Introduction

What is Pure Exploration?
Relation to Reinforcement Learning?
Why care?
  ▶ Pure Exploration problems occur as sub-problems in RL!
  ▶ Novel/interesting/powerful PE learning algorithms! “Fresh”
My focus: Pure Exploration Renaissance (2016)
  1. Track-and-Stop algorithm for Best Arm Identification
  2. BAI-MCTS approach for Game Tree Search
  3. Murphy Sampling for Games Trees of “depth 1.5”
1 Introduction

2 Relation of RL and PE

3 Pure Exploration Intro: Best Arm Identification
   - Model
   - Sample Complexity Lower Bound
   - Algorithms

4 Game Tree Search
   - Game Trees of Arbitrary Depth
   - Confidence Intervals on Min/Max
   - Game Trees of Depth 1.5 (Maximum/Minimum)
     - Results

5 Conclusion
Relation of RL and PE

Reinforcement Learning
- MDP
- Exploration/Exploitation
- State

Relation of RL and PE
- Regret minimisation
  - MAB
  - Exploration/Exploitation
  - No State

Pure Exploration
- MAB + Structured Query
- Exploration only
- Structure

Best Arm Identification
- MAB
- Exploration only
- No State

Koolen (CWI)
Example 1: Phased Q-Learning

[Even-Dar, Mannor, and Mansour, 2002]

Early example(s) of Pure Exploration as sub-module in RL

Initialise $V_0(s) := 0$ for each state $s$.

for phase $i = 1, 2, \ldots$ do

  for each state $s$ do

    Run **Best Arm Identification** algorithm on

    $$a \mapsto r + \gamma V_i(s') \quad \text{where} \quad (r, s') \sim \mathbb{P}(r, s'|s, a)$$

    Store estimate in $V_{i+1}(s)$.

  end for

end for

Example 2: AlphaZero

MCTS as Policy/Value Improvement Operator

Randomise network parameters $\theta_1$.

\begin{itemize}
  \item for iteration $i = 1, 2, \ldots$ do
    \item for training games $j = 1, 2, \ldots$ do
      \item $\pi_t$ result of MCTS from $s_t$ with network $\theta_i$
      \item Store $(s_1, \pi_1, z), \ldots, (s_T, \pi_T, z)$.
  \end{itemize}

end for

Train network $\theta_{i+1}$ on stored data.

end for
Environment (Multi-armed bandit model)

$K$ distributions parameterised by their means $\mu = (\mu_1, \ldots, \mu_K)$. The best arm is

$$i^* = \arg\max_{i \in [K]} \mu_i$$
Formal model

Environment (Multi-armed bandit model)

$K$ distributions parameterised by their means $\mu = (\mu_1, \ldots, \mu_K)$. The best arm is

$$i^* = \arg\max_{i \in [K]} \mu_i$$

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** $\hat{i} \in [K]$. 
Formal model

Environment (Multi-armed bandit model)

$K$ distributions parameterised by their means $\mu = (\mu_1, \ldots, \mu_K)$. The best arm is

$$i^* = \arg\max_{i \in [K]} \mu_i$$

Strategy

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

Realisation of interaction: $(l_1, X_1), \ldots, (l_\tau, X_\tau), \hat{i}$. 
Formal model

Environment (Multi-armed bandit model)

\( K \) distributions parameterised by their means \( \mu = (\mu_1, \ldots, \mu_K) \). The best arm is

\[
i^* = \arg\max_{i \in [K]} \mu_i
\]

Strategy

- **Stopping rule** \( \tau \in \mathbb{N} \)
- In round \( t \leq \tau \) sampling rule picks \( l_t \in [K] \). See \( X_t \sim \mu_{l_t} \).
- **Recommendation rule** \( \hat{i} \in [K] \).

Realisation of interaction: \( (l_1, X_1), \ldots, (l_\tau, X_\tau), \hat{i} \).

Two objectives: sample efficiency \( \tau \) and correctness \( \hat{i} = i^* \).
Objective

On bandit $\mu$, strategy $(\tau, (l_t)_t, \hat{i})$ has

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

Idea: constrain one, optimise the other.
Objective

On bandit $\mu$, strategy $(\tau, (l_t)_t, \hat{l})$ has

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

Idea: constrain one, optimise the other.

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 $\mu$.

