In-Context Learning for Pure Exploration

Alessio Russo[1], Ryan Welch[2], Aldo Pacchiano[1,3]



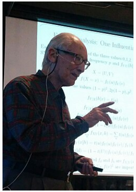
October 2025

[1] Boston University. [2] Stanford University. [3] Broad Institute of MIT and Harvard.

Overview i

Introduction

What is Pure Exploration? [Chernoff, 1959]



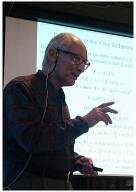
SEQUENTIAL DESIGN OF EXPERIMENTS

By Herman Chernoff[1]

Stanford University

1. Introduction. Considerable scientific research is characterized as follows. The scientist is interested in studying a phenomenon. At first he is quite ignorant and his initial experiments are preliminary and tentative. As he gathers relevant data, he becomes more definite in his impression of the underlying theory. This more definite impression is used to construct more informative experiments. Finally after a certain point he is satisfied that his evidence is sufficient to allow him to announce certain conclusions and he does so.

What is Pure Exploration? [Chernoff, 1959]



SEQUENTIAL DESIGN OF EXPERIMENTS

BY HERMAN CHERNOFF[1]

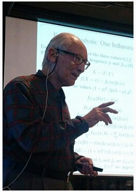
Stanford University

1. Introduction. Considerable scientific research is characterized as follows. The scientist is interested in studying a phenomenon. At first he is quite ignorant and his initial experiments are preliminary and tentative. As he gathers relevant data, he becomes more definite in his impression of the underlying theory. This more definite impression is used to construct more informative experiments. Finally after a certain point he is satisfied that his evidence is sufficient to allow him to announce certain conclusions and he does so.

Pure Exploration is the Machine Learning term **for what a statistician would call active sequential hypothesis testing**[1] [Naghshvar and Javidi, 2013].

[1] Wouter Koolen, https://homepages.cwi.nl/~wmkoolen/PureExploration18/#why.

What is Pure Exploration? [Chernoff, 1959]



SEQUENTIAL DESIGN OF EXPERIMENTS

By Herman Chernoff[1]

Stanford University

1. Introduction. Considerable scientific research is characterized as follows. The scientist is interested in studying a phenomenon. At first he is quite ignorant and his initial experiments are preliminary and tentative. As he gathers relevant data, he becomes more definite in his impression of the underlying theory. This more definite impression is used to construct more informative experiments. Finally after a certain point he is satisfied that his evidence is sufficient to allow him to announce certain conclusions and he does so.

Pure exploration is about inferring an underlying **true hypothesis**. How **should you allocate experiments** based on what you can **infer from their outcomes**?

Pure Exploration vs Regret Minimization

Probably you are familiar with the "**Exploration/Exploitation**" trade-off in Reinforcement Learning (RL) [Lai and Robbins, 1985].

Pure Exploration vs Regret Minimization

Probably you are familiar with the "**Exploration/Exploitation**" trade-off in Reinforcement Learning (RL) [Lai and Robbins, 1985].

Exploration vs Exploitation: accumulate reward by choosing good actions!
Yet, to know that an action is good, you need to explore...

Pure Exploration vs Regret Minimization

Probably you are familiar with the "**Exploration/Exploitation**" trade-off in Reinforcement Learning (RL) [Lai and Robbins, 1985].

Exploration vs Exploitation: accumulate reward by choosing good actions 😄!
Yet, to know that an action is good, you need to explore... 😟

But...what is the difference between *Pure Exploration* and *Regret Minimization*?

Pure Exploration vs Regret Minimization

Probably you are familiar with the "**Exploration/Exploitation**" trade-off in Reinforcement Learning (RL) [Lai and Robbins, 1985].

> **Exploration vs Exploitation**: accumulate reward by choosing good actions !
>
> Yet, to know that an action is good, you need to explore...

But...what is the difference between *Pure Exploration* and *Regret Minimization*?

- Pure Exploration is the **rebellious counter-movement**: "pure" refers to how fast it learns, with no regard for how much it earns.

Motivation

For what problems is it useful?

Motivation

For what problems is it useful?

1. Minimizing the number of DNA-based tests performed to accurately detect cancer [Gan et al., 2021] .

Motivation

For what problems is it useful?

1. Minimizing the number of DNA-based tests performed to accurately detect cancer [Gan et al., 2021] .
2. Improving recommendations in recommender systems [Resnick and Varian, 1997] .

Motivation

For what problems is it useful?

1. Minimizing the number of DNA-based tests performed to accurately detect cancer [Gan et al., 2021] .
2. Improving recommendations in recommender systems [Resnick and Varian, 1997] 

.
3. Quickly identifying a faulty sensor [Hero and Cochran, 2011] .

Motivation

For what problems is it useful?

1. Minimizing the number of DNA-based tests performed to accurately detect cancer [Gan et al., 2021] .
2. Improving recommendations in recommender systems [Resnick and Varian, 1997] .
3. Quickly identifying a faulty sensor [Hero and Cochran, 2011] .
4. Active search on the celestial sphere .

Motivation

For what problems is it useful?

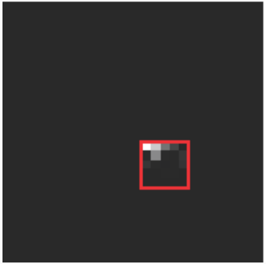
1. Minimizing the number of DNA-based tests performed to accurately detect cancer [Gan et al., 2021].
2. Improving recommendations in recommender systems [Resnick and Varian, 1997].
3. Quickly identifying a faulty sensor [Hero and Cochran, 2011].
4. Active search on the celestial sphere.

Let's check some examples more in detail.

Example: Digit Detection

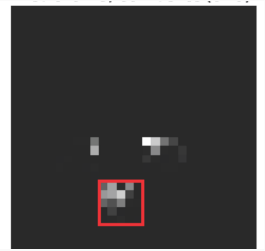
Classify the image using the least number of pixel patches.

