The Pure Exploration Renaissance

Deep Dive @ Booking.com

Wouter Koolen
December 10th, 2020

CWI
Centrum Wiskunde & Informatica
Lay of the Land

Bandits

- reward maximisation
- pure exploration
  - fixed budget
  - fixed confidence
  - simple regret
    - confidence intervals
    - Thompson sampling
    - Track and Stop
Pure Exploration is statistical hypothesis testing on steroids:

- Multiple
- Composite
- Sequential
- Active
• Introduce Pure Exploration problems.
• Sketch the GLR stopping rule
• Sketch the TaS sampling rule
• Highlight some recent lessons learned.
Pure Exploration
Introduction
Best Arm Identification (BAI)

Assumption: Bernoulli Multi-Armed Bandit

$K$ Bernoulli arms with unknown means $\mu = (\mu_1, \ldots, \mu_K) \in [0, 1]^K$. 
Assumption: Bernoulli Multi-Armed Bandit

$K$ Bernoulli arms with unknown means $\mu = (\mu_1, \ldots, \mu_K) \in [0, 1]^K$.

BAI-MAB Protocol

\begin{verbatim}
for $t = 1, 2, \ldots$ until Learner decides to stop
    • Learner picks arm $A_t \in [K]$
    • Learner observes $X_t \sim \text{Bernoulli} (\mu_{A_t})$

Learner recommends $\hat{i} \in [K]$.
\end{verbatim}
Best Arm Identification

**BAI-MAB Protocol**

```plaintext
for $t = 1, 2, \ldots$ until Learner decides to stop
  • Learner picks arm $A_t \in [K]$
  • Learner observes $X_t \sim \text{Bernoulli} (\mu_{A_t})$

Learner recommends $\hat{i} \in [K]$. 
```

Let $\tau \in \mathbb{N} \cup \{\infty\}$ denote the # rounds after which Learner stops.
Best Arm Identification

BAI-MAB Protocol

**for** $t = 1, 2, \ldots \text{ until Learner decides to stop}$

- Learner picks arm $A_t \in [K]$
- Learner observes $X_t \sim \text{Bernoulli}(\mu_{A_t})$

Learner recommends $\hat{i} \in [K]$.

Let $\tau \in \mathbb{N} \cup \{\infty\}$ denote the # rounds after which Learner stops.

**Definition**

Learner is $\delta$-PAC if

$$\mathbb{P}_\mu \left\{ \tau < \infty \text{ and } \hat{i} \neq \arg \max_i \mu_i \right\} \leq \delta \quad \text{for all } \mu \in [0, 1]^K.$$

a mistake
Best Arm Identification

**BAI-MAB Protocol**

```plaintext
for \( t = 1, 2, \ldots \) until Learner decides to stop
  • Learner picks arm \( A_t \in [K] \)
  • Learner observes \( X_t \sim \text{Bernoulli}(\mu_{A_t}) \)

Learner recommends \( \hat{i} \in [K] \).
```

Let \( \tau \in \mathbb{N} \cup \{\infty\} \) denote the number of rounds after which Learner stops.

**Definition**

Learner is \( \delta \)-PAC if

\[
\mathbb{P}_\mu \left\{ \tau < \infty \text{ and } \hat{i} \neq \arg \max_i \mu_i \right\} \leq \delta \quad \text{for all } \mu \in [0, 1]^K.
\]

**Definition**

We call \( \mathbb{E}_\mu[\tau] \) the sample complexity of Learner in bandit \( \mu \).
Best Arm Identification

**BAI-MAB Protocol**

for $t = 1, 2, \ldots$ until Learner decides to stop
  - Learner picks arm $A_t \in [K]$
  - Learner observes $X_t \sim \text{Bernoulli}(\mu_{A_t})$

Learner recommends $\hat{i} \in [K]$.

Let $\tau \in \mathbb{N} \cup \{\infty\}$ denote the # rounds after which Learner stops.

