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

1 Ideas
2 Problem Settings
3 Lower Bounds
4 Algorithms
5 Iterative Saddle-Point Methods
6 Experiments
7 Conclusion
Stochastic Bandit

\[ P(I) = \frac{1}{6} \]
\[ P(I) = \frac{2}{3} \]
\[ P(I) = \frac{1}{2} \]
Stochastic Bandit

Model (Unknown)

\[
P(\text{Smiley} | \text{B}) = \frac{1}{6}
\]

\[
P(\text{Smiley} | \text{VC}) = \frac{2}{3}
\]

\[
P(\text{Smiley} | \text{Plus}) = \frac{1}{2}
\]
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)

\[ P(\text{arm} 1) = \frac{1}{6} \]
\[ P(\text{arm} 2) = \frac{3}{6} \]
\[ P(\text{arm} 3) = \frac{5}{6} \]
\[ P(\text{arm} 4) = \frac{4}{6} \]
\[ P(\text{arm} 5) = \frac{2}{6} \]
Structured Stochastic Bandit Interaction
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

1. Ideas
2. Problem Settings
3. Lower Bounds
4. Algorithms
5. Iterative Saddle-Point Methods
6. Experiments
7. Conclusion
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 $\mu = (\mu_1, \ldots, \mu_K)$. 
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 $\mu = (\mu_1, \ldots, \mu_K)$.

**Learning Target**

The best arm for $\mu$ is

$$i^*(\mu) := \arg\max_i \mu_i$$
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 $\mu = (\mu_1, \ldots, \mu_K)$.

Learning Target

The best arm for $\mu$ is

$$i^*(\mu) := \arg\max_i \mu_i$$

Structure

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

Interaction

\[ X_t \sim \mu A_t \]

Learner \rightarrow \text{Bandit } \mu \rightarrow A_t \rightarrow \text{Learner}
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: $H := (A_1, X_1, \ldots, A_{\tau}, X_{\tau}, \hat{i})$.
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: $\mathcal{H} := (A_1, X_1, \ldots, A_\tau, X_\tau, \hat{i})$. 
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: $\mathcal{H} := (A_1, X_1, \ldots, A_\tau, X_\tau, \hat{i})$.

Two objectives: **sample efficiency** $\tau$ and **correctness** $\hat{i} = i^*(\mu)$. 
Best Arm Identification Goal: PAC learning

Definition

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

$$\mathbb{P}_\mu(\hat{l} \neq i^*(\mu)) \leq \delta$$

for every bandit model $\mu \in \mathcal{M}$. 

Best Arm Identification Goal: PAC learning

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

$$\mathbb{P}_\mu(\hat{i} \neq i^*(\mu)) \leq \delta$$

for every bandit model $\mu \in \mathcal{M}$.

Goal: minimise sample complexity $\mathbb{E}_\mu[\tau]$ over all $\delta$-correct strategies.
Best Arm Identification Goal: PAC learning

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

$$\Pr_{\mu}(\hat{I} \neq i^*(\mu)) \leq \delta$$

for every bandit model $\mu \in M$.

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

Hope
Efficient $\delta$-correct algorithm with instance-optimal sample complexity

$$\mathbb{E}_{\mu}[\tau] \preceq \Box_{\mu} \ln \frac{1}{\delta}$$

for all $\mu \in M$. 
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: \( \mathcal{H} := (A_1, X_1, \ldots, A_T, X_T) \).

**Definition**

The objective is

\[
R_T(\mu) := \sum_{k=1}^{K} \mathbb{E}[N_T^k] \Delta^k
\]

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: $\mathcal{H} := (A_1, X_1, \ldots, A_T, X_T)$.

Definition

The objective is

$$R_T(\mu) := \sum_{k=1}^{K} \mathbb{E}[N^k_T] \Delta^k$$

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

Hope

Efficient algorithm with instance-optimal regret

$$R_T(\mu) \leq \square_\mu \ln T \quad \text{for all } \mu \in \mathcal{M}.$$
Outline

1. Ideas
2. Problem Settings
3. Lower Bounds
4. Algorithms
5. Iterative Saddle-Point Methods
6. Experiments
7. Conclusion
Instance-Dependent Sample Complexity Lower Bound

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

Define the alternative to $\mu$ by $\text{Alt}(\mu) := \{\lambda \in M | i^*(\lambda) \neq i^*(\mu)\}$. 
Instance-Dependent Sample Complexity Lower Bound

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

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

**Theorem (Castro 2014, Garivier and Kaufmann 2016)**

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

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

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

$$\frac{1}{T^*(\mu)} = \max_{w \in \Delta_K} \min_{\lambda \in \text{Alt}(\mu)} \sum_{i=1}^{K} w_i d(\mu_i, \lambda_i) \propto N^k \text{pulls}$$
Example

