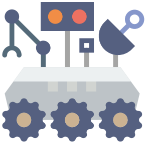
Pure Exploration Problems



Wouter M. Koolen
CWI and University of Twente

Download these slides from
`https://wouterkoolen.info/Talks/2022-07-15.pdf`!

- Pure Exploration:
  - PAC Learning
  - Best Arm Identification
  - Minimax Strategies in Noisy Games (Zero-Sum, Extensive Form)

Introduction and Motivation

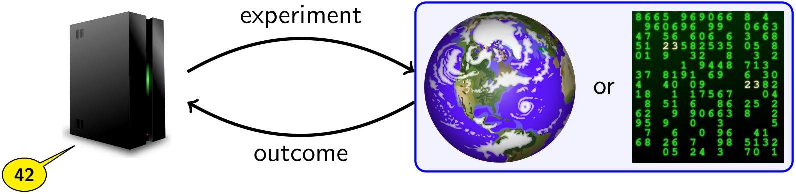
Grand Goal: Interactive Machine Learning

Query:

most effective drug dose?

most appealing website layout?

safest next robot action?



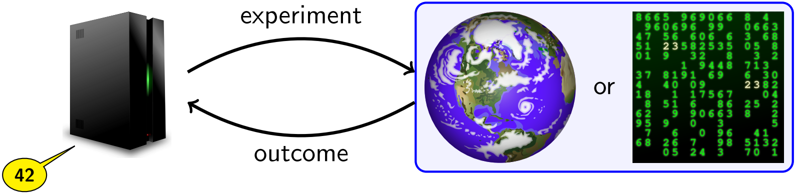
Grand Goal: Interactive Machine Learning

Query:

most effective drug dose?

most appealing website layout?

safest next robot action?



Main scientific questions

- *Efficient* systems
- **Sample complexity** as function of query and environment

Pure Exploration and Reinforcement Learning

Both are about *learning in uncertain environments*.

Pure Exploration and Reinforcement Learning

Both are about *learning in uncertain environments*.

Pure Exploration focuses on the statistical problem (**learn the truth**), while Reinforcement Learning focuses on behaviour (**maximise reward**).

Pure Exploration and Reinforcement Learning

Both are about *learning in uncertain environments*.

Pure Exploration focuses on the statistical problem (**learn the truth**), while Reinforcement Learning focuses on behaviour (**maximise reward**).

Pure Exploration occurs as **sub-module** in some RL algorithms (i.e. Phased Q-Learning by Even-Dar, Mannor, and Mansour, 2002)

Pure Exploration and Reinforcement Learning

Both are about *learning in uncertain environments*.

Pure Exploration focuses on the statistical problem (**learn the truth**), while Reinforcement Learning focuses on behaviour (**maximise reward**).

Pure Exploration occurs as **sub-module** in some RL algorithms (i.e. Phased Q-Learning by Even-Dar, Mannor, and Mansour, 2002)

Some problems approached with RL are in fact **better modelled** as pure exploration problems. Most notably MCTS for playing games.

Intuition with Cartoon Statistics

Thinking About One Arm

Consider T samples $X_1, X_2, \ldots, X_T$ drawn i.i.d. from a probability distribution with mean $\mu = \mathbb{E}[X]$. Let $\hat{\mu}_T = \frac{1}{T}\sum_{t=1}^{T} X_t$ be the empirical mean.

Hoeffding (bounded) or Chernoff (sub-Gaussian) give

Confidence Width At error probability δ,

$$\mathbb{P}\left\{\hat{\mu}_T - \mu \geq \sqrt{\frac{\ln \frac{1}{\delta}}{T}}\right\} \;\leq\; \delta$$

Exponential Error Decay If you take T samples, then

$$\mathbb{P}\left\{\hat{\mu}_T - \mu \geq \epsilon\right\} \;\leq\; e^{-T\epsilon^2}$$

Sample Complexity Seeing an effect of size ϵ at confidence δ requires

$$T \;=\; \frac{\ln \frac{1}{\delta}}{\epsilon^2} \quad \text{samples.}$$

More Than One Arm: Bandit Model

K-armed bandit:

- Arm k has mean μ_k.
- The best arm is $\arg\max_k \mu_k$
- The highest mean is $\mu_* = \max_k \mu_k$.
- The *gap* of arm k is $\Delta_k = \mu_* - \mu_k$.
- After t rounds, we write
  - $N_k(t)$ for the number of samples collected from arm k
  - $\hat{\mu}_k(t)$ for the average reward collected from arm k.

Maximising Reward

Task

Collect as much **reward** as possible. Hope: collect roughly $T\mu_*$.