Example: Digit Detection

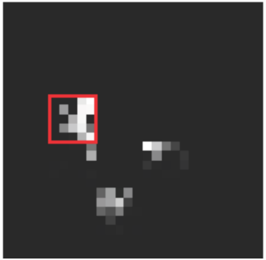


Example: Digit Detection

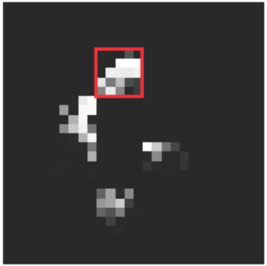
Example: Digit Detection



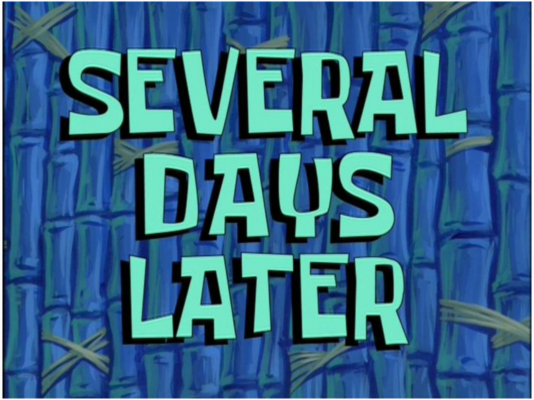
Example: Digit Detection



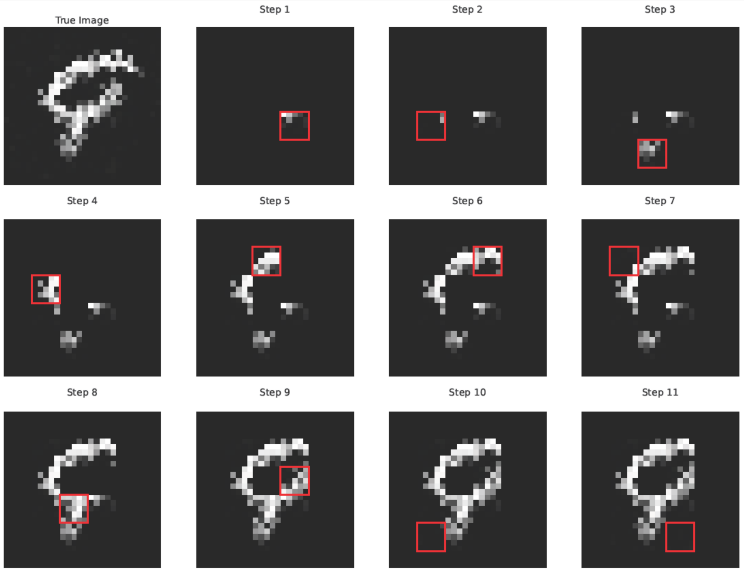
Example: Digit Detection



Example: Digit Detection



Example: Digit Detection



Example: Best Arm Identification (BAI) [Audibert and Bubeck, 2010]

Consider a Multi-Armed Bandit problem with K arms [Lattimore and Szepesvári, 2020]:



- You can select one arm a at the time and observe a random reward with mean value μ_a.
- The optimal arm is $a^\star = \arg\max_a \mu_a$.
- How do we quickly identify the optimal arm $a^\star$ with some confidence level $\delta \in (0, 1)$?

Example: Best Arm Identification (BAI) [Audibert and Bubeck, 2010]

Consider a Multi-Armed Bandit problem with K arms [Lattimore and Szepesvári, 2020]:

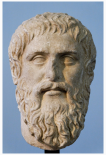
- You can select one arm a at the time and observe a random reward with mean value μ_a.
- The optimal arm is $a^\star = \arg\max_a \mu_a$.
- How do we quickly identify the optimal arm $a^\star$ with some confidence level $\delta \in (0, 1)$?

Example: Best Arm Identification (BAI) [Audibert and Bubeck, 2010]

Consider a Multi-Armed Bandit problem with K arms [Lattimore and Szepesvári, 2020]:

- You can select one arm a at the time and observe a random reward with mean value μ_a.
- The optimal arm is $a^\star = \arg\max_a \mu_a$.
- How do we quickly identify the optimal arm $a^\star$ with some confidence level $\delta \in (0, 1)$?

Some Problems and Limitations



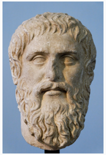
How do we compute adaptive exploration strategies?

Adaptive Exploration

A truly intelligent agent should tailor exploration to the difficulty of the problem; treating all problems the same is not a sign of intelligent behavior. (**self cit**.)

$\Longrightarrow$ The solution should adapt to the problem at hand.

Some Problems and Limitations



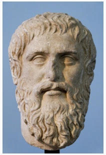
How do we compute adaptive exploration strategies?

Adaptive Exploration

A truly intelligent agent should tailor exploration to the difficulty of the problem; treating all problems the same is not a sign of intelligent behavior. (**self cit**.)

$\Longrightarrow$ The solution should adapt to the problem at hand.

Some Problems and Limitations



How do we compute adaptive exploration strategies?

Adaptive Exploration

A truly intelligent agent should tailor exploration to the difficulty of the problem; treating all problems the same is not a sign of intelligent behavior. (**self cit**.)

$\Longrightarrow$ The solution should adapt to the problem at hand.

Some Problems and Limitations

It's **hard** to find optimal solutions to pure exploration problems.

How does the solution look like? Informally, the solution is characterized by this problem:

$$\underbrace{\sup_{\pi}}_{\text{Exploration policy}} \overbrace{\inf_{P' \in \mathrm{Alt}(P)}}^{\text{Confusing model}} \mathbb{E}_{x\sim\rho^\pi}[\mathrm{KL}(P(x), P'(x))]$$

- ρ^π is the data distribution induced by π and P is the true data model.
- P' is a *confusing* model: different from P, but optimized so that it's statistically similar when collecting data with π.

