Bandit Algorithms for Pure Exploration: Best Arm Identification and Game Tree Search

Wouter M. Koolen

CWI
Centrum Wiskunde & Informatica

Machine Learning and Statistics for Structures
Friday 23rd February, 2018

Outline

Question

Complexity of interactive learning.

Question

Complexity of interactive learning.

- Medical testing [Villar et al., 2015]
- Online advertising and website optimisation [Zhou et al., 2014]
- Monte Carlo planning [Grill et al., 2016], and
- Game-playing AI [Silver et al., 2016]

Pure Exploration

Query: **Which is . . .**

- **the most effective drug dose?**
- **the most appealing website layout?**
- **the safest next robot action?**

Pure Exploration

Query: **Which is . . .**

- **the most effective drug dose?**
- **the most appealing website layout?**
- **the safest next robot action?**

Method

- statistical **experiments** in physical or simulated environment, **interactively** and **adaptively**.

Pure Exploration

Query: **Which is . . .**

- **the most effective drug dose?**
- **the most appealing website layout?**
- **the safest next robot action?**

Method

- statistical **experiments** in physical or simulated environment, **interactively** and **adaptively**.

Main scientific questions:

- **sample complexity** of interactive learning
  # experiments as function of query structure and environment
- Design of **efficient** pure exploration systems

Outline

Formal model

Environment (Multi-armed bandit model)

K distributions $\nu_1, \ldots, \nu_K$ with means $\mu_1, \ldots, \mu_K$.

Best arm

$$i^* = \operatorname*{argmax}_{i \in [K]} \mu_i$$

Formal model

Environment (Multi-armed bandit model)

K distributions $\nu_1, \ldots, \nu_K$ with means $\mu_1, \ldots, \mu_K$.
Best arm

$$i^* = \operatorname*{argmax}_{i \in [K]} \mu_i$$

Strategy

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

Formal model

Environment (Multi-armed bandit model)

K distributions $\nu_1, \ldots, \nu_K$ with means $\mu_1, \ldots, \mu_K$.
Best arm

$$i^* = \operatorname*{argmax}_{i \in [K]} \mu_i$$

Strategy

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

Realisation of interaction: $(I_1, X_1), \ldots, (I_\tau, X_\tau), \hat{I}$.

Formal model

Environment (Multi-armed bandit model)

K distributions $\nu_1, \ldots, \nu_K$ with means $\mu_1, \ldots, \mu_K$.
Best arm

$$i^* = \operatorname*{argmax}_{i \in [K]} \mu_i$$

Strategy

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

Realisation of interaction: $(I_1, X_1), \ldots, (I_\tau, X_\tau), \hat{I}$.

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

Objective

Two main flavours:

- **fixed budget** : fix $\tau = T$, optimise $\mathbb{P}(\hat{I} = i^*)$
- **fixed confidence** : fix $\mathbb{P}(\hat{I} = i^*) \leq \delta$, optimise $\mathbb{E}[\tau]$.

Objective

Two main flavours:

- **fixed budget** : fix $\tau = T$, optimise $\mathbb{P}(\hat{I} = i^*)$
- **fixed confidence** : fix $\mathbb{P}(\hat{I} = i^*) \leq \delta$, optimise $\mathbb{E}[\tau]$.

Notation

$$N_i(t) \;=\; \sum_{s=1}^{t} \mathbf{1}\{I_s = i\} \qquad \text{and} \qquad \hat{\mu}_i(t) \;=\; \frac{1}{N_i(t)} \sum_{s=1}^{t} X_s \mathbf{1}\{I_s = i\}$$

Outline

A statistician's view

Active Sequential Multiple Composite Hypothesis Testing

$$\mathcal{H}_i \;=\; \{\boldsymbol{\nu} \mid i^*(\boldsymbol{\nu}) = i\} \qquad i \in [K]$$

(Frequentist) uniform type 1 error control.