A suboptimal arm k looks attractive only with probability

$$\mathbb{P}\left\{\hat{\mu}_k(t) \geq \mu_*\right\} \;=\; \mathbb{P}\left\{\hat{\mu}_k(t) - \mu_k \geq \Delta_k\right\} \leq e^{-N_k(t)\Delta_k^2}$$

To make this probability not important, it suffices to ensure

$$e^{-N_k(t)\Delta_k^2} = \frac{1}{t} \qquad \text{i.e.} \qquad N_k(t) = \frac{\ln t}{\Delta_k^2}$$

and the total regret thus incurred by time t is at most

$$\sum_{k \neq *} N_k(t)\Delta_k \;=\; \sum_{k \neq *} \frac{\ln t}{\Delta_k}$$

Fixed Budget

Task

Allocate T samples among K arms to **maximise the probability** of identifying the best arm, $\arg\max_k \mu_k$.

After t rounds

$$\begin{aligned}\mathbb{P}\left\{\hat{\mu}_k(t) \geq \hat{\mu}_*(t)\right\} &\leq \mathbb{P}\left\{\hat{\mu}_k(t) \geq (\mu_k + \mu_*)/2\right\} + \mathbb{P}\left\{(\mu_k + \mu_*)/2 \geq \hat{\mu}_*(t)\right\} \\ &\leq e^{-\frac{N_k}{4}\Delta_k^2} + e^{-\frac{N_*}{4}\Delta_k^2}\end{aligned}$$

Equalising these contributions to the error dictates taking

$$N_k \propto \frac{1}{\Delta_k^2} \qquad \text{i.e} \qquad N_k = \frac{T\frac{1}{\Delta_k^2}}{\sum_j \frac{1}{\Delta_j^2}}$$

and hence the overall error probability decays like

$$K \cdot \exp\left(-\frac{T}{4}\frac{1}{\sum_k \frac{1}{\Delta_k^2}}\right)$$

Fixed Confidence

Task

Take as **few samples** as possible to identify the best arm with error probability at most $\delta \in (0, 1)$.

Inverting the fixed budget bound

$$\delta \;=\; K \cdot \exp\left(-\frac{T}{4}\frac{1}{\sum_k \frac{1}{\Delta_k^2}}\right)$$

results in δ-correct identification after number of samples given by

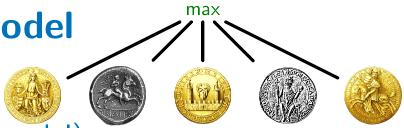
$$T \;=\; \frac{\ln \frac{K}{\delta}}{\frac{1}{4}\frac{1}{\sum_k \frac{1}{\Delta_k^2}}}.$$

A Stark Contrast

- For reward maximisation, sample each sub-optimal arm **logarithmically** often
$$N_k(t) \;=\; \frac{\ln t}{\Delta_k^2}$$
(**error decay of $1/t$** sufficient for exploitation)
- For best arm identification, sample each sub-optimal arm **linearly** often
$$N_k(t) \;=\; t \frac{\frac{1}{\Delta_k^2}}{\sum_j \frac{1}{\Delta_j^2}} \qquad \text{or} \qquad N_k(t) \;=\; \frac{4 \ln \frac{K}{\delta}}{\Delta_k^2}$$
(**error decays as e^{-t}** for pure exploration)

Best Arm Identification

Formal model



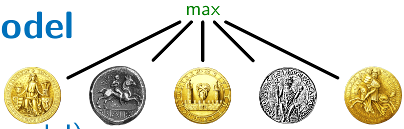
Environment (Multi-armed bandit model)

K distributions parameterised by their means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$.

The *best arm* is

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

Formal model



Environment (Multi-armed bandit model)

K distributions parameterised by their means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$.

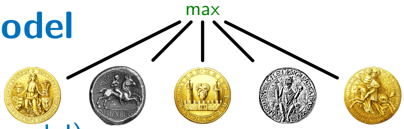
The *best arm* is

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

Strategy

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

Formal model



Environment (Multi-armed bandit model)

K distributions parameterised by their means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$.

The *best arm* is

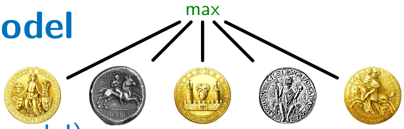
$$i^* = \operatorname*{argmax}_{i \in [K]} \mu_i$$

Strategy

- *Stopping rule* $\tau \in \mathbb{N}$
- In round $t \leq \tau$ *sampling rule* picks $I_t \in [K]$. See $X_t \sim \mu_{I_t}$.
- *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 parameterised by their means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$.