$\Rightarrow$ Find π that maximizes the collected evidence!

Some Problems and Limitations

It's **hard** to find optimal solutions to pure exploration problems.

How does the solution look like? Informally, the solution is characterized by this problem:

$$\underbrace{\sup_{\pi}}_{\text{Exploration policy}} \overbrace{\inf_{P' \in \mathrm{Alt}(P)}}^{\text{Confusing model}} \mathbb{E}_{x \sim \rho^\pi}[\mathrm{KL}(P(x), P'(x))]$$

- ρ^π is the data distribution induced by π and P is the true data model.
- P' is a *confusing* model: different from P, but optimized so that it's statistically similar when collecting data with π.

$\Rightarrow$ Find π that maximizes the collected evidence!

Some Problems and Limitations

It's **hard** to find optimal solutions to pure exploration problems.

How does the solution look like? Informally, the solution is characterized by this problem:

$$\underbrace{\sup_{\pi}}_{\text{Exploration policy}} \overbrace{\inf_{P' \in \mathrm{Alt}(P)}}^{\text{Confusing model}} \mathbb{E}_{x \sim \rho^{\pi}}[\mathrm{KL}(P(x), P'(x))]$$

- ρ^{π} is the data distribution induced by π and P is the true data model.
- P' is a *confusing* model: different from P, but optimized so that it's statistically similar when collecting data with π.

$\Rightarrow$ Find π that maximizes the collected evidence!

Some Problems and Limitations

It's **hard** to find optimal solutions to pure exploration problems.

How does the solution look like? Informally, the solution is characterized by this problem:

$$\underbrace{\sup_{\pi}}_{\text{Exploration policy}} \overbrace{\inf_{P' \in \mathrm{Alt}(P)}}^{\text{Confusing model}} \mathbb{E}_{x\sim\rho^{\pi}}[\mathrm{KL}(P(x), P'(x))]$$

- ρ^{π} is the data distribution induced by π and P is the true data model.
- P' is a *confusing* model: different from P, but optimized so that it's statistically similar when collecting data with π.

$\Rightarrow$ Find π that maximizes the collected evidence!

Some Problems and Limitations

It's **hard** to find optimal solutions to pure exploration problems.

How does the solution look like? Informally, the solution is characterized by this problem:

$$\underbrace{\sup_{\pi}}_{\text{Exploration policy}} \overbrace{\inf_{P' \in \mathrm{Alt}(P)}}^{\text{Confusing model}} \mathbb{E}_{x \sim \rho^{\pi}}[\mathrm{KL}(P(x), P'(x))]$$

- ρ^{π} is the data distribution induced by π and P is the true data model.
- P' is a *confusing* model: different from P, but optimized so that it's statistically similar when collecting data with π.

$\Rightarrow$ Find π that maximizes the collected evidence!

Some Problems and Limitations

It's **hard** to find optimal solutions to pure exploration problems.

How does the solution look like? Informally, the solution is characterized by this problem:

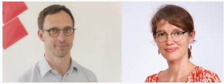
$$\underbrace{\sup_{\pi}}_{\text{Exploration policy}} \overbrace{\inf_{P' \in \mathrm{Alt}(P)}}^{\text{Confusing model}} \mathbb{E}_{x \sim \rho^{\pi}}[\mathrm{KL}(P(x), P'(x))]$$

- ρ^{π} is the data distribution induced by π and P is the true data model.
- P' is a *confusing* model: different from P, but optimized so that it's statistically similar when collecting data with π.

$\Rightarrow$ Find π that maximizes the collected evidence!

Some Problems and Limitations

- Recent advances in BAI showed how to characterize the optimal exploration strategy in simple i.i.d. Bandit models [Garivier and Kaufmann, 2016]

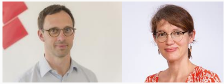


- Most results are limited to tabular problems, and extending to more complex settings is difficult.
  - For tabular Markov Decision Processes (MDPs) the optimal exploration strategy is characterized by a non-convex problem [Al Marjani et al., 2021].
  - In some cases it is not possible to identify the underlying true hypothesis [Russo et al., 2025].

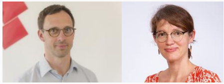
Some Problems and Limitations

- Recent advances in BAI showed how to characterize the optimal exploration strategy in simple i.i.d. Bandit models [Garivier and Kaufmann, 2016]  
- Most results are limited to tabular problems, and extending to more complex settings is difficult.
  - For tabular Markov Decision Processes (MDPs) the optimal exploration strategy is characterized by a non-convex problem [Al Marjani et al., 2021].
  - In some cases it is not possible to identify the underlying true hypothesis [Russo et al., 2025].

Some Problems and Limitations

- Recent advances in BAI showed how to characterize the optimal exploration strategy in simple i.i.d. Bandit models [Garivier and Kaufmann, 2016] 
- Most results are limited to tabular problems, and extending to more complex settings is difficult.
  - For tabular Markov Decision Processes (MDPs) the optimal exploration strategy is characterized by a non-convex problem [Al Marjani et al., 2021].
  - In some cases it is not possible to identify the underlying true hypothesis [Russo et al., 2025].

Some Problems and Limitations

- Recent advances in BAI showed how to characterize the optimal exploration strategy in simple i.i.d. Bandit models [Garivier and Kaufmann, 2016] 
- Most results are limited to tabular problems, and extending to more complex settings is difficult.
  - For tabular Markov Decision Processes (MDPs) the optimal exploration strategy is characterized by a non-convex problem [Al Marjani et al., 2021].
  - In some cases it is not possible to identify the underlying true hypothesis [Russo et al., 2025].

Our Solution: ICPE

Can we design a simple, but general, method that learns how to solve pure exploration problems efficiently?

Our Solution: ICPE

> Can we design a simple, but general, method that learns how to solve pure exploration problems efficiently?