**Definition**

Learner is $\delta$-PAC if

$$\mathbb{P}_\mu \left\{ \tau < \infty \text{ and } \hat{i} \neq \arg \max_i \mu_i \right\} \leq \delta \quad \text{for all } \mu \in [0, 1]^K.$$  

**Definition**

We call $\mathbb{E}_\mu[\tau]$ the *sample complexity* of Learner in bandit $\mu$.

**Goal:** efficient $\delta$-PAC algorithms with minimal sample complexity.
Goal: efficient $\delta$-PAC algorithms with minimal sample complexity.
**Goal**: efficient $\delta$-PAC algorithms with minimal sample complexity.

<table>
<thead>
<tr>
<th>Fancy Algorithm($\delta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop when ...</td>
</tr>
<tr>
<td>Sample arm $A_t = ...$</td>
</tr>
<tr>
<td>Recommend $\hat{i} = ...$</td>
</tr>
</tbody>
</table>
Best Arm Identification, prototypical solution

Goal: efficient $\delta$-PAC algorithms with minimal sample complexity.

<table>
<thead>
<tr>
<th>Fancy Algorithm($\delta$)</th>
<th>Theorem (safe)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop when ...</td>
<td>Fancy Algorithm($\delta$) is $\delta$-PAC</td>
</tr>
<tr>
<td>Sample arm $A_t = \ldots$</td>
<td></td>
</tr>
<tr>
<td>Recommend $\hat{i} = \ldots$</td>
<td></td>
</tr>
</tbody>
</table>
**Best Arm Identification, prototypical solution**

**Goal:** efficient $\delta$-PAC algorithms with minimal sample complexity.

<table>
<thead>
<tr>
<th>Fancy Algorithm$(\delta)$</th>
<th>Theorem (safe)</th>
<th>Theorem (comput. eff.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop when $\ldots$</td>
<td>$Fancy\ Algorithm(\delta)$ is $\delta$-PAC</td>
<td>$\ldots$ runs in time $O(\ldots)$</td>
</tr>
<tr>
<td>Sample arm $A_t = \ldots$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recommend $\hat{i} = \ldots$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Best Arm Identification, prototypical solution

**Goal**: efficient $\delta$-PAC algorithms with minimal sample complexity.

<table>
<thead>
<tr>
<th>Fancy Algorithm($\delta$)</th>
<th>Theorem (safe)</th>
<th>Theorem (comput. eff.)</th>
<th>Theorem (statistic. eff.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop when ...</td>
<td><em>Fancy Algorithm($\delta$) is $\delta$-PAC</em></td>
<td>... runs in time $O(\ldots)$</td>
<td>... has sample complexity</td>
</tr>
<tr>
<td>Sample arm $A_t = \ldots$</td>
<td></td>
<td></td>
<td>$\mathbb{E}_\mu[\tau] \leq f(\mu) \ln \frac{1}{\delta} + o(\ln \frac{1}{\delta})$.</td>
</tr>
<tr>
<td>Recommend $\hat{I} = \ldots$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Best Arm Identification, prototypical solution

**Goal**: efficient $\delta$-PAC algorithms with minimal sample complexity.

<table>
<thead>
<tr>
<th>Fancy Algorithm($\delta$)</th>
<th>Theorem (safe)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop when ...</td>
<td><em>Fancy Algorithm($\delta$) is $\delta$-PAC</em></td>
</tr>
<tr>
<td>Sample arm $A_t = \ldots$</td>
<td>... runs in time $O(\ldots)$</td>
</tr>
<tr>
<td>Recommend $\hat{i} = \ldots$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Theorem (lower bd)</th>
<th>Theorem (comput. eff.)</th>
<th>Theorem (statistic. eff.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any $\delta$-PAC algorithm needs sample complexity at least $\mathbb{E}_\mu[\tau] \geq f(\mu) \ln \frac{1}{\delta}$</td>
<td>... has sample complexity $\mathbb{E}_\mu[\tau] \leq f(\mu) \ln \frac{1}{\delta} + o(\ln \frac{1}{\delta})$.</td>
<td></td>
</tr>
</tbody>
</table>
Assumptions about arm distributions

- Exponential Family:
  Gaussian, Gamma, Poisson, Geometric, ...

