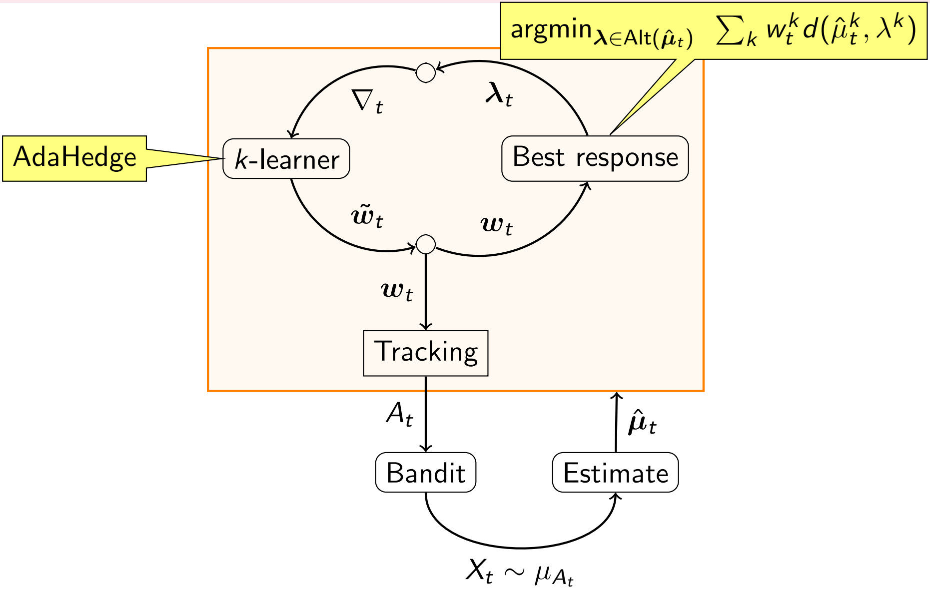
Main pipeline [Degenne, Koolen, and Ménard, 2019]:

- Plug-in estimate $\hat{\boldsymbol{\mu}}_t$ (so problem is **shifting**).
- Advance the saddle point solver by **one** iteration for every bandit interaction.
- Add optimism to gradients to induce exploration

Sampling Rule for Best Arm Identification



Sampling Rule for Regret Minimisation



Compositionality

The "overheads" of the ingredients **compose**: Tracking $O(1)$, concentration $\sqrt{T}$, regret $\sqrt{T}$, optimism $\sqrt{T}$, perturbation $\sqrt{\cdot}$.

Theorem (Degenne, Koolen, and Ménard 2019)

The sample complexity is at most

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \;\leq\; T^*(\boldsymbol{\mu}) \ln \frac{1}{\delta} + \textit{small}$$

Theorem (Degenne, Shao, and Koolen 2020)

The regret is at most

$$\mathrm{R}_T(\boldsymbol{\mu}) \;\leq\; V^*(\boldsymbol{\mu}) \ln T + \textit{small}$$

Proof ideas (cheating with optimism)