The *best arm* is

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

Strategy

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

On bandit $\boldsymbol{\mu}$, strategy $(\tau, (I_t)_t, \hat{I})$ has

- *error probability* $\mathbb{P}_{\boldsymbol{\mu}}\big(\hat{I} \neq i^*(\boldsymbol{\mu})\big)$, and
- *sample complexity* $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$.

Idea: constrain one, optimise the other.

Objective

On bandit $\boldsymbol{\mu}$, strategy $(\tau, (I_t)_t, \hat{I})$ has

- *error probability* $\mathbb{P}_{\boldsymbol{\mu}}\big(\hat{I} \neq i^*(\boldsymbol{\mu})\big)$, and
- *sample complexity* $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$.

Idea: constrain one, optimise the other.

Definition

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

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

(Generalisation: output ϵ-best arm)

Objective

On bandit $\boldsymbol{\mu}$, strategy $(\tau, (I_t)_t, \hat{I})$ has

- *error probability* $\mathbb{P}_{\boldsymbol{\mu}}\left(\hat{I} \neq i^*(\boldsymbol{\mu})\right)$, and
- *sample complexity* $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$.

Idea: constrain one, optimise the other.

Definition

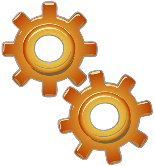
Fix small confidence $\delta \in (0, 1)$. A strategy is δ-*correct* (aka δ-*PAC*) if

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\hat{I} \neq i^*(\boldsymbol{\mu})\right) \leq \delta \qquad \text{for every bandit model } \boldsymbol{\mu}.$$

(Generalisation: output ϵ-best arm)

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

Algorithms

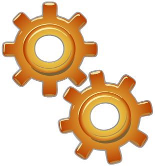


- Sampling rule I_t?
- Stopping rule τ?
- Recommendation rule $\hat{I}$?

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

where $\hat{\boldsymbol{\mu}}(t)$ is *empirical mean*.

Algorithms



- Sampling rule I_t?
- Stopping rule τ?
- Recommendation rule $\hat{I}$?

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

where $\hat{\boldsymbol{\mu}}(t)$ is *empirical mean*.

Approach: start investigating **lower bounds**

Instance-Dependent Sample Complexity Lower bound

Define the *alternatives* to $\boldsymbol{\mu}$ by $\mathsf{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} | i^*(\boldsymbol{\lambda}) \neq i^*(\boldsymbol{\mu})\}$.

Instance-Dependent Sample Complexity Lower bound

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

Theorem (Castro 2014; Garivier and Kaufmann 2016)

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

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

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

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

Instance-Dependent Sample Complexity Lower bound

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

Theorem (Castro 2014; Garivier and Kaufmann 2016)

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

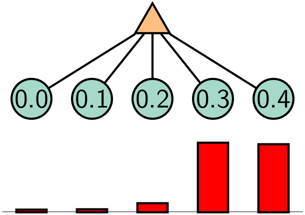
$$\mathop{\mathbb{E}}_{\boldsymbol{\mu}}[\tau] \;\geq\; T^*(\boldsymbol{\mu}) \ln \frac{1}{\delta}$$

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

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

Intuition (going back 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.

Example



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

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

At $\delta = 0.05$, the time gets multiplied by $\ln \frac{1}{\delta} = 3.0$.

Sampling Rule

Look at the lower bound again. Any good algorithm **must** sample with optimal (*oracle*) proportions

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

Sampling Rule

Look at the lower bound again. Any good algorithm **must** sample with optimal (*oracle*) proportions

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

Track-and-Stop

Idea: draw $I_t \sim \boldsymbol{w}^*(\hat{\boldsymbol{\mu}}(t-1))$.

- Ensure $\hat{\boldsymbol{\mu}}(t) \to \boldsymbol{\mu}$ hence $N_i(t)/t \to w_i^*$ by "forced exploration"
- Draw arm with $N_i(t)/t$ below w_i^* (tracking)
- Computation of $\boldsymbol{w}^*$ (reduction to 1d line search)

Stopping

When can we stop?

Stopping

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

Stopping

When can we stop?
When can we stop and give answer $\hat{\imath}$?
There is no plausible bandit model $\boldsymbol{\lambda}$ on which $\hat{\imath}$ is wrong.

Stopping

When can we stop?
When can we stop and give answer $\hat{\imath}$?
There is no plausible bandit model $\boldsymbol{\lambda}$ on which $\hat{\imath}$ is wrong.