Families of approaches to BAI

- **Upper and Lower confidence bounds** [Bubeck et al., 2011, Kalyanakrishnan et al., 2012, Gabillon et al., 2012, Kaufmann and Kalyanakrishnan, 2013, Jamieson et al., 2014],
- **Racing or Successive Rejects/Eliminations** [Maron and Moore, 1997, Even-Dar et al., 2006, Audibert et al., 2010, Kaufmann and Kalyanakrishnan, 2013, Karnin et al., 2013],
- **Thompson Sampling** [Russo, 2016] (hemidemisemiBayesian)
- **Track-and-Stop** [Garivier and Kaufmann, 2016].

Outline

Change of Measure

If the observations are likely under both $\boldsymbol{\nu}$ and $\boldsymbol{\lambda}$, yet $i^*(\boldsymbol{\nu}) \neq i^*(\boldsymbol{\lambda})$, then the algorithm cannot stop and be correct.

Theorem (Kaufmann, Cappé, and Garivier [2016])

For bandit models $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$, stopping time τ, and event $\mathcal{E} \in \mathcal{F}_\tau$,

$$\sum_{i=1}^{K} \mathbb{E}_{\boldsymbol{\nu}}\left[N_i(\tau)\right] \mathrm{KL}(\nu_i \| \lambda_i) \;\geq\; d\left(\mathbb{P}_{\boldsymbol{\nu}}(\mathcal{E}), \mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E})\right)$$

Change of Measure

If the observations are likely under both $\boldsymbol{\nu}$ and $\boldsymbol{\lambda}$, yet $i^*(\boldsymbol{\nu}) \neq i^*(\boldsymbol{\lambda})$, then the algorithm cannot stop and be correct.

Theorem (Kaufmann, Cappé, and Garivier [2016])

For bandit models $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$, stopping time τ, and event $\mathcal{E} \in \mathcal{F}_\tau$,

$$\sum_{i=1}^{K} \mathbb{E}_{\boldsymbol{\nu}}[N_i(\tau)] \, \mathrm{KL}(\nu_i \| \lambda_i) \;\geq\; d\left(\mathbb{P}_{\boldsymbol{\nu}}(\mathcal{E}), \mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E})\right)$$

Theorem (Garivier and Kaufmann [2016])

Let $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ be bandit models with $i^*(\boldsymbol{\nu}) \neq i^*(\boldsymbol{\lambda})$. Then for any δ-correct algorithm

$$\sum_{i=1}^{K} \mathbb{E}_{\boldsymbol{\nu}}[N_i(\tau)] \, \mathrm{KL}(\nu_i \| \lambda_i) \;\geq\; d(\delta, 1-\delta)$$

Sample Complexity Consequence

Starting point:

$$\mathop{\mathbb{E}}_{\boldsymbol{\nu}}[\tau] \sum_{i=1}^{K} \frac{\mathbb{E}_{\boldsymbol{\nu}}\left[N_i(\tau)\right]}{\mathbb{E}_{\boldsymbol{\nu}}[\tau]} \mathrm{KL}(\nu_i \| \lambda_i) \;\geq\; d(\delta, 1-\delta)$$

hence

$$\mathop{\mathbb{E}}_{\boldsymbol{\nu}}[\tau] \max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\nu})} \sum_{i=1}^{K} w_i \, \mathrm{KL}(\nu_i \| \lambda_i) \;\geq\; d(\delta, 1-\delta)$$

so

Theorem (Garivier and Kaufmann [2016])

$$\mathop{\mathbb{E}}_{\boldsymbol{\nu}}[\tau] \;\geq\; T^*(\boldsymbol{\nu}) d(\delta, 1-\delta)$$

where

$$\frac{1}{T^*(\boldsymbol{\nu})} \;=\; \max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in Alt(\boldsymbol{\nu})} \sum_{i=1}^{K} w_i \, \mathrm{KL}(\nu_i \| \lambda_i)$$

Example

$K = 5$ arms, $\boldsymbol{\mu} = (.3, .4, .5, .6, .7)$.