In-Context Learning for Pure Exploration

Alessio Russo* Boston University arusso2@bu.edu

Ryan Welch*,† Stanford University rcwelch@stanford.edu

Aldo Pacchiano Boston University, Broad Institute of MIT and Harvard pacchian@bu.edu

Abstract

We study the problem *active sequential hypothesis testing*, also known as *pure exploration*: given a new task, the learner *adaptively collects data* from the environment to efficiently determine an underlying correct hypothesis. A classical instance of this problem is the task of identifying the best arm in a multi-armed bandit problem (a.k.a. BAI, Best-Arm Identification), where actions index hypotheses. Another important case is generalized search, a problem of determining the correct label through a sequence of strategically selected queries that indirectly reveal information about the label. In this work, we introduce *In-Context Pure Exploration* (`ICPE`), which meta-trains Transformers to map *observation histories* to *query actions* and a *predicted hypothesis*, yielding a model that transfers in-context. At inference time, `ICPE` actively gathers evidence on new tasks and infers the true hypothesis without parameter updates. Across deterministic, stochastic, and structured benchmarks, including BAI and generalized search, `ICPE` is competitive with adaptive baselines while requiring no explicit modeling of information structure. Our results support Transformers as practical architectures for *general sequential testing*.

Can we design a simple, but general, method that learns how to solve pure exploration problems efficiently?

In-Context Learning for Pure Exploration

Alessio Russo* Boston University arusso2@bu.edu
Ryan Welch*,† Stanford University rcwelch@stanford.edu
Aldo Pacchiano Boston University, Broad Institute of MIT and Harvard pacchian@bu.edu

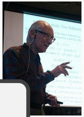
Abstract

We study the problem *active sequential hypothesis testing*, also known as *pure exploration*: given a new task, the learner *adaptively collects data* from the environment to efficiently determine an underlying correct hypothesis. A classical instance of this problem is the task of identifying the best arm in a multi-armed bandit problem (a.k.a. BAI, Best-Arm Identification), where actions index hypotheses. Another important case is generalized search, a problem of determining the correct label through a sequence of strategically selected queries that indirectly reveal information about the label. In this work, we introduce *In-Context Pure Exploration* (ICPE), which meta-trains Transformers to map *observation histories* to *query actions* and a *predicted hypothesis*, yielding a model that transfers in-context. At inference time, ICPE actively gathers evidence on new tasks and infers the true hypothesis without parameter updates. Across deterministic, stochastic, and structured benchmarks, including BAI and generalized search, ICPE is competitive with adaptive baselines while requiring no explicit modeling of information structure. Our results support Transformers as practical architectures for *general sequential testing*.

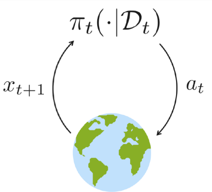
- In-Context Pure Explorer (ICPE) is a Transformer-based architecture meta-trained on a family of tasks to learn an exploration policy.
- ICPE is a model that transfers in-context: at inference time, gathers evidence on new tasks and infers the true hypothesis without parameter updates .

ICPE: Modeling

ICPE: Intro



Pure exploration is about inferring an underlying **true hypothesis**. How **should you allocate experiments** based on what you can **infer from their outcomes**?

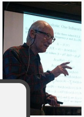


An agent π interacts with the environment and collects data.

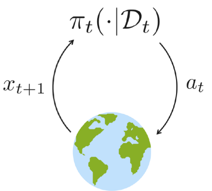
It's a **sequential problem**. In each round t you observe x_t and choose an experiment (action) a_t. We model it:

1. Define the hypothesis space $\mathcal{H}$.
2. Define the query/action/experiment space $\mathcal{A}$ (i.e., the ways you can interact).
3. Define the observation space $\mathcal{X}$ (what you observe after an interaction).
4. Define the dynamics P of the environment: $P(x_{t+1}|x_1, a_1, \ldots, x_t, a_t)$ (i.e., how does the environment react after selecting a_t? how does it depend on past interactions?)

ICPE: Intro



> Pure exploration is about inferring an underlying **true hypothesis**. How **should you allocate experiments** based on what you can **infer from their outcomes**?

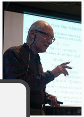


An agent π interacts with the environment and collects data.

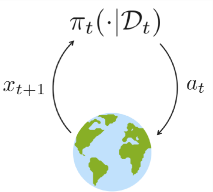
It's a **sequential problem**. In each round t you observe x_t and choose an experiment (action) a_t. We model it:

1. Define the hypothesis space $\mathcal{H}$.
2. Define the query/action/experiment space $\mathcal{A}$ (i.e., the ways you can interact).
3. Define the observation space $\mathcal{X}$ (what you observe after an interaction).
4. Define the dynamics P of the environment: $P(x_{t+1}|x_1, a_1, \ldots, x_t, a_t)$ (i.e., how does the environment react after selecting a_t? how does it depend on past interactions?)

ICPE: Intro



Pure exploration is about inferring an underlying **true hypothesis**. How **should you allocate experiments** based on what you can **infer from their outcomes**?

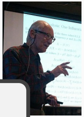


An agent π interacts with the environment and collects data.

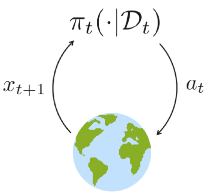
It's a **sequential problem**. In each round t you observe x_t and choose an experiment (action) a_t. We model it:

1. Define the hypothesis space $\mathcal{H}$.
2. Define the query/action/experiment space $\mathcal{A}$ (i.e., the ways you can interact).
3. Define the observation space $\mathcal{X}$ (what you observe after an interaction).
4. Define the dynamics P of the environment: $P(x_{t+1}|x_1, a_1, \ldots, x_t, a_t)$ (i.e., how does the environment react after selecting a_t? how does it depend on past interactions?)