Definition

Generalized Likelihood Ratio (GLR) measure of evidence

$$\mathrm{GLR}_n(\hat{\imath}) \;:=\; \ln \frac{\sup_{\boldsymbol{\mu}:\hat{\imath}\in i^*(\boldsymbol{\mu})} P\left(X^n | A^n, \boldsymbol{\mu}\right)}{\sup_{\boldsymbol{\lambda}:\hat{\imath}\notin i^*(\boldsymbol{\lambda})} P\left(X^n | A^n, \boldsymbol{\lambda}\right)}$$

Stopping

When can we stop?
When can we stop and give answer $\hat{\imath}$?
There is no plausible bandit model $\boldsymbol{\lambda}$ on which $\hat{\imath}$ is wrong.

Definition

Generalized Likelihood Ratio (GLR) measure of evidence

$$\mathrm{GLR}_n(\hat{\imath}) \;:=\; \ln \frac{\sup_{\boldsymbol{\mu}:\hat{\imath}\in i^*(\boldsymbol{\mu})} P\left(X^n|A^n, \boldsymbol{\mu}\right)}{\sup_{\boldsymbol{\lambda}:\hat{\imath}\notin i^*(\boldsymbol{\lambda})} P\left(X^n|A^n, \boldsymbol{\lambda}\right)}$$

Idea: stop when $\mathrm{GLR}_n(\hat{\imath})$ is big for some answer $\hat{\imath}$.

GLR Stopping

For any plausible answer $\hat{\imath} \in i^*(\hat{\boldsymbol{\mu}}(n))$, the GLR_n simplifies to

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

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

GLR Stopping, Treshold

What is a suitable threshold for GLR_n so that we do not make mistakes?

GLR Stopping, Treshold

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

GLR Stopping, Treshold

What is a suitable threshold for GLR_n so that we do not make mistakes?
A mistake is made when $\mathrm{GLR}_n(\hat{\imath})$ is big while $\hat{\imath} \notin i^*(\boldsymbol{\mu})$.
But then

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

GLR Stopping, Treshold

What is a suitable threshold for GLR_n so that we do not make mistakes?
A mistake is made when $\mathrm{GLR}_n(\hat{\imath})$ is big while $\hat{\imath} \notin i^*(\boldsymbol{\mu})$.
But then

$$\mathrm{GLR}_n(\hat{\imath}) \;=\; \inf_{\boldsymbol{\lambda}:\hat{\imath}\notin i^*(\boldsymbol{\lambda})} \sum_{a=1}^{K} N_a(n)\,\mathrm{KL}(\hat{\mu}_a(n), \lambda_a) \;\leq\; \boxed{\sum_{a=1}^{K} N_a(n)\,\mathrm{KL}(\hat{\mu}_a(n), \mu_a)}.$$

Good **anytime deviation inequalities** exist for that upper bound.

Theorem (Kaufmann and Koolen, 2018)

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

for $C(K,\delta) \approx \ln\frac{1}{\delta} + K\ln\ln\frac{1}{\delta}$.

All in all

Final result: lower and upper bound meet on **every problem instance**.

Theorem (Garivier and Kaufmann 2016)

For the Track-and-Stop algorithm, for any bandit $\boldsymbol{\mu}$

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

All in all

Final result: lower and upper bound meet on **every problem instance**.

Theorem (Garivier and Kaufmann 2016)

For the Track-and-Stop algorithm, for any bandit $\boldsymbol{\mu}$

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

Very similar optimality result for *Top Two Thompson Sampling* by Russo (2016). Here $N_i(t)/t \to w_i^*$ result of posterior sampling.

Problem Variations and Algorithms

Variations

- Prior knowledge about $\boldsymbol{\mu}$
  - Shape constraints: linear, convex, unimodal, etc. bandits
  - Non-parametric (and heavy-tailed) reward distributions (Agrawal, Koolen, and Juneja, 2021)
  - ...
- Questions beyond Best Arm
  - A/B/n testing (Russac et al., 2021)
  - Robust best arm (part 2 today)
  - Thresholding
  - Best VaR, CVaR and other tail risk measures (Agrawal, Koolen, and Juneja, 2021)
  - ...
- Multiple correct answers
  - ϵ-best arm
  - In general (Degenne and Koolen, 2019)
    (Requires a change in lower bound and upper bound)

Lazy Iterative Optimisation of w^*

Instead of computing at every round the plug-in oracle weights

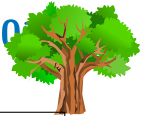
$$\boldsymbol{w}^*(\hat{\boldsymbol{\mu}}) = \underset{\boldsymbol{w} \in \triangle_K}{\operatorname{argmax}} \underbrace{\min_{\boldsymbol{\lambda} \in \operatorname{Alt}(\hat{\boldsymbol{\mu}})} \sum_{i=1}^{K} w_i \operatorname{KL}(\hat{\mu}_i \| \lambda_i)}_{\text{concave in } \boldsymbol{w}}$$