- Non-parametric:
  bounded support, sub-Gaussian, \((1 + \epsilon)^{\text{th}}\) moment, ...
Many Variations of BAI Problem in Literature

Assumptions about arm distributions

- Exponential Family:
  Gaussian, Gamma, Poisson, Geometric, …

- Non-parametric:
  bounded support, sub-Gaussian, \((1 + \epsilon)^{\text{th}}\) moment, …

Assume prior knowledge of structured bandit model class

- Linear, Lipschitz, Sparse, Categorical, Unimodal, …

Identify a near-optimal arm: Learner is \((\epsilon, \delta)\)-PAC if

\[
P_{\mu} \{ \tau < \infty \text{ and } \mu^\hat{I} < \max_i \mu_i - \epsilon \} \leq \delta \quad \text{for all } \mu \in [0, 1]^K.
\]
Many Variations of BAI Problem in Literature

Assumptions about arm distributions

- Exponential Family:
  Gaussian, Gamma, Poisson, Geometric, . . .
- Non-parametric:
  bounded support, sub-Gaussian, \((1 + \epsilon)^{th}\) moment, . . .

Assume prior knowledge of structured bandit model class

- Linear, Lipschitz, Sparse, Categorical, Unimodal, . . .

Identify a near-optimal arm: Learner is \((\epsilon, \delta)\)-PAC if

\[
P_{\mu} \left\{ \tau < \infty \text{ and } \mu_j < \max_i \mu_i - \epsilon \right\} \leq \delta \quad \text{for all } \mu \in [0, 1]^K.\]
Many Variations of BAI Problem in Literature

Assumptions about arm distributions

- Exponential Family:
  Gaussian, Gamma, Poisson, Geometric, ...
- Non-parametric:
  bounded support, sub-Gaussian, \((1 + \epsilon)^{th}\) moment, ...

Assume prior knowledge of structured bandit model class

- Linear, Lipschitz, Sparse, Categorical, Unimodal, ...

Identify a near-optimal arm: Learner is \((\epsilon, \delta)\)-PAC if

\[
\mathbb{P}_{\mu} \left\{ \tau < \infty \text{ and } \mu_{\hat{i}} < \max_{i} \mu_{i} - \epsilon \right\} \leq \delta \quad \text{for all } \mu \in [0, 1]^K.
\]

Feedback graphs (semi-bandit, ...)
A/B Testing Problems
A/B Testing Problems

Input: baseline arm $\mu_0$, candidates $\mu_1, \ldots, \mu_K$.

Problem (Best Arm Identification)

$$i^*(\mu) = \arg\max_{i \in \{0, \ldots, K\}} \mu_i$$
**A/B Testing Problems**

Input: baseline arm $\mu_0$, candidates $\mu_1, \ldots, \mu_K$.

**Problem (Best Arm Identification)**

$$i^*(\mu) = \arg \max_{i \in \{0, \ldots, K\}} \mu_i$$

**Problem (A/B test, decision version)**

$$i^*(\mu) = 1 \left\{ \max_{i \in \{1, \ldots, K\}} \mu_i > \mu_0 \right\}$$
A/B Testing Problems

Input: baseline arm $\mu_0$, candidates $\mu_1, \ldots, \mu_K$.

Problem (A/B test, identification version)

$$i^*(\mu) = \begin{cases} 
\{0\} & \mu_0 > \max_{i \in \{1, \ldots, K\}} \mu_i \\
\{i \in \{1, \ldots, K\} | \mu_i > \mu_0\} & \text{o.w.}
\end{cases}$$

*this is a multiple-answer problem*
A/B Testing Problems

Input: baseline arm $\mu_0$, candidates $\mu_1, \ldots, \mu_K$.