As long as we do not stop, $t < \tau$,

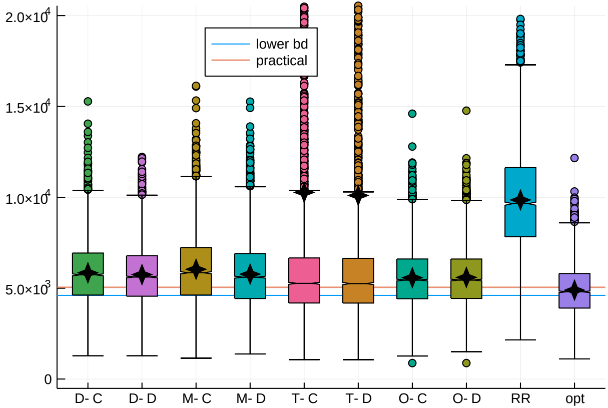
$$
\begin{aligned}
\ln \frac{1}{\delta} \approx \beta(t,\delta) &\geq \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{k=1}^{K} N_t^k d(\mu^k, \lambda^k) && \text{(stop rule)} \\
&\approx \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{s=1}^{t} \sum_{k=1}^{K} w_s^k d(\mu^k, \lambda^k) && \text{(tracking)} \\
&\geq \sum_{s=1}^{t} \sum_{k=1}^{K} w_s^k \, \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}_s} d(\mu^k, \lambda^k) - R_t^{\lambda} && \text{(regret } \boldsymbol{\lambda}\text{)} \\
&\geq \max_k \sum_{s=1}^{t} \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}_s} d(\mu^k, \lambda^k) - R_t^{\lambda} - R_t^k && \text{(regret } k\text{)} \\
&\geq t \inf_{\boldsymbol{q} \in \mathcal{P}(\mathrm{Alt}(\boldsymbol{\mu}))} \max_k \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}} d(\mu^k, \lambda^k) - O(\sqrt{t})
\end{aligned}
$$

Find maximal t to get bound on τ.

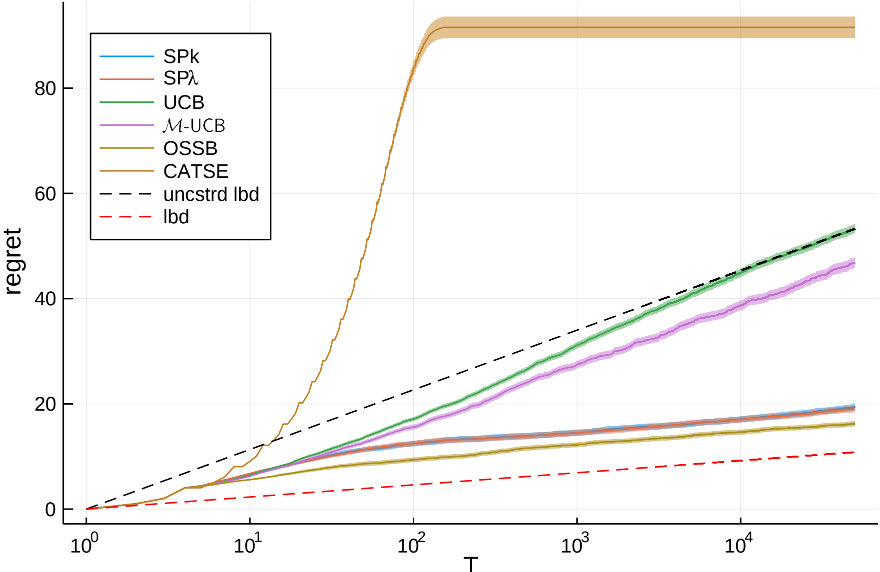
Outline

Pure Exploration Experiment: Minimum Threshold

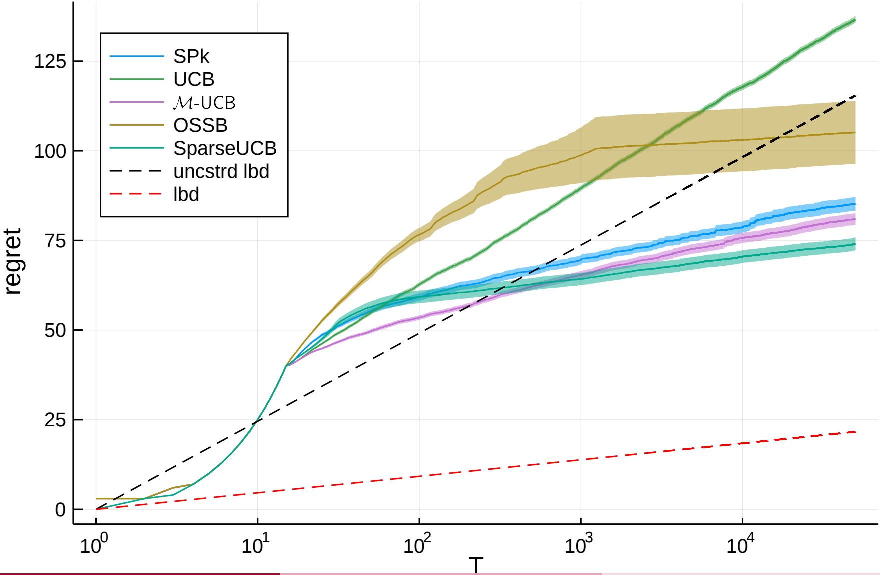
Minimum Threshold for Gaussian bandit model $\boldsymbol{\mu} = (0.5, 0.6)$ with threshold $\gamma = 0.6$, $\boldsymbol{w}^* = (1, 0)$. Note the excessive sample complexity of T-C/T-D. $\delta = 10^{-10}$.