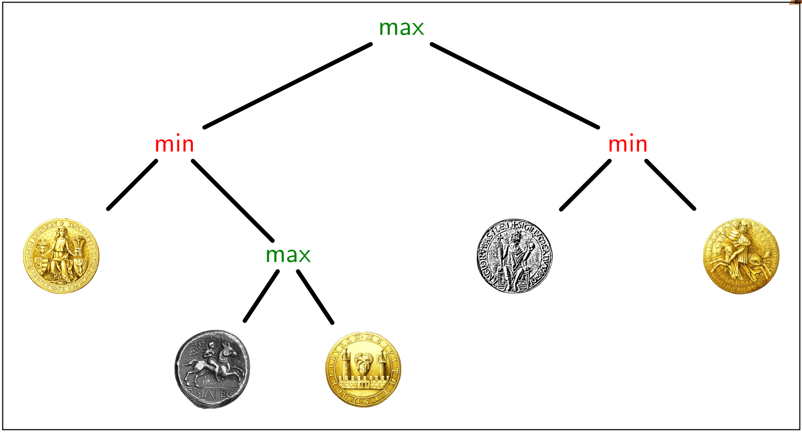
We may work as follows

- The inner problem is concave in $\boldsymbol{w}$.
- It can be maximised iteratively, i.e. with gradient descent.
- We may **interleave** **sampling** and **gradient** steps.
- A **single** gradient step per sample is enough (Degenne, Koolen, and Ménard, 2019)

Minimax Action Identification

Model (Teraoka, Hatano, and Takimoto, 20

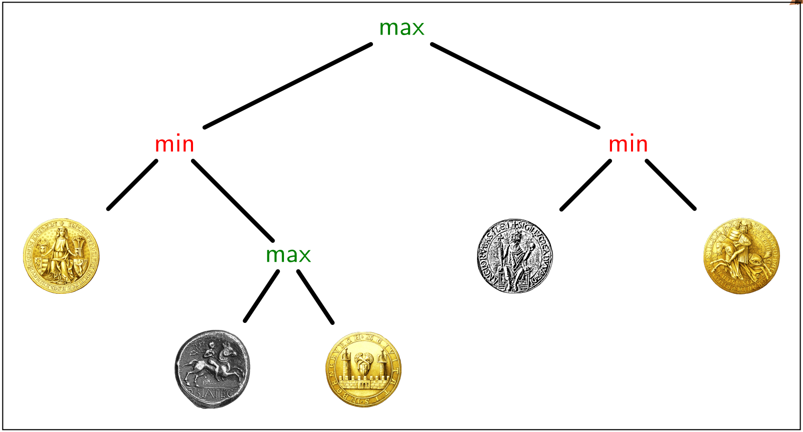




Maximin Action Identification Problem

Find **best move at root** from samples of **leaves**.

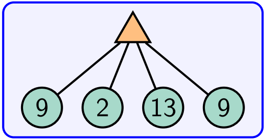
Model (Teraoka, Hatano, and Takimoto, 20



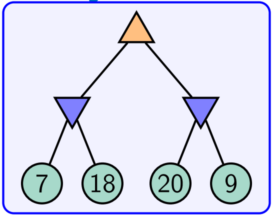
Maximin Action Identification Problem

Find **best move at root** from samples of **leaves**.

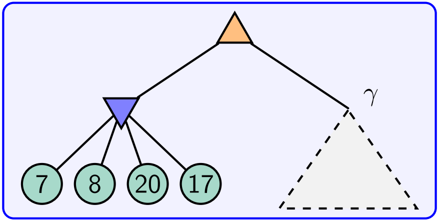
My Brief History



Best Arm Identification
(Garivier and Kaufmann, 2016)
Solved, continuous



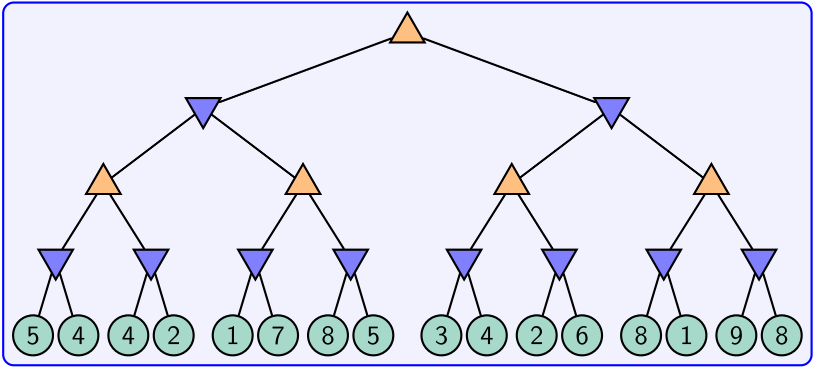
Depth 2 Game
(Garivier, Kaufmann, and Koolen, 2016)
Open, continuous?