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

\[ T^*(\mu) = 200.4 \quad w^*(\mu) = (0.45, 0.46, 0.06, 0.02, 0.01) \]

At confidence \( \delta = 0.05 \) we have \( \ln \frac{1}{\delta} = 3.0 \) and hence \( \mathbb{E}_\mu[\tau] \geq 601.2 \).
Instance-Dependent Regret Lower Bound

Theorem (Graves and Lai 1997)

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

$$R_T(\mu) \geq V(\mu) \ln T$$

where the characteristic regret rate is given by

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

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

1. Ideas
2. Problem Settings
3. Lower Bounds
4. Algorithms
5. Iterative Saddle-Point Methods
6. Experiments
7. Conclusion
Recall sample complexity/regret lower bound governed by

\[
\max_{\boldsymbol{w} \in \Delta_K} \min_{\lambda \in \text{Alt}(\mu)} \sum_{i=1}^{K} w_i d(\mu_i, \lambda_i) \quad \text{or} \quad \max_{\tilde{\boldsymbol{w}} \in \Delta} \inf_{\lambda \in \text{Alt}(\mu)} \sum_{k} \tilde{w}_k \frac{d(\mu_k, \lambda_k)}{\Delta_k}
\]
Lower Bounds Inspire Strategies

Recall sample complexity/regret lower bound governed by

$$\max_{w \in \Delta_K} \min_{\lambda \in \text{Alt}(\mu)} \sum_{i=1}^{K} w_i d(\mu_i, \lambda_i)$$

or

$$\max_{\tilde{w} \in \Delta} \inf_{\lambda \in \text{Alt}(\mu)} \sum_{k} \tilde{w}_k \frac{d(\mu^k, \lambda^k)}{\Delta^k}$$

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** $w^*(\hat{\mu}_t)$ or $\bar{w}^*(\hat{\mu}_t)$.
- sample arm $A_t$ to track that solution
- **force exploration** to ensure $\hat{\mu}_t \rightarrow \mu$. 

Iteratively solve lower bounds by full information online learning.
Use iterates to drive sampling rule.
Add optimism to induce exploration.
Cap gap estimates $\hat{\Delta}_t$ from below to reduce estimation variance.

Compose regret bound from saddle-point regret + estimation regret.
Lower Bounds Inspire Strategies

Earlier work [Combes et al., 2017, Garivier and Kaufmann, 2016]
At each time step
- compute plug-in oracle solution $w^*(\hat{\mu}_t)$ or $\tilde{w}^*(\hat{\mu}_t)$.
- sample arm $A_t$ to track that solution
- force exploration to ensure $\hat{\mu}_t \to \mu$.

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 $\Delta_t$ from below to reduce estimation variance
- Compose regret bound from saddle-point regret + estimation regret
Iterative Saddle-Point Methods

Outline

1. Ideas
2. Problem Settings
3. Lower Bounds
4. Algorithms
5. Iterative Saddle-Point Methods
6. Experiments
7. Conclusion
Interleaved Iterative Solution

Standard technique: can approximately solve saddle point problems like

\[
\max_{\vec{w} \in \Delta_K} \min_{\lambda \in \text{Alt}(\mu)} \sum_{i=1}^{K} w_i d(\mu_i, \lambda_i)
\]

or

\[
\max_{\vec{\tilde{w}} \in \Delta} \inf_{\lambda \in \text{Alt}(\mu)} \sum_{k} \tilde{w}_k \frac{d(\mu^k, \lambda^k)}{\Delta_k}
\]

iteratively using two online learners.

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

- Plug-in estimate \( \hat{\mu}_t \) (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

$$\text{argmin}_{\lambda \in \text{Alt}(\hat{\mu}_t)} \sum_k w_t^k d(\hat{\mu}_t^k, \lambda^k)$$
Sampling Rule for Regret Minimisation

$$\text{argmin}_{\lambda \in \text{Alt}(\hat{\mu}_t)} \sum_k w_t^k d(\hat{\mu}^k_t, \lambda^k)$$

AdaHedge

$k$-learner

Best response

$\lambda_t$

$\nabla_t$

$\tilde{w}_t$

$w_t$

Tracking

$w_t$

$A_t$

Bandit

$\hat{\mu}_t$

Estimate

$\mathcal{X}_t \sim \mu_{A_t}$
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*

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

**Theorem (Degenne, Shao, and Koolen 2020)**

*The regret is at most*

$$
R_T(\mu) \leq V^*(\mu) \ln T + \text{small}
$$
Proof ideas (cheating with optimism)