ICPE: Intro



Pure exploration is about inferring an underlying **true hypothesis**. How **should you allocate experiments** based on what you can **infer from their outcomes**?

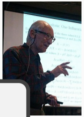


An agent π interacts with the environment and collects data.

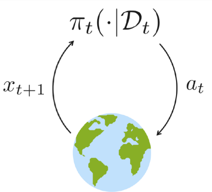
It's a **sequential problem**. In each round t you observe x_t and choose an experiment (action) a_t. We model it:

1. Define the hypothesis space $\mathcal{H}$.
2. Define the query/action/experiment space $\mathcal{A}$ (i.e., the ways you can interact).
3. Define the observation space $\mathcal{X}$ (what you observe after an interaction).
4. Define the dynamics P of the environment: $P(x_{t+1}|x_1, a_1, \ldots, x_t, a_t)$ (i.e., how does the environment react after selecting a_t? how does it depend on past interactions?)

ICPE: Intro



> Pure exploration is about inferring an underlying **true hypothesis**. How **should you allocate experiments** based on what you can **infer from their outcomes**?



An agent π interacts with the environment and collects data.

It's a **sequential problem**. In each round t you observe x_t and choose an experiment (action) a_t. We model it:

1. Define the hypothesis space $\mathcal{H}$.
2. Define the query/action/experiment space $\mathcal{A}$ (i.e., the ways you can interact).
3. Define the observation space $\mathcal{X}$ (what you observe after an interaction).
4. Define the dynamics P of the environment: $P(x_{t+1}|x_1, a_1, \ldots, x_t, a_t)$ (i.e., how does the environment react after selecting a_t? how does it depend on past interactions?)

ICPE: Intro

What about the true hypothesis?

$\Rightarrow$ we assume there exists a functional $h^\star$ mapping $P \mapsto \mathcal{H}$, dynamics to hypotheses. We set $H^\star = h^\star(P)$.

Definition (Environment)

We define an environment to be $M = (\mathcal{X}, \mathcal{A}, P, H^\star)$.

What about the true hypothesis?

$\Rightarrow$ we assume there exists a functional $h^\star$ mapping $P \mapsto \mathcal{H}$, dynamics to hypotheses. We set $H^\star = h^\star(P)$.

Definition (Environment)

We define an environment to be $M = (\mathcal{X}, \mathcal{A}, P, H^\star)$.

What about the true hypothesis?

$\Rightarrow$ we assume there exists a functional $h^\star$ mapping $P \mapsto \mathcal{H}$, dynamics to hypotheses. We set $H^\star = h^\star(P)$.

Definition (Environment)

We define an environment to be $M = (\mathcal{X}, \mathcal{A}, P, H^\star)$.

ICPE: Modeling Example

How do we model the Best Arm Identification problem?




- $H^\star = \arg\max_a \mu_a$ (arm with highest mean reward).
- $P(\cdot|a)$ is the reward distribution when you select action $a \in \{1, \ldots, K\}$.
- Upon selecting action a you observe $x \sim P(\cdot|a)$, the reward.

How do we model the Best Arm Identification problem?





- $H^\star = \arg\max_a \mu_a$ (arm with highest mean reward).
- $P(\cdot|a)$ is the reward distribution when you select action $a \in \{1, \dots, K\}$.
- Upon selecting action a you observe $x \sim P(\cdot|a)$, the reward.

ICPE: Modeling Example

How do we model the Best Arm Identification problem?


- $H^\star = \arg\max_a \mu_a$ (arm with highest mean reward).
- $P(\cdot|a)$ is the reward distribution when you select action $a \in \{1, \ldots, K\}$.
- Upon selecting action a you observe $x \sim P(\cdot|a)$, the reward.

How do we model the Best Arm Identification problem?


- $H^\star = \arg\max_a \mu_a$ (arm with highest mean reward).
- $P(\cdot|a)$ is the reward distribution when you select action $a \in \{1, \ldots, K\}$.
- Upon selecting action a you observe $x \sim P(\cdot|a)$, the reward.

ICPE: Prior and Objectives

Last ingredient is the assumption that we have a prior $\mathcal{P}$ over a set of environments $\mathcal{M}$, representing our belief of M.

:

1. Fixed Budget: maximize evidence over a horizon.
2. Fixed Confidence: quickly collect evidence.

ICPE: Prior and Objectives

Last ingredient is the assumption that we have a prior $\mathcal{P}$ over a set of environments $\mathcal{M}$, representing our belief of M.

What should be the objective? 

:

1. Fixed Budget: maximize evidence over a horizon.
2. Fixed Confidence: quickly collect evidence.

ICPE: Prior and Objectives

Last ingredient is the assumption that we have a prior $\mathcal{P}$ over a set of environments $\mathcal{M}$, representing our belief of M.

What should be the objective?

SEQUENTIAL DESIGN OF EXPERIMENTS

BY HERMAN CHERNOFF[1]

Stanford University

1. Introduction. Considerable scientific research is characterized as follows. The scientist is interested in studying a phenomenon. At first he is quite ignorant and his initial experiments are preliminary and tentative. As he gathers relevant data, he becomes more definite in his impression of the underlying theory. This more definite impression is used to construct more informative experiments. Finally after a certain point he is satisfied that his evidence is sufficient to allow him to announce certain conclusions and he does so.

:

1. Fixed Budget: maximize evidence over a horizon.
2. Fixed Confidence: quickly collect evidence.

ICPE: Prior and Objectives

Last ingredient is the assumption that we have a prior $\mathcal{P}$ over a set of environments $\mathcal{M}$, representing our belief of M.

What should be the objective?

SEQUENTIAL DESIGN OF EXPERIMENTS

BY HERMAN CHERNOFF[1]

Stanford University

1. Introduction. Considerable scientific research is characterized as follows. The scientist is interested in studying a phenomenon. At first he is quite ignorant and his initial experiments are preliminary and tentative. As he gathers relevant data, he becomes more definite in his impression of the underlying theory. This more definite impression is used to construct more informative experiments. Finally after a certain point he is satisfied that his evidence is sufficient to allow him to announce certain conclusions and he does so.