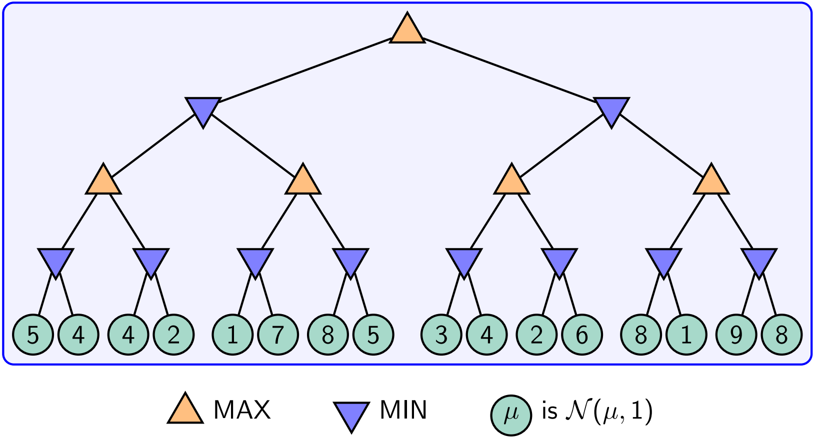
Depth 1.5 Game
(Kaufmann, Koolen, and Garivier, 2018)
Solved, discontinuous

What we are able to solve today

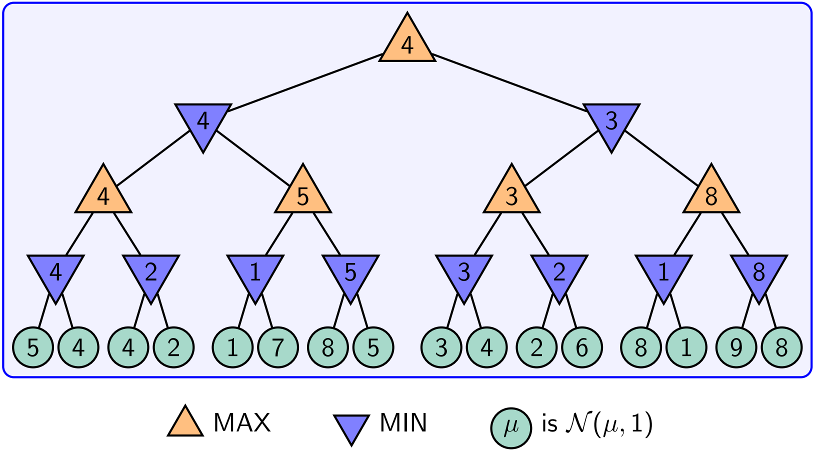
Noisy games of any depth



Example Backward Induction Computation



Example Backward Induction Computation



Model

Definition

A **game tree** is a min-max tree with leaves $\mathcal{L}$. A **bandit model** $\boldsymbol{\mu}$ assigns a distribution μ_ℓ to each leaf $\ell \in \mathcal{L}$.

Model

Definition

A **game tree** is a min-max tree with leaves $\mathcal{L}$. A **bandit model** $\boldsymbol{\mu}$ assigns a distribution μ_ℓ to each leaf $\ell \in \mathcal{L}$.

The maximin action (best action at the root) is

$$i^*(\boldsymbol{\mu}) \;:=\; \operatorname*{argmax}_{a_1} \min_{a_2} \max_{a_3} \min_{a_4} \cdots \; \mu_{a_1 a_2 a_3 a_4 \ldots}$$

Model

Definition

A **game tree** is a min-max tree with leaves $\mathcal{L}$. A **bandit model** $\boldsymbol{\mu}$ assigns a distribution μ_ℓ to each leaf $\ell \in \mathcal{L}$.