Regret Experiment: Categorised Bandit



Regret Experiment: Sparse Bandit



Outline

Conclusion

Game equilibrium based techniques for matching **instance dependent lower bounds** for structured stochastic bandits.

Run-time determined by **Best Response oracle** for your structure.

Topics Skipped

- Pure Exploration problems with multiple correct answers (incl. ϵ-Best Arm) [Degenne and Koolen, 2019] $\Leftarrow$ **surprisingly subtle**.
- Optimal algorithms based on variations of Thompson Sampling
  - Top-Two for Best Arm [Russo, 2016]
  - Murphy Sampling for Minimum Threshold [Kaufmann et al., 2018].

Where to Next?

- Fine tuning
- What about "lower-order" terms not scaling with $\ln T$ or $\ln \frac{1}{\delta}$ [Simchowitz et al., 2017]?
- Is minigame interaction "easy data"? OMD/OFTRL? MetaGrad [van Erven and Koolen, 2016]?
- Pure Exploration Beyond Best Arm (understand sparsity patterns). Currently working on game trees. RL on the horzon.
- Minigames for other problems?
- Fixed Budget? Simple Regret?

Where to Next?

- Fine tuning
- What about "lower-order" terms not scaling with $\ln T$ or $\ln \frac{1}{\delta}$ [Simchowitz et al., 2017]?
- Is minigame interaction "easy data"? OMD/OFTRL? MetaGrad [van Erven and Koolen, 2016]?
- Pure Exploration Beyond Best Arm (understand sparsity patterns). Currently working on game trees. RL on the horzon.
- Minigames for other problems?
- Fixed Budget? Simple Regret?

Thank you!

Outline

8 Proof Ideas

9 Noise Free Case

10 The Real Deal

11 Pictures

Outline

Noise-free result

Let $\mathcal{B}_n^k$ be regret of full information online learning (AdaHedge) w. linear losses on the simplex.

Theorem

Consider running our algorithm until $\inf_{\boldsymbol{\lambda} \in \Lambda} \sum_{t=1}^n \sum_k w_t^k d(\mu^k, \lambda^k) \geq \ln T$. *The iterates* $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ *satisfy*

$$R_n \;=\; \sum_{t=1}^{n} \langle \boldsymbol{w}_t, \Delta \rangle \;\leq\; V_T + \frac{\mathcal{B}_n^k}{D^*}$$

Note

- Can get $A_1, \ldots, A_n$ using tracking (at cost $\Delta^{\max} \ln K$)
- Standard choice gives $n = O(\ln T)$ and $\mathcal{B}_n^k = O(\sqrt{n}) = O(\sqrt{\ln T}) = o(\ln T)$.