Problem (A/B test, identification version)

$$i^*(\mu) = \begin{cases} 
\{0\} & \mu_0 > \max_{i \in \{1, \ldots, K\}} \mu_i \\
\{i \in \{1, \ldots, K\} | \mu_i > \mu_0\} & \text{o.w.}
\end{cases}$$

this is a multiple-answer problem

Problem (A/B test, thresholding version)

$$i^*(\mu) = \{i \in \{1, \ldots, K\} | \mu_i > \mu_0\}$$

this is a set-valued single-answer problem
More Advanced Testing Problems

Input: arms $\mu_{i,j}$, $i \in \{1, \ldots, K\}$, $j \in \{1, \ldots, M\}$

Problem (Maximin Action Identification)

$$i^*(\mu) = \arg \max_{i \in \{1, \ldots, K\}} \min_{j \in \{1, \ldots, M\}} \mu_{i,j}$$
GLRT Stopping
Stopping

When can we stop?

There is no plausible bandit model $\lambda$ on which $\hat{\mathbf{i}}$ is wrong.

Definition

Generalized Likelihood Ratio (GLR) measure of evidence

$$GLR_n(\hat{\mathbf{i}}) := \ln \sup_{\mu : \hat{\mathbf{i}} \in \mathbf{i}^*} \frac{P(X_n | A_n, \mu)}{P(X_n | A_n, \lambda)}$$

Idea: stop when $GLR_n(\hat{\mathbf{i}})$ is big for some answer $\hat{\mathbf{i}}$. \

When can we stop?

When can we stop and give answer ✰?
When can we stop?

When can we stop and give answer $\hat{i}$?

There is no plausible bandit model $\lambda$ on which $\hat{i}$ is wrong.
Stopping

When can we stop?

When can we stop and give answer $\hat{i}$?

There is no plausible bandit model $\lambda$ on which $\hat{i}$ is wrong.

**Definition**

Generalized Likelihood Ratio (GLR) measure of evidence

$$\text{GLR}_n(\hat{i}) := \ln \frac{\sup_{\mu: \hat{i} \in i^*(\mu)} P(X^n | A^n, \mu)}{\sup_{\lambda: \hat{i} \not\in i^*(\lambda)} P(X^n | A^n, \lambda)}$$
Stopping

When can we stop?
When can we stop and give answer $\hat{i}$?

There is no plausible bandit model $\lambda$ on which $\hat{i}$ is wrong.

**Definition**

Generalized Likelihood Ratio (GLR) measure of evidence

$$\text{GLR}_n(\hat{i}) := \ln \frac{\sup_{\mu : \hat{i} \in i^*(\mu)} P(X^n | A^n, \mu)}{\sup_{\lambda : \hat{i} \not\in i^*(\lambda)} P(X^n | A^n, \lambda)}$$

Idea: stop when $\text{GLR}_n(\hat{i})$ is big for some answer $\hat{i}$. 
For any plausible answer $\hat{i} \in i^*(\hat{\mu}(n))$, the GLR$_n$ simplifies to

$$\text{GLR}_n(\hat{i}) = \inf_{\lambda: \hat{i} \notin i^*(\lambda)} \sum_{a=1}^{K} N_a(n) \text{KL}(\hat{\mu}_a(n), \lambda_a)$$

where $\text{KL}(x, y)$ is the Kullback-Leibler divergence in the exponential family.
GLR Stopping, Threshold

What is a suitable threshold for $\text{GLR}_n$ so that we do not make mistakes?
GLR Stopping, Threshold

What is a suitable threshold for $\text{GLR}_n$ so that we do not make mistakes?

A mistake is made when $\text{GLR}_n(\hat{i})$ is big while $\hat{i} \not\in i^*(\mu)$.
What is a suitable threshold for $\text{GLR}_n$ so that we do not make mistakes?

A mistake is made when $\text{GLR}_n(\hat{i})$ is big while $\hat{i} \notin i^*(\mu)$.