Bernoulli

$$T^*(\boldsymbol{\mu}) \ = \ 203.4$$

Gaussian ($\sigma^2 = 1$)

$$T^*(\boldsymbol{\mu}) \ = \ 893.5$$

Outline

Algorithms

- Sampling rule ?
- Stopping rule ?
- Recommendation rule ?

$$\hat{I} = \underset{i \in [K]}{\text{argmax}} \ \hat{\mu}_i(\tau)$$

Sampling Rule

Look at the lower bound again. Any good algorithm must sample with optimal proportions

$$\boldsymbol{w}^*(\boldsymbol{\nu}) \;=\; \operatorname*{argmax}_{\boldsymbol{w}\in\triangle_K} \min_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\nu})} \sum_{i=1}^{K} w_i \, \mathrm{KL}(\nu_i \| \lambda_i)$$

Sampling Rule

Look at the lower bound again. Any good algorithm must sample with optimal proportions

$$\boldsymbol{w}^*(\boldsymbol{\nu}) \;=\; \underset{\boldsymbol{w}\in\triangle_K}{\operatorname{argmax}} \; \min_{\boldsymbol{\lambda}\in\mathrm{Alt}(\boldsymbol{\nu})} \; \sum_{i=1}^{K} w_i \, \mathrm{KL}(\nu_i \| \lambda_i)$$

Idea: draw $I_t \sim \boldsymbol{w}^*(\hat{\boldsymbol{\mu}}(t))$.

- Ensure $N_i(t)/t \to w_i^*$ by "forced exploration"
- Draw arm with $N_i(t)/t$ below w_i^* (tracking)
- Computation

Stopping Rule

When do we have enough evidence to stop?
Generalized Likelihood Ratio Test (GLRT)

$$Z_t \;=\; \ln \frac{P_{\hat{\boldsymbol{\mu}}(t)}(data)}{\max_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} P_{\boldsymbol{\lambda}}(data)}$$

Stopping Rule

When do we have enough evidence to stop?
Generalized Likelihood Ratio Test (GLRT)

$$Z_t \;=\; \ln \frac{P_{\hat{\boldsymbol{\mu}}(t)}(data)}{\max_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} P_{\boldsymbol{\lambda}}(data)}$$

Turns out, GLRT statistic equals

$$Z_t \;=\; \min_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \sum_{i=1}^{K} N_i(t) \, \mathrm{KL}(\hat{\mu}_i(t) \| \lambda_i)$$

i.e. lower bound with $\hat{\boldsymbol{\mu}}(t)$ plug-in.

Stopping Rule

When do we have enough evidence to stop?
Generalized Likelihood Ratio Test (GLRT)

$$Z_t \;=\; \ln \frac{P_{\hat{\boldsymbol{\mu}}(t)}(data)}{\max_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} P_{\boldsymbol{\lambda}}(data)}$$

Turns out, GLRT statistic equals

$$Z_t \;=\; \min_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \; \sum_{i=1}^{K} N_i(t) \, \mathrm{KL}(\hat{\mu}_i(t) \| \lambda_i)$$

i.e. lower bound with $\hat{\boldsymbol{\mu}}(t)$ plug-in.

Roughly: stop when $Z_t \geq \ln \frac{1}{\delta}$. Make precise with careful universal coding (MDL) argument.