Regret analysis

Given moves $\boldsymbol{w}_t \in \triangle_K$ and $\boldsymbol{\lambda}_t \in \Lambda$, we instantiate a k-learner for the gain function

$$g_t(\tilde{\boldsymbol{w}}) \;=\; \langle \boldsymbol{w}_t, \Delta \rangle \sum_k \tilde{w}^k \frac{d(\mu^k, \lambda_t^k)}{\Delta^k}$$

to provide regret bound

$$\sum_{t=1}^n g_t(\tilde{\boldsymbol{w}}_t) \;\geq\; \max_k \sum_{t=1}^n \langle \boldsymbol{w}_t, \Delta \rangle \frac{d(\mu^k, \lambda_t^k)}{\Delta^k} - \mathcal{B}_n^k. \qquad (1)$$

Regret analysis (ctd)

Given $\tilde{\boldsymbol{w}}_t$ from the k-learner, we define player and opponent by

$$w_t^k \;\propto\; \tilde{w}_t^k / \Delta^k \tag{2}$$

$$\boldsymbol{\lambda}_t \;\in\; \operatorname*{argmin}_{\boldsymbol{\lambda} \in \Lambda} \; \sum_k w_t^k d(\mu^k, \lambda^k) \tag{3}$$

to obtain

$$\sum_{t=1}^n g_t(\tilde{\boldsymbol{w}}_t) \;=\; \sum_{t=1}^n \langle \boldsymbol{w}_t, \Delta \rangle \sum_k \tilde{w}_t^k \frac{d(\mu^k, \lambda_t^k)}{\Delta^k} \;\overset{(2)}{=}\; \sum_{t=1}^n \sum_k w_t^k d(\mu^k, \lambda_t^k)$$
$$\overset{(3)}{=}\; \sum_{t=1}^n \inf_{\boldsymbol{\lambda} \in \Lambda} \sum_k w_t^k d(\mu^k, \lambda^k) \;\le\; \inf_{\boldsymbol{\lambda} \in \Lambda} \; \sum_{t=1}^n \sum_k w_t^k d(\mu^k, \lambda^k) \tag{4}$$

Regret analysis (ctd)

The stopping condition plus regret bounds (1) and (4) result in

$$\begin{aligned}
\ln T + \mathcal{B}_n^k \;&\geq\; \max_k \sum_{t=1}^n \langle \boldsymbol{w}_t, \Delta \rangle \frac{d(\mu^k, \lambda_t^k)}{\Delta^k} \;=\; R_n \max_k \sum_{t=1}^n \frac{\langle \boldsymbol{w}_t, \Delta \rangle}{R_n} \frac{d(\mu^k, \lambda_t^k)}{\Delta^k} \\
&\geq\; R_n \inf_{\boldsymbol{q} \in \triangle(\Lambda)} \max_k \frac{\mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}} \left[d(\mu^k, \lambda^k) \right]}{\Delta^k} \;=\; R_n D^*
\end{aligned}$$

where we abbreviated $R_n = \sum_{t=1}^n \langle \boldsymbol{w}_t, \Delta \rangle$. All in all we showed

$$R_n \;\leq\; V_T + \frac{\mathcal{B}_n^k}{D^*}$$

Outline

Scaling up

Can use what we developed so far to compute oracle weights every round (OSSB). Efficient for **every** bandit structure for which best response is tractable.

Scaling up

Can use what we developed so far to compute oracle weights every round (OSSB). Efficient for **every** bandit structure for which best response is tractable.

But we can do much better!

Scaling up

Can use what we developed so far to compute oracle weights every round (OSSB). Efficient for **every** bandit structure for which best response is tractable.

But we can do much better!
Idea:

- Run only one iteration every round.
- Deal with unknown $\boldsymbol{\mu}$.
- Exploitation.

some issues . . .

First Issue

Actually, $\Delta^* = 0$. And we were dividing by it all over the place.

First Issue

Actually, $\Delta^* = 0$. And we were dividing by it all over the place.

Idea: run on $\Delta^k_\epsilon = \max\left\{\Delta^k, \epsilon\right\}$.

Theorem

$$\lim_{\epsilon \to 0} V^\epsilon_T \ = \ V_T$$

In several cases we can show perturbed value is $V^\epsilon_T \le V_T + \sqrt{2\epsilon V_T}$.