But then

$$\text{GLR}_n(\hat{i}) = \inf_{\lambda: \hat{i} \notin i^*(\lambda)} \sum_{a=1}^{K} N_a(n) \text{KL}(\hat{\mu}_a(n), \lambda_a) \leq \sum_{a=1}^{K} N_a(n) \text{KL}(\hat{\mu}_a(n), \mu_a).$$
GLR Stopping, Threshold

What is a suitable threshold for $\text{GLR}_n$ so that we do not make mistakes?

A mistake is made when $\text{GLR}_n(\hat{i})$ is big while $\hat{i} \not\in i^*(\mu)$.

But then

$$\text{GLR}_n(\hat{i}) = \inf_{\lambda : \hat{i} \not\in i^*(\lambda)} \sum_{a=1}^{K} N_a(n) \text{KL}(\hat{\mu}_a(n), \lambda_a) \leq \sum_{a=1}^{K} N_a(n) \text{KL}(\hat{\mu}_a(n), \mu_a).$$

Good anytime deviation inequalities exist for that upper bound.

**Theorem (Kaufmann and Koolen, 2018)**

$$\mathbb{P} \left( \exists n : \sum_{a=1}^{K} N_a(n) \text{KL}(\hat{\mu}_a(n), \mu_a) - \sum_{n} \ln \ln N_a(n) \geq C(K, \delta) \right) \leq \delta$$

for $C(K, \delta) \approx \ln \frac{1}{\delta} + K \ln \ln \frac{1}{\delta}$. 
• Tight criterion for stopping.
• We will see asymptotically matches lower bound.
• Often relatively easy to compute
• Typically reduces sample complexity by factor $\approx 2$ for BAI problems compared to confidence-interval based stopping (Why? Confidence region)
• Performs very well in practise
Track-and-Stop Algorithm Template
Intuition, going back at least to Lai and Robbins (1985)

A (spectacular) difference in behaviour must be due to a (spectacular) difference in the observations.

So being $\delta$-PAC on $\mu$ and also on $\lambda$ with $i^*(\mu) \neq i^*(\lambda)$ requires collecting enough discriminating information.
Intuition, going back at least to Lai and Robbins (1985)

A (spectacular) difference in behaviour **must** be due to a (spectacular) difference in the observations.

So being $\delta$-PAC on $\mu$ and also on $\lambda$ with $i^*(\mu) \neq i^*(\lambda)$ requires collecting enough discriminating information.

Define the *alternative* to $\mu$ by $\text{Alt}(\mu) := \{\text{bandit } \lambda | i^*(\lambda) \neq i^*(\mu)\}$. 
Intuition, going back at least to Lai and Robbins (1985)

A (spectacular) difference in behaviour must be due to a (spectacular) difference in the observations.

So being $\delta$-PAC on $\mu$ and also on $\lambda$ with $i^*(\mu) \neq i^*(\lambda)$ requires collecting enough discriminating information.

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

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

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

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

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

$$\frac{1}{T^*(\mu)} = \max_{w \in \Delta_K} \inf_{\lambda \in \text{Alt}(\mu)} \sum_{i=1}^{K} w_i KL(\mu_i, \lambda_i)$$
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 $E_{\mu}[\tau] \geq 601.2$. 
Recall sample complexity lower bound at bandit $\mu$ governed by

$$\max_{w \in \Delta_K} \inf_{\lambda \in \text{Alt}(\mu)} \sum_{i=1}^{K} w_i \text{KL}(\mu_i, \lambda_i)$$
Recall sample complexity lower bound at bandit $\mu$ governed by

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

Matching algorithms must sample with $\text{argmax}$ (oracle) proportions $w^*(\mu)$. 
Track-and-Stop scheme (Garivier and Kaufmann, 2016)

At each time step $t$

- compute plug-in oracle solution $w^*(\hat{\mu}_t)$
- sample arm $A_t$ to track (ensure $N_{a}(t)/t \rightarrow w_{a}^*(\hat{\mu}_t)$)
- force exploration to ensure $\hat{\mu}_t \rightarrow \mu$.