(Generalisation: output $\epsilon$-best arm)
Objective

On bandit \( \mu \), strategy \((\tau, (I_t)_t, \hat{i})\) has

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

Idea: constrain one, optimise the other.

Definition

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

\[
\mathbb{P}_\mu(\hat{i} \neq i^*(\mu)) \leq \delta \quad \text{for every bandit model } \mu.
\]

(Generalisation: output \( \epsilon \)-best arm)

Goal: minimise \( \mathbb{E}_\mu[\tau] \) over all \( \delta \)-correct strategies.
Algorithms

- Sampling rule $l_t$?
- Stopping rule $\tau$?
- Recommendation rule $\hat{l}$?

\[ \hat{l} = \arg\max_{i \in [K]} \hat{\mu}_i(\tau) \]

where $\hat{\mu}(t)$ is empirical mean.
Instance-Dependent Sample Complexity Lower Bound

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

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

*Fix a $\delta$-correct strategy. Then for every bandit model $\mu$*

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

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

$$
\frac{1}{T^*(\mu)} = \max_{w \in \Delta_K} \min_{\lambda \in \text{Alt}(\mu)} \sum_{i=1}^{K} w_i \text{KL}(\mu_i \| \lambda_i).
$$
Instance-Dependent Sample Complexity Lower bound

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

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

*Fix a $\delta$-correct strategy. Then for every bandit model $\mu$*

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

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

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

Intuition (going back 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.
Example

$K = 5$ arms, Bernoulli $\mu = (0, 0.1, 0.2, 0.3, 0.4)$.

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

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

\[ w^*(\mu) = \arg\max_{w \in \Delta_K} \min_{\lambda \in \text{Alt}(\mu)} \sum_{i=1}^{K} w_i \text{KL}(\mu_i || \lambda_i) \]
Sampling Rule

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

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

Track-and-Stop

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

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

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

**Theorem (Garivier and Kaufmann 2016)**

For Track-and-Stop algorithm, for any bandit $\mu$

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

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

**Theorem (Garivier and Kaufmann 2016)**

For Track-and-Stop algorithm, for any bandit $\mu$

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

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

2. Relation of RL and PE

3. Pure Exploration Intro: Best Arm Identification
   - Model
   - Sample Complexity Lower Bound
   - Algorithms

4. Game Tree Search
   - Game Trees of Arbitrary Depth
   - Confidence Intervals on Min/Max
   - Game Trees of Depth 1.5 (Maximum/Minimum)
     - Results

5. Conclusion
Challenge Environment: Stochastic Game Tree Search

Problem
Determine the optimal move at the root
From sample access to the leaf payoffs

\[ \mu \text{ is } \mathcal{N}(\mu, 1) \]
Challenge Environment: Stochastic Game Tree Search

Problem

Determine the **optimal move** at the root
From *sample access* to the leaf payoffs
Maximin Action Identification Problem

Find **best move at root** from samples of **leaves**.
Maximin Action Identification Problem

Find best move at root from samples of leaves.

Model [Teraoka et al., 2014]
Methods for Best Arm Identification

\[ \text{max} \]
Methods for Best Arm Identification

\[ \text{max} \]

LUCB, UGapE, ...
Methods for Best Arm Identification

max

LUCB, UGapE,
Reduction of MCTS to BAI
Reduction of MCTS to BAI
Reduction of MCTS to BAI
Reduction of MCTS to BAI
Reduction of MCTS to BAI

\[ \text{max} \quad \text{min} \quad \text{max} \quad \text{min} \]
Reduction of MCTS to BAI
Correctness

\((\epsilon, \delta)\)-PAC algorithms

Efficiency

Sample complexity function of leaf gaps \(\Delta_\ell\)

\[
O \left( \sum_{\ell \in L} \frac{1}{\Delta_\ell^2 \vee \epsilon^2 \log \left( \frac{1}{\delta} \right)} \right)
\]
How to build confidence intervals on min/max nodes

More principled approach in [Kaufmann, Koolen, and Garivier, 2018].