The maximin action (best action at the root) is

$$i^*(\boldsymbol{\mu}) \;:=\; \underset{a_1}{\operatorname{argmax}}\, \min_{a_2} \max_{a_3} \min_{a_4} \cdots \; \mu_{a_1 a_2 a_3 a_4 \dots}$$

Protocol

For $t = 1, 2, \dots, \tau$:

- Learner picks a leaf $L_t \in \mathcal{L}$.
- Learner sees $X_t \sim \mu_{L_t}$

Learner recommends action $\hat{I}$

Model

Definition

A **game tree** is a min-max tree with leaves $\mathcal{L}$. A **bandit model** $\boldsymbol{\mu}$ assigns a distribution μ_ℓ to each leaf $\ell \in \mathcal{L}$.

The maximin action (best action at the root) is

$$i^*(\boldsymbol{\mu}) \;\coloneqq\; \operatorname*{argmax}_{a_1} \min_{a_2} \max_{a_3} \min_{a_4} \cdots \; \mu_{a_1 a_2 a_3 a_4 \ldots}$$

Protocol

For $t = 1, 2, \ldots, \tau$:

- Learner picks a leaf $L_t \in \mathcal{L}$.
- Learner sees $X_t \sim \mu_{L_t}$

Learner recommends action $\hat{I}$

Learner is δ-PAC if

$$\forall \boldsymbol{\mu} : \mathbb{P}_{\boldsymbol{\mu}}\left(\tau < \infty \wedge \hat{I} \neq i^*(\boldsymbol{\mu})\right) \leq \delta$$

Main Theorem I: Lower Bound

Define the *alternatives* to $\boldsymbol{\mu}$ by $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} | i^*(\boldsymbol{\lambda}) \neq i^*(\boldsymbol{\mu})\}$.
NB here i^* is **best action at the root**

Main Theorem I: Lower Bound

Define the *alternatives* to $\boldsymbol{\mu}$ by $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} | i^*(\boldsymbol{\lambda}) \neq i^*(\boldsymbol{\mu})\}$.
NB here i^* is **best action at the root**

Theorem (Castro 2014; Garivier and Kaufmann 2016)

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

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

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

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

Main Theorem II: Algorithm

Idea is still to consider the oracle weight map

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

and track the plug-in estimate: $L_t \sim \boldsymbol{w}^*(\hat{\boldsymbol{\mu}}(t-1))$.

Main Theorem II: Algorithm

Idea is still to consider the oracle weight map

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

and track the plug-in estimate: $L_t \sim \boldsymbol{w}^*(\hat{\boldsymbol{\mu}}(t-1))$.

But what about **continuity**? Does $\hat{\boldsymbol{\mu}}(t) \to \boldsymbol{\mu}$ imply $\boldsymbol{w}^*(\boldsymbol{\mu}(t)) \to \boldsymbol{w}^*(\boldsymbol{\mu})$?

Main Theorem II: Algorithm

Idea is still to consider the oracle weight map

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

and track the plug-in estimate: $L_t \sim \boldsymbol{w}^*(\hat{\boldsymbol{\mu}}(t-1))$.

But what about **continuity**? Does $\hat{\boldsymbol{\mu}}(t) \to \boldsymbol{\mu}$ imply $\boldsymbol{w}^*(\boldsymbol{\mu}(t)) \to \boldsymbol{w}^*(\boldsymbol{\mu})$?

But $\boldsymbol{w}^*$ is **not continuous**. Even at depth "1.5" with 2 arms.

Main Theorem II: Algorithm

Idea is still to consider the oracle weight map

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

and track the plug-in estimate: $L_t \sim \boldsymbol{w}^*(\hat{\boldsymbol{\mu}}(t-1))$.

But what about **continuity**? Does $\hat{\boldsymbol{\mu}}(t) \to \boldsymbol{\mu}$ imply $\boldsymbol{w}^*(\boldsymbol{\mu}(t)) \to \boldsymbol{w}^*(\boldsymbol{\mu})$?

But $\boldsymbol{w}^*$ is **not continuous**. Even at depth "1.5" with 2 arms.

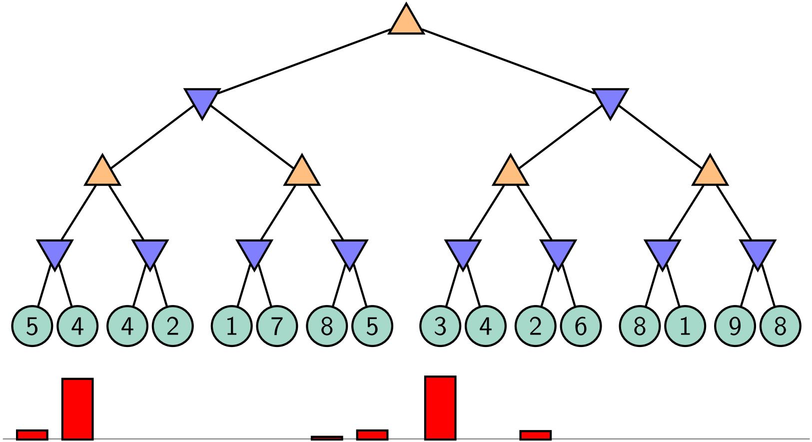
Theorem (Degenne and Koolen, 2019)