Objectives:

1. Fixed Budget: maximize evidence over a horizon.
2. Fixed Confidence: quickly collect evidence.

ICPE: Prior and Objectives

Last ingredient is the assumption that we have a prior $\mathcal{P}$ over a set of environments $\mathcal{M}$, representing our belief of M.

What should be the objective? 

SEQUENTIAL DESIGN OF EXPERIMENTS

BY HERMAN CHERNOFF[1]

Stanford University

1. Introduction. Considerable scientific research is characterized as follows. The scientist is interested in studying a phenomenon. At first he is quite ignorant and his initial experiments are preliminary and tentative. As he gathers relevant data, he becomes more definite in his impression of the underlying theory. This more definite impression is used to construct more informative experiments. Finally after a certain point he is satisfied that his evidence is sufficient to allow him to announce certain conclusions and he does so.

Objectives:

1. Fixed Budget: maximize evidence over a horizon.
2. Fixed Confidence: quickly collect evidence.

ICPE: Prior and Objectives

First objective:

Definition (Fixed Budget Problem)

Given a horizon $N \in \mathbb{N}$, the learner chooses a **policy** π and **inference rule** I maximizing the evidence after N queries:

$$\sup_{\pi, I} \mathbb{P}^{\pi}\left(I_N(\mathcal{D}_N) = H^{\star}\right). \tag{1}$$

where

- $\mathcal{D}_N = (x_1, a_1, \ldots, a_N, x_N)$ is the **set of data** collected up to time N [1].
- $\pi(\cdot|\mathcal{D}_t)$ is the **exploration policy** of the agent (mapping $\mathcal{D}_t \to \Delta(\mathcal{A})$ from data to probabilities over actions).
- $I : \mathcal{D}_t \to \mathcal{H}$ is the **inference rule** of the agent, mapping data to a hypothesis.

[1] x_1 is the initial observation.

ICPE: Prior and Objectives

First objective:

Definition (Fixed Budget Problem)

Given a horizon $N \in \mathbb{N}$, the learner chooses a **policy** π and **inference rule** I maximizing the evidence after N queries:

$$\sup_{\pi, I} \mathbb{P}^{\pi}\left(I_N(\mathcal{D}_N) = H^\star\right). \tag{1}$$

where

- $\mathcal{D}_N = (x_1, a_1, \dots, a_N, x_N)$ is the **set of data** collected up to time N [1].
- $\pi(\cdot|\mathcal{D}_t)$ is the **exploration policy** of the agent (mapping $\mathcal{D}_t \to \Delta(\mathcal{A})$ from data to probabilities over actions).
- $I : \mathcal{D}_t \to \mathcal{H}$ is the **inference rule** of the agent, mapping data to a hypothesis.

[1] x_1 is the initial observation.

ICPE: Prior and Objectives

First objective:

Definition (Fixed Budget Problem)

Given a horizon $N \in \mathbb{N}$, the learner chooses a **policy** π and **inference rule** I maximizing the evidence after N queries:

$$\sup_{\pi, I} \mathbb{P}^{\pi}\left(I_N(\mathcal{D}_N) = H^{\star}\right). \tag{1}$$

where

- $\mathcal{D}_N = (x_1, a_1, \dots, a_N, x_N)$ is the **set of data** collected up to time N [1].
- $\pi(\cdot|\mathcal{D}_t)$ is the **exploration policy** of the agent (mapping $\mathcal{D}_t \to \Delta(\mathcal{A})$ from data to probabilities over actions).
- $I : \mathcal{D}_t \to \mathcal{H}$ is the **inference rule** of the agent, mapping data to a hypothesis.

[1] x_1 is the initial observation.

ICPE: Prior and Objectives

First objective:

Definition (Fixed Budget Problem)

Given a horizon $N \in \mathbb{N}$, the learner chooses a **policy** π and **inference rule** I maximizing the evidence after N queries:

$$\sup_{\pi, I} \mathbb{P}^{\pi}\left(I_N(\mathcal{D}_N) = H^{\star}\right). \tag{1}$$

where

- $\mathcal{D}_N = (x_1, a_1, \dots, a_N, x_N)$ is the **set of data** collected up to time N [1].
- $\pi(\cdot|\mathcal{D}_t)$ is the **exploration policy** of the agent (mapping $\mathcal{D}_t \to \Delta(\mathcal{A})$ from data to probabilities over actions).
- $I : \mathcal{D}_t \to \mathcal{H}$ is the **inference rule** of the agent, mapping data to a hypothesis.

[1] x_1 is the initial observation.

ICPE: Prior and Objectives

Second objective:

Definition (Fixed Confidence Problem)

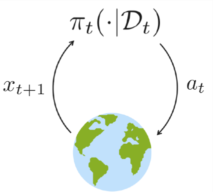
Given a target error level $\delta \in (0,1)$, minimize the number of samples τ needed to identify $H^\star$ with confidence $1-\delta$:

$$\inf_{\tau,\pi,I} \mathbb{E}^{\pi}[\tau] \quad \text{s.t.} \quad \mathbb{P}^{\pi}\left(I_\tau(\mathcal{D}_\tau) = H^\star\right) \geq 1-\delta. \tag{2}$$

where

- $\mathcal{D}_\tau = (x_1, a_1, \ldots, a_\tau, x_\tau)$ is the **set of data** collected up to time τ [2].
- $\pi(\cdot|\mathcal{D}_t)$ is the **policy** of the agent (mapping $\mathcal{D}_t \to \Delta(\mathcal{A})$ from data to probabilities over actions).
- $I : \mathcal{D}_t \to \mathcal{H}$ is the **inference** rule of the agent, mapping data to a hypothesis.