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

$$\ln \frac{1}{\delta} \approx \beta(t, \delta) \geq \inf_{\lambda \in \text{Alt}(\mu)} \sum_{k=1}^{K} N_t^k d(\mu^k, \lambda^k)$$  \hspace{1cm} \text{(stop rule)}

$$\approx \inf_{\lambda \in \text{Alt}(\mu)} \sum_{s=1}^{t} \sum_{k=1}^{K} w_s^k d(\mu^k, \lambda^k)$$  \hspace{1cm} \text{(tracking)}

$$\geq \sum_{s=1}^{t} \sum_{k=1}^{K} w_s^k \mathbb{E}_{\lambda \sim q} d(\mu^k, \lambda^k) - R^\lambda_t$$  \hspace{1cm} \text{(regret $\lambda$)}

$$\geq \max_{k} \sum_{s=1}^{t} \mathbb{E}_{\lambda \sim q} d(\mu^k, \lambda^k) - R^\lambda_t - R^k_t$$  \hspace{1cm} \text{(regret $k$)}

$$\geq t \inf_{q \in \mathcal{P}(\text{Alt}(\mu))} \max_k \mathbb{E}_{\lambda \sim q} d(\mu^k, \lambda^k) - O(\sqrt{t})$$

Find maximal $t$ to get bound on $\tau$. 
Outline

1. Ideas
2. Problem Settings
3. Lower Bounds
4. Algorithms
5. Iterative Saddle-Point Methods
6. Experiments
7. Conclusion
Minimum Threshold for Gaussian bandit model $\mu = (0.5, 0.6)$ with threshold $\gamma = 0.6$, $\mathbf{w}^* = (1, 0)$. Note the excessive sample complexity of $T$-C/T-D. $\delta = 10^{-10}$. 
Regret Experiment: Categorised Bandit

The graph shows the regret over time for different strategies in a categorised bandit problem. The x-axis represents time (T), and the y-axis represents regret. The strategies compared include SPk, SPλ, UCB, $\mathcal{M}$-UCB, OSSB, CATSE, uncstrd lbd, and lbd.

- SPk: Blue line
- SPλ: Red line
- UCB: Green line
- $\mathcal{M}$-UCB: Pink line
- OSSB: Orange line
- CATSE: Yellow line
- uncstrd lbd: Black dashed line
- lbd: Red dashed line

The figure illustrates how each strategy performs over time, with different lines indicating the performance of each approach. The regret increases as time progresses, showing the trade-off between exploration and exploitation in decision-making processes.
Regret Experiment: Sparse Bandit

![Graph showing regret curves for different algorithms including SPk, UCB, OSS, and SparseUCB, with curves for unstrd lbd and lbd marked.

Title: Bandits, Games, Explore/Exploit

Author: Wouter Koolen

Page: 35 / 53
Outline

1. Ideas
2. Problem Settings
3. Lower Bounds
4. Algorithms
5. Iterative Saddle-Point Methods
6. Experiments
7. 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. $\epsilon$-Best Arm) [Degenne and Koolen, 2019] $\iff$ *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 horizon.
- 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 horizon.
- 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

8 Proof Ideas
9 Noise Free Case
10 The Real Deal
11 Pictures
Noise-free result

Let $B_n^k$ be regret of full information online learning (AdaHedge) w. linear losses on the simplex.

Theorem

Consider running our algorithm until
$$\inf_{\lambda \in \Lambda} \sum_{t=1}^{n} \sum_{k} w_t^k d(\mu^k, \lambda^k) \geq \ln T.$$ The iterates $w_1, \ldots, w_n$ satisfy

$$R_n = \sum_{t=1}^{n} \langle w_t, \Delta \rangle \leq V_T + \frac{B_n^k}{D^*}$$

Note

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

Given moves $\mathbf{w}_t \in \Delta_K$ and $\lambda_t \in \Lambda$, we instantiate a $k$-learner for the gain function

$$g_t(\tilde{w}) = \langle \mathbf{w}_t, \Delta \rangle \sum_k \tilde{w}_k \frac{d(\mu^k, \lambda^k_t)}{\Delta_k}$$

to provide regret bound

$$\sum_{t=1}^{n} g_t(\tilde{w}_t) \geq \max_k \sum_{t=1}^{n} \langle \mathbf{w}_t, \Delta \rangle \frac{d(\mu^k, \lambda^k_t)}{\Delta_k} - B_n^k. \quad (1)$$
Given $\tilde{w}_t$ from the $k$-learner, we define player and opponent by

$$w_t^k \propto \tilde{w}_t^k / \Delta^k$$  \hspace{1cm} (2)

$$\lambda_t \in \arg\min_{\lambda \in \Lambda} \sum_k w_t^k d(\mu_k^k, \lambda_k^t)$$  \hspace{1cm} (3)