One iteration every round

- Replace $\boldsymbol{\mu}$ by **estimate** $\hat{\boldsymbol{\mu}}_t$.
- Add **optimism** to force exploration.
  We introduce upper confidence bounds on the ratio KL/gap.

$$\mathrm{UCB}_s^k = \sup_{\xi \in \mathcal{C}_{s-1}^k} \frac{d(\xi, \lambda_t^k)}{\max\left\{\epsilon_s, 1\{k \neq j_s\}\left[\mu_{s-1}^+ - \xi\right]\right\}}$$

$$\text{where } \mathcal{C}_{s-1}^k = \left[\hat{\mu}_{s-1}^k \pm \sqrt{\frac{\overline{\ln}(n_{s-1}^{j_s}, N_{s-1}^k)}{N_{s-1}^k}}\right].$$

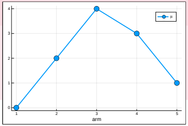
- We do not know **identity of the best arm**, and hence Λ (domain of $\boldsymbol{\lambda}$) Estimate best arm, and run K independent interactions.

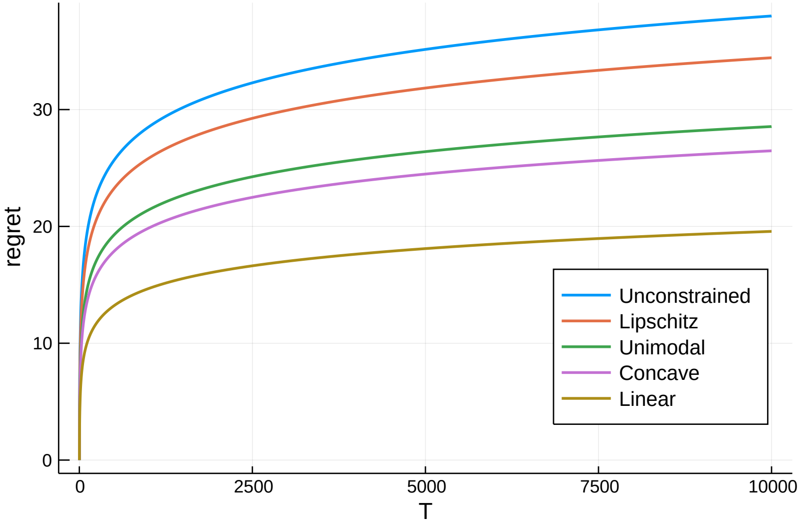
Algorithm

Pull each arm once and get $\hat{\boldsymbol{\mu}}_K$.
for $t = K+1, \cdots, T$ **do**
  if $\exists i \in [K]$, $\min_{\boldsymbol{\lambda} \in \neg i} \sum_k N^k_{t-1} d(\hat{\mu}^k_{t-1}, \lambda^k) > f(t-1)$ **then**
    $A_t = i$ (if there are several suitable i, pull any one of them)
  else
    $\mu^+_{t-1}, j_t = (\arg)\max_{j \in [K]} \hat{\mu}^j_{t-1} + \sqrt{\frac{\overline{\ln}(n^j_{t-1}, N^j_{t-1})}{N^j_{t-1}}}$.
    get $\tilde{\boldsymbol{w}}_t$ from learner $\mathcal{A}^k_{j_t}$, compute $w^k_t \propto \tilde{w}^k_t / \tilde{\Delta}^k$.
    compute best response $\boldsymbol{\lambda}_t$.
    Compute $\mathrm{UCB}^k_t = \max_{\xi \in [\hat{\mu}^k_{t-1} - \ldots, \hat{\mu}^k_{t-1} + \ldots]} \left[\frac{d(\xi, \lambda^k_t)}{\max\{\varepsilon_t, 1\{k \neq j_t\}[\mu^+_{t-1} - \xi]\}} \right]$
    $A_t = \operatorname{argmin}_{k \in [K]} N^k_{t-1} - \sum_{s=1}^t w^k_s$. $\triangleright$ Tracking
  end if
  Access $X^{A_t}_t$, update $\hat{\boldsymbol{\mu}}_t$ and $\boldsymbol{N}_t$
end for

Outline

Desired behaviour





Illustration