Children  Max node
How to build confidence intervals on min/max nodes

More principled approach in [Kaufmann, Koolen, and Garivier, 2018]. Many equal intervals ⇒ higher lower bound.
1. Introduction

2. Relation of RL and PE

3. Pure Exploration Intro: Best Arm Identification
   - Model
   - Sample Complexity Lower Bound
   - Algorithms

4. Game Tree Search
   - Game Trees of Arbitrary Depth
   - Confidence Intervals on Min/Max
   - Game Trees of Depth 1.5 (Maximum/Minimum)
     - Results

5. Conclusion
Simplify

Best Arm Identification
[Garivier and Kaufmann, 2016]  
\textbf{Solved}

Depth 2
[Garivier, Kaufmann, and Koolen, 2016]  
\textbf{Open}
Simple Instance: Minimum Threshold Identification

Fix threshold $\gamma$.

$$\mu^* := \min_i \mu_i \leq \gamma?$$

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

- Pick leaf $A_t$
- See $X_t \sim \mu_{A_t}$

Recommend $\hat{m} \in \{<, >\}$

Goal: fixed confidence $\mathbb{P}_\mu \{\text{error}\} < \delta$
and small sample complexity $\mathbb{E}_\mu[\tau]$
Generic lower bound [Castro, 2014, Garivier and Kaufmann, 2016] shows sample complexity for any $\delta$-correct algorithm is at least

$$\mathbb{E}_\mu[\tau] \geq T^*(\mu) \ln \frac{1}{\delta}.$$
Generic lower bound [Castro, 2014, Garivier and Kaufmann, 2016] shows sample complexity for any $\delta$-correct algorithm is at least

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

For our problem the characteristic time and oracle weights are

$$T^*(\mu) = \begin{cases} \frac{1}{\text{KL}(\mu^*, \gamma)} & \mu^* < \gamma, \\ \sum_a \frac{1}{\text{KL}(\mu_a, \gamma)} & \mu^* > \gamma, \end{cases}$$

$$w^*_a(\mu) = \begin{cases} 1_{a=a^*} & \mu^* < \gamma, \\ \frac{1}{\text{KL}(\mu_a, \gamma)} \sum_j \frac{1}{\text{KL}(\mu_j, \gamma)} & \mu^* > \gamma. \end{cases}$$
Dichotomous Oracle Behaviour! Sampling Rule?

\[ \mu \cdot \gamma \]
**Sampling Rules**

- **Lower Confidence Bounds**
  Play $A_t = \text{arg min}_a \ LCB_a(t)$

- **Thompson Sampling** (π_{t-1} is posterior after t − 1 rounds)
  Sample $\theta \sim \Pi_{t-1}$, then play $A_t = \text{arg min}_a \ \theta_a$.

- **Murphy Sampling**  **condition on low minimum mean**
  Sample $\theta \sim \Pi_{t-1} (\cdot | \text{min}_a \theta_a < \gamma)$, then play $A_t = \text{arg min}_a \ \theta_a$.  

*new*
Intuition for Murphy Sampling

- When $\mu^* < \gamma$ conditioning is immaterial: $\theta \approx \mu$ and MS \equiv TS.

- When $\mu^* > \gamma$ conditioning results in $\theta \approx (\mu_1, \ldots, \gamma, \ldots, \mu_K)$. Index $a$ lowered to $\gamma$ with probability $\propto \frac{1}{\text{KL}(\mu_a, \gamma)}$ [Russo, 2016].
Murphy Sampling Rule [KKG, NIPS’18]

Theorem

Asymptotic optimality: \( N_a(t)/t \to w^*_a(\mu) \) for all \( \mu \)

<table>
<thead>
<tr>
<th>Sampling rule</th>
<th>&lt;</th>
<th>&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thompson Sampling</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Lower Confidence Bounds</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Murphy Sampling</strong></td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Lemma

Any anytime sampling strategy \( (A_t)_t \) ensuring \( \frac{N_t}{t} \to w^*(\mu) \) and good stopping rule \( \tau_\delta \) guarantee

\[
\limsup_{\delta \to 0} \frac{\tau_\delta}{\ln \frac{T^*}{\delta}} \leq T^*(\mu).
\]
Conclusion

- Pure Exploration currently going through a renaissance
- Instance-optimal identification algorithms
  - Best Arm
  - Game Tree Search
  - …
- Moving toward more complex queries. RL on the horizon …
- Useful submodules
Conclusion

- Pure Exploration currently going through a renaissance
- Instance-optimal identification algorithms
  - Best Arm
  - Game Tree Search
  - …
- Moving toward more complex queries. RL on the horizon . . .
- Useful submodules

Thank you! And let’s talk!