[2] x_1 is the initial observation.



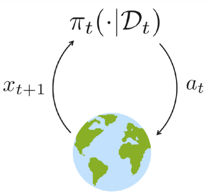
An agent π interacts with the environment and collects data.

- All very good...but how do we learn π, I, τ? How do we optimize

$$\inf_{\tau,\pi,I} \mathbb{E}^{\pi}[\tau] \quad \text{s.t.} \quad \mathbb{P}^{\pi}(I_\tau(\mathcal{D}_\tau) = H^\star) \geq 1-\delta.$$

  or the fixed budget objective?
- Intuitively, optimizing π seems an RL problem... and learning I seems like a supervised learning problem.
- We need a bit of theory that can point us out in the right direction.

ICPE: How to learn π, I?



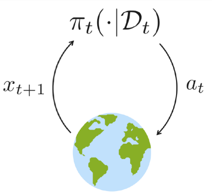
An agent π interacts with the environment and collects data.

- All very good...but how do we learn π, I, τ? How do we optimize

$$\inf_{\tau,\pi,I} \mathbb{E}^\pi[\tau] \quad \text{s.t.} \quad \mathbb{P}^\pi(I_\tau(\mathcal{D}_\tau) = H^\star) \geq 1-\delta.$$

  or the fixed budget objective?
- Intuitively, optimizing π seems an RL problem... and learning I seems like a supervised learning problem.
- We need a bit of theory that can point us out in the right direction.

ICPE: How to learn π, I?



An agent π interacts with the environment and collects data.

- All very good...but how do we learn π, I, τ? How do we optimize

$$\inf_{\tau,\pi,I} \mathbb{E}^{\pi}[\tau] \quad \text{s.t.} \quad \mathbb{P}^{\pi}\left(I_\tau(\mathcal{D}_\tau) = H^\star\right) \geq 1 - \delta.$$

  or the fixed budget objective?
- Intuitively, optimizing π seems an RL problem... and learning I seems like a supervised learning problem.
- We need a bit of theory that can point us out in the right direction.

ICPE: Some Theory

Why Theory?

- First, we see that optimizing the inference rule I amounts to computing a posterior distribution[3].
- Secondly, the policy π can be learned using RL with an appropriate reward function.

Importantly, the reward function used for training π *emerges* naturally from the problem formulation, and it *is not* a user-chosen criterion, making it a principled information-theoretical reward function.

[3] The results we present here hold for general continuous observation/action spaces (but finite number of hypotheses).

Why Theory?

- First, we see that optimizing the inference rule I amounts to computing a posterior distribution[3].
- Secondly, the policy π can be learned using RL with an appropriate reward function.

Importantly, the reward function used for training π *emerges* naturally from the problem formulation, and it *is not* a user-chosen criterion, making it a principled information-theoretical reward function.

[3] The results we present here hold for general continuous observation/action spaces (but finite number of hypotheses).

Why Theory?

- First, we see that optimizing the inference rule I amounts to computing a posterior distribution[3].
- Secondly, the policy π can be learned using RL with an appropriate reward function.

Importantly, the reward function used for training π *emerges* naturally from the problem formulation, and it *is not* a user-chosen criterion, making it a principled information-theoretical reward function.

[3] The results we present here hold for general continuous observation/action spaces (but finite number of hypotheses).

Optimal Inference Rule

Idea: can we find an inference rule that is π-independent to simplify the optimization problem?

Proposition

Let $t \geq 1$ and a policy π. The optimal inference rule maximizing $\sup_I \mathbb{P}^\pi(H^\star = I_t(\mathcal{D}_t))$ is given by

$$I^\star(\mathcal{D}_t) = \arg\max_{H \in \mathcal{H}} \mathbb{P}(H^\star = H | \mathcal{D}_t)^4,$$

where $\mathbb{P}(H^\star \in \cdot | \mathcal{D}_t)$ is the posterior distribution of $H^\star$.

In other words, the optimal prediction is the hypothesis that maximizes the posterior distribution of the true hypothesis.

[4] $\mathcal{D}_t$ is a possible trajectory $(x_1, a_1, \ldots, x_t)$.

Optimal Inference Rule

Idea: can we find an inference rule that is π-independent to simplify the optimization problem?

Proposition

Let $t \geq 1$ and a policy π. The optimal inference rule maximizing $\sup_I \mathbb{P}^\pi(H^\star = I_t(\mathcal{D}_t))$ is given by

$$I^\star(\mathcal{D}_t) = \arg\max_{H \in \mathcal{H}} \mathbb{P}(H^\star = H | \mathcal{D}_t)^4,$$

where $\mathbb{P}(H^\star \in \cdot | \mathcal{D}_t)$ is the posterior distribution of $H^\star$.

In other words, the optimal prediction is the hypothesis that maximizes the posterior distribution of the true hypothesis.

[4] $\mathcal{D}_t$ is a possible trajectory $(x_1, a_1, \ldots, x_t)$.

Optimal Policy in Fixed Budget

Consider the fixed budget setting. Using the previous result, we can immediately conclude that for all $t \geq 1$

$$\sup_{\pi, I} \mathbb{P}^{\pi}(H^\star = I(\mathcal{D}_t)) = \sup_{\pi} \mathbb{E}^{\pi}[r(\mathcal{D}_t)]$$

where $r(\mathcal{D}_t) := \max_{H \in \mathcal{H}} \mathbb{P}_t(H^\star = H|\mathcal{D}_t)$. Does this ring a

We can use RL!

Optimal Policy in Fixed Budget

Consider the fixed budget setting. Using the previous result, we can immediately conclude that for all $t \geq 1$

$$\sup_{\pi, I} \mathbb{P}^\pi(H^\star = I(\mathcal{D}_t)) = \sup_\pi \mathbb{E}^\pi[r(\mathcal{D}_t)]$$

where $r(\mathcal{D}_t) := \max_{H \in \mathcal{H}} \mathbb{P}_t(H^\star = H | \mathcal{D}_t)$. Does this ring a

We can use RL!