to obtain

$$\sum_{t=1}^n g_t(\tilde{w}_t) = \sum_{t=1}^n \langle w_t, \Delta \rangle \sum_k \tilde{w}_t^k d(\mu_k^k, \lambda_k^t) \quad \overset{(2)}{=} \quad \sum_{t=1}^n \sum_k w_t^k d(\mu_k^k, \lambda_k^t)$$

$$\overset{(3)}{=} \quad \sum_{t=1}^n \inf_{\lambda \in \Lambda} \sum_k w_t^k d(\mu_k^k, \lambda_k^t) \leq \inf_{\lambda \in \Lambda} \sum_{t=1}^n \sum_k w_t^k d(\mu_k^k, \lambda_k^t)$$  \hspace{1cm} (4)
Regret analysis (ctd)

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

$$\ln T + \mathcal{B}_n^k \geq \max_k \sum_{t=1}^n \langle w_t, \Delta \rangle \frac{d(\mu^k, \lambda^k_t)}{\Delta^k} = R_n \max_k \sum_{t=1}^n \frac{\langle w_t, \Delta \rangle}{R_n} \frac{d(\mu^k, \lambda^k_t)}{\Delta^k}$$

$$\geq R_n \inf_{q \in \triangle(\Lambda)} \max_k \frac{\mathbb{E}_{\lambda \sim q} [d(\mu^k, \lambda^k)]}{\Delta^k} = R_n D^*$$

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

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

8 Proof Ideas

9 Noise Free Case

10 The Real Deal

11 Pictures
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 $\mu$.
- Exploitation.

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

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

**Theorem**

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

In several cases we can show perturbed value is $V^\epsilon_T \leq V_T + \sqrt{2\epsilon V_T}$. 
One iteration every round

- Replace $\mu$ by estimate $\hat{\mu}_t$.
- Add **optimism** to force exploration.

We introduce upper confidence bounds on the ratio $\text{KL}/\text{gap}$.

$$\text{UCB}^k_s = \sup_{\xi \in C_{s-1}^k} d(\xi, \lambda^k_t) \max \left\{ \epsilon_s, 1 \{ k \neq j_s \} \left[ \mu^+_{s-1} - \xi \right] \right\}$$

where $C_{s-1}^k = \left\lfloor \hat{\mu}_{s-1}^k \pm \sqrt{\ln\left( n_{s-1}^{j_s}, N_{s-1}^k \right) \frac{\ln(n_{s-1}^{j_s}, N_{s-1}^k)}{N_{s-1}^k}} \right\rfloor$.

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

1: Pull each arm once and get $\hat{\mu}_K$.
2: for $t = K + 1, \cdots, T$ do
3: \hspace{1em} if $\exists i \in [K], \min_{\lambda \in \Lambda} \sum_k N_{t-1}^k d(\hat{\mu}_{t-1}^k, \lambda^k) > f(t - 1)$ then
4: \hspace{2em} $A_t = i$ (if there are several suitable $i$, pull any one of them)
5: \hspace{1em} else
6: \hspace{2em} $\mu_{t-1}^+, j_t = (\arg) \max_{j \in [K]} \hat{\mu}_{t-1}^j + \sqrt{\frac{\ln(n_{t-1}^i, N_{t-1}^i)}{N_{t-1}^j}}$.
7: \hspace{1em} get $\tilde{w}_t$ from learner $A_{j_t}^k$, compute $w_t^k \propto \tilde{w}_t^k / \tilde{\Delta}^k$.
8: \hspace{1em} compute best response $\lambda_t$.
9: \hspace{1em} Compute $\text{UCB}_t^k = \max_{\xi \in [\hat{\mu}_{t-1}^k - \ldots, \hat{\mu}_{t-1}^k + \ldots]} \left[ \frac{d(\xi, \lambda_t^k)}{\max\{\epsilon_t, 1\{k \neq j_t\}[\mu_{t-1}^+ - \xi]\}} \right]$.
10: \hspace{1em} $A_t = \arg\min_{k \in [K]} N_{t-1}^k - \sum_{s=1}^t w_s^k$. $\triangleright$ Tracking
11: \hspace{1em} end if
12: \hspace{1em} Access $X_t^{A_t}$, update $\hat{\mu}_t$ and $N_t$
13: end for
Outline

8 Proof Ideas

9 Noise Free Case

10 The Real Deal

11 Pictures
Desired behaviour

![Graph showing regret over time for different types of functions: Unconstrained, Lipschitz, Unimodal, Concave, Linear. Each function type is represented by a different color line, indicating the trend of regret as time progresses.](#)
Illustration

Unconstrained

Lipschitz

Unimodal

Concave

Linear

Support for Lipschitz

Wouter Koolen

Bandits, Games, Explore/Exploit

Noberwolfach