All in all

Final result: for **Track-and-Stop** algorithms

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}{\ln \frac{1}{\delta}} \; = \; T^*(\boldsymbol{\mu})$$

Very similar optimality result for **Top Two Thompson Sampling** by Russo [2016]

Outline

Beyond asymptotic bounds

Okay, so good algorithms have

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

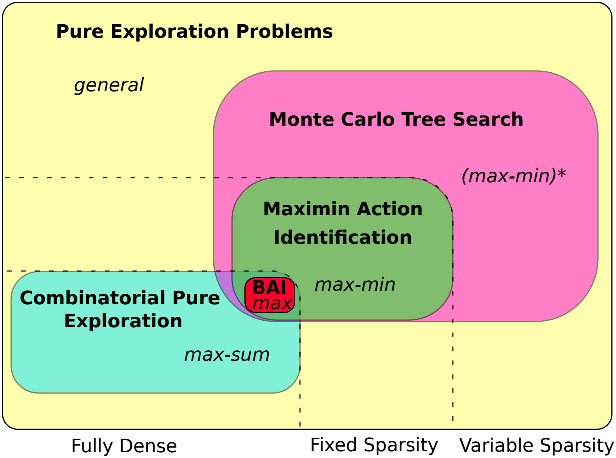
What about lower-order terms? "Moderate confidence" regime!

- Dependence on $\ln \ln \frac{1}{\delta}$
- Dependence on $\ln K$.

[Simchowitz et al., 2017, Chen et al., 2017b]

Beyond Best Arm

Practical and fundamental question: solving more complex pure exploration problems.



Combinatorial Pure Exploration

- Best k-set
- Shortest path
- Spanning tree
- . . .

Combinatorial Pure Exploration

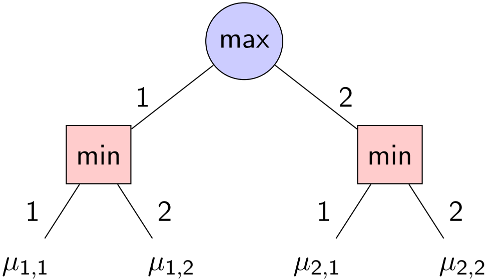
- Best k-set
- Shortest path
- Spanning tree
- ...

Combinatorial collection $\mathcal{F}$ of subsets of $[K]$.

$$i^* = i^*(\boldsymbol{\nu}) = \underset{S \in \mathcal{F}}{\operatorname{argmax}} \sum_{i \in S} \mu_i.$$

[Chen et al., 2017a]
Track-and-stop-like algorithms. Can compute lower bound. Dense.

Game Tree Search



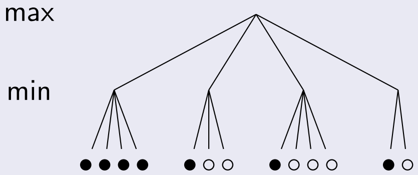
Goal: find **maximin action**

$$i^* := \arg\max_i \min_j \mu_{i,j}$$

Sparsity in the Lower Bound (depth 2)

Kaufmann and Koolen [2017]

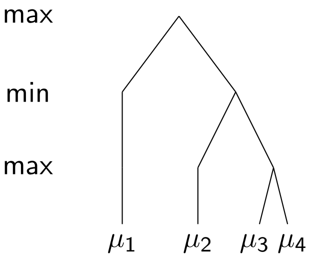
Sparsity Pattern

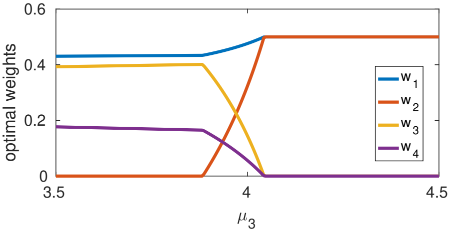


Oracle weights $\boldsymbol{w}^*$ supported on only 7 of the 13 leaves.

Open problem: algorithms incorporating appropriate pruning?

Sparsity in the Lower Bound (depth 3)





oracle weights $\boldsymbol{w}^* = (w_1, w_2, w_3, w_4)$ as a function of μ_3 for $\boldsymbol{\mu} = (16, 5, \mu_3, 0)$

It's complicated. But not intractable.

Conclusion

Pure Exploration is an interesting, hot, promising area.