Stop using GLRT stopping rule, recommend single non-rejected arm.
Track-and-Stop scheme (Garivier and Kaufmann, 2016)

At each time step $t$

- compute plug-in **oracle solution** $w^*(\hat{\mu}_t)$
- sample arm $A_t$ to track (ensure $N_a(t)/t \to w^*_a(\hat{\mu}_t)$)
- force exploration to ensure $\hat{\mu}_t \to \mu$.

Stop using GLRT stopping rule, recommend single non-rejected arm.

**Theorem (Asymptotic Instance-Optimality)**

The sample complexity of Track-and-Stop for BAI is bounded by

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

**Analysis**

Convergence $\hat{\mu}_t \to \mu$ and **continuity** of $\mu \mapsto w^*(\mu)$ ensures sampling proportion $N_a(t)/t$ approximates oracle $w^*_a(\mu)$. 
Why interested in asymptotically optimal algorithms?

- State-of-the-art performance in practise (some problems)
  - Best Arm Identification
  - Minimax Game Tree Search
Why interested in asymptotically optimal algorithms?

- State-of-the-art performance in practise (some problems)
  - Best Arm Identification
  - Minimax Game Tree Search
- Different ("fresh") structure compared to other techniques (confidence intervals, elimination, Thompson sampling, ...
Why interested in asymptotically optimal algorithms?

- State-of-the-art performance in practise (some problems)
  - Best Arm Identification
  - Minimax Game Tree Search
- Different ("fresh") structure compared to other techniques (confidence intervals, elimination, Thompson sampling, ...)
- TaS reduces the identification problem to efficiently computing $w^*(\mu)$. 
TaS Conclusion

• “Track” instance-optimal sampling rule.
• Often relatively easy to compute
• Works very well for BAI. Not many experiments beyond.
• Performs very well in practise
Two Interesting Points
Q: Is $\mu \mapsto w^*(\mu)$ always continuous? No!
Q: Is $\mu \mapsto w^*(\mu)$ always continuous? No!

For single-answer problems, can escalate to set-valued mappings and upper hemi-continuity. Tracking requires care.
Q: Is $\mu \mapsto w^*(\mu)$ always continuous? No!

For single-answer problems, can escalate to set-valued mappings and upper hemi-continuity. Tracking requires care.

For multiple-answer problems (including $\epsilon$-BAI), continuity is unsalvageable.
Q: Is $\mu \mapsto w^*(\mu)$ always continuous? No!

For single-answer problems, can escalate to **set-valued mappings** and **upper hemi-continuity**. Tracking requires care.

For multiple-answer problems (including $\epsilon$-BAI), continuity is **unsalvageable**.

Contributions in (Degenne and Koolen, 2019)

- A lower-bound with multiple correct answers (now $\max \max \max \inf$).
- A new algorithm **Sticky Track-and-Stop** that asymptotically matches the lower bound.
- Explicit example where vanilla TaS fails (arcsine law)
Saddle Point Techniques

Standard technique: approximately solve saddle point problem

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

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.
• Compose regret bound, concentration and optimism to get finite-confidence guarantee.

Implementation available in tidnabbil library.

Analogue for regret in (Degenne, Shao, and Koolen, 2020)
Saddle Point Techniques

Standard technique: approximately solve saddle point problem

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

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.
• Compose regret bound, concentration and optimism to get finite-confidence guarantee.
Saddle Point Techniques

Standard technique: approximately solve saddle point problem

\[
\max_{w \in \Delta K} \inf_{\lambda \in \text{Alt}(\mu)} \sum_{i=1}^{K} w_i \text{KL}(\mu_i, \lambda_i)
\]

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.
- Compose regret bound, concentration and optimism to get finite-confidence guarantee.

Implementation available in tidnabbi library.

Analogue for regret in (Degenne, Shao, and Koolen, 2020).
Thanks!