Take set-valued interpretation of argmax *defining* $\boldsymbol{w}^*$*. Then* $\boldsymbol{\mu} \mapsto \boldsymbol{w}^*(\boldsymbol{\mu})$ *is upper-hemicontinuous and convex-valued. Suitable tracking ensures that as* $\hat{\boldsymbol{\mu}}(t) \to \boldsymbol{\mu}$*, any* $\boldsymbol{w}_t \in \boldsymbol{w}^*(\hat{\boldsymbol{\mu}}(t-1))$ *have*

$$\min_{\boldsymbol{w} \in \boldsymbol{w}^*(\boldsymbol{\mu})} \left\| \boldsymbol{w}_t - \boldsymbol{w} \right\|_\infty \to 0$$

Track-and-Stop is asymptotically optimal.

Example



On Computation

To compute a gradient (in $\boldsymbol{w}$) we need to differentiate

$$\boldsymbol{w} \;\mapsto\; \min_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \; \sum_{i=1}^{K} w_i \, \mathrm{KL}(\mu_i \| \lambda_i)$$

An optimal $\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})$ can be found by binary search for common value plus tree reasoning in $O(|\mathcal{L}|)$.

Conclusion

Conclusion

We saw

- Cartoon statistics to sharpen intuition
- Lower bounds for instance-optimal best-arm identification
- Algorithms for instance-optimal best-arm identification
- Extensions to game trees

Thank you!

Agrawal, S., W. M. Koolen, and S. Juneja (Dec. 2021). “Optimal Best-Arm Identification Methods for Tail-Risk Measures”. In: *Advances in Neural Information Processing Systems (NeurIPS) 34*. Accepted.

Castro, R. M. (Nov. 2014). “Adaptive sensing performance lower bounds for sparse signal detection and support estimation”. In: *Bernoulli* 20.4, pp. 2217–2246.

Degenne, R. and W. M. Koolen (Dec. 2019). “Pure Exploration with Multiple Correct Answers”. In: *Advances in Neural Information Processing Systems (NeurIPS) 32*, pp. 14591–14600.

Degenne, R., W. M. Koolen, and P. Ménard (Dec. 2019). “Non-Asymptotic Pure Exploration by Solving Games”. In: *Advances in Neural Information Processing Systems (NeurIPS) 32*, pp. 14492–14501.

Even-Dar, E., S. Mannor, and Y. Mansour (2002). “PAC Bounds for Multi-armed Bandit and Markov Decision Processes”. In: *Computational Learning Theory, 15th Annual Conference on Computational Learning Theory, COLT 2002, Sydney, Australia, July 8-10, 2002, Proceedings*. Vol. 2375. Lecture Notes in Computer Science, pp. 255–270.

- Garivier, A. and E. Kaufmann (2016). "Optimal Best arm Identification with Fixed Confidence". In: *Proceedings of the 29th Conference On Learning Theory (COLT)*.
- Garivier, A., E. Kaufmann, and W. M. Koolen (June 2016). "Maximin Action Identification: A New Bandit Framework for Games". In: *Proceedings of the 29th Annual Conference on Learning Theory (COLT)*.
- Kaufmann, E. and W. M. Koolen (Oct. 2018). "Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals". Preprint.
- Kaufmann, E., W. M. Koolen, and A. Garivier (Dec. 2018). "Sequential Test for the Lowest Mean: From Thompson to Murphy Sampling". In: *Advances in Neural Information Processing Systems (NeurIPS) 31*, pp. 6333–6343.
- Russac, Y., C. Katsimerou, D. Bohle, O. Cappé, A. Garivier, and W. M. Koolen (Dec. 2021). "A/B/n Testing with Control in the Presence of Subpopulations". In: *Advances in Neural Information Processing Systems (NeurIPS) 34*. Accepted.

Russo, D. (2016). "Simple Bayesian Algorithms for Best Arm Identification". In: *CoRR* abs/1602.08448.

Teraoka, K., K. Hatano, and E. Takimoto (2014). "Efficient Sampling Method for Monte Carlo Tree Search Problem". In: *IEICE Transactions on Infomation and Systems*, pp. 392–398.