Optimal Policy in Fixed Budget

Consider the **fixed budget** **setting.** Using the previous result, we can immediately conclude that for all $t \geq 1$

$$\sup_{\pi, I} \mathbb{P}^{\pi}(H^\star = I(\mathcal{D}_t)) = \sup_{\pi} \mathbb{E}^{\pi}[r(\mathcal{D}_t)]$$

where $r(\mathcal{D}_t) := \max_{H \in \mathcal{H}} \mathbb{P}_t(H^\star = H | \mathcal{D}_t)$. Does this ring a  ?

We can use RL!

Optimal Policy in Fixed Budget

Consider the fixed budget setting. Using the previous result, we can immediately conclude that for all $t \geq 1$

$$\sup_{\pi, I} \mathbb{P}^{\pi}(H^{\star} = I(\mathcal{D}_t)) = \sup_{\pi} \mathbb{E}^{\pi}[r(\mathcal{D}_t)]$$

where $r(\mathcal{D}_t) := \max_{H \in \mathcal{H}} \mathbb{P}_t(H^{\star} = H | \mathcal{D}_t)$. Does this ring a  ?

We can use RL!

Wait... are we back to RL?

We can use RL!

Wait.. we use RL to solve an exploration problem? *"It's like hiring a tour guide... who asks you to show them around first!"*

Is it always an RL problem at the end?

Wait... are we back to RL?

We can use RL!

Wait.. we use RL to solve an exploration problem? *"It's like hiring a tour guide... who asks you to show them around first!"*

Is it always an RL problem at the end?

Wait... are we back to RL?

We can use RL!

Wait.. we use RL to solve an exploration problem? *"It's like hiring a tour guide... who asks you to show them around first!"*

Is it always an RL problem at the end?

Wait... are we back to RL?

We can use RL!

Wait.. we use RL to solve an exploration problem? *"It's like hiring a tour guide... who asks you to show them around first!"*

Is it always an RL problem at the end?

Wait... are we back to RL?

We can use RL!

Wait.. we use RL to solve an exploration problem? *"It's like hiring a tour guide... who asks you to show them around first!"*

Is it always an RL problem at the end?

Optimal Exploration Policy in Fixed Budget

$$\sup_{\pi,I} \mathbb{P}^\pi(H^\star = I(\mathcal{D}_t)) = \sup_{\pi} \mathbb{E}^\pi[r(\mathcal{D}_t)]$$

where $r(\mathcal{D}_t) := \max_{H\in\mathcal{H}} \mathbb{P}_t(H^\star = H|\mathcal{D}_t)$.

Proposition (Policy Optimality)

For any $t \geq 1$ the policy $\pi^\star(\mathcal{D}_t) = \arg\max_{a\in\mathcal{A}} Q_t(\mathcal{D}_t, a)$ is optimal, where

$$V_t(\mathcal{D}_t) = \max_{a\in\mathcal{A}} Q_t(\mathcal{D}_t, a), \quad Q_t(\mathcal{D}_t, a) = \int_{\mathcal{X}} V_{t+1}\underbrace{(\mathcal{D}_t, a, x')}_{=\mathcal{D}_{t+1}} \underbrace{\bar{P}_t(\mathrm{d}x'|\mathcal{D}_t, a)}_{\text{posterior distribution of } x'}, \quad t < N,$$

with $V_N(\mathcal{D}_N) = \max_{H\in\mathcal{H}} \mathbb{P}_t(H^\star = H|\mathcal{D}_N)$.

Optimal Exploration Policy in Fixed Budget

$$\sup_{\pi,I} \mathbb{P}^\pi(H^\star = I(\mathcal{D}_t)) = \sup_{\pi} \mathbb{E}^\pi[r(\mathcal{D}_t)]$$

where $r(\mathcal{D}_t) := \max_{H\in\mathcal{H}} \mathbb{P}_t(H^\star = H|\mathcal{D}_t)$.

Proposition (Policy Optimality)

For any $t \geq 1$ the policy $\pi^\star(\mathcal{D}_t) = \arg\max_{a\in\mathcal{A}} Q_t(\mathcal{D}_t, a)$ is optimal, where

$$V_t(\mathcal{D}_t) = \max_{a\in\mathcal{A}} Q_t(\mathcal{D}_t, a), \quad Q_t(\mathcal{D}_t, a) = \int_{\mathcal{X}} V_{t+1}\underbrace{(\mathcal{D}_t, a, x')}_{=\mathcal{D}_{t+1}} \underbrace{\bar{P}_t(\mathrm{d}x'|\mathcal{D}_t, a)}_{\text{posterior distribution of } x'}, \quad t < N,$$

with $V_N(\mathcal{D}_N) = \max_{H\in\mathcal{H}} \mathbb{P}_t(H^\star = H|\mathcal{D}_N)$.

The Fixed Confidence Setting

Recall the **fixed confidence** setting

$$\inf_{\tau,\pi,I} \mathbb{E}^{\pi}[\tau] \quad \text{s.t.} \quad \mathbb{P}^{\pi}\left(I_\tau(\mathcal{D}_\tau) = H^\star\right) \geq 1-\delta.$$

- τ indicates the number of samples... formally, it's a *stopping rule*. **Informally,** this rule tells you if you should **stop** or **continue** given the data $\mathcal{D}_t$.
- **First simplification**: we define a stopping action a_{stop} and set $\mathcal{A} \leftarrow \mathcal{A} \cup \{a_{\text{stop}}\}$, $\tau = \inf\{t \in \mathbb{N} : a_t = a_{\text{stop}}\} \Rightarrow$ *the policy π decides when to stop.*
- **Second simplification**: we study the dual problem!.

The Fixed Confidence Setting

Recall the **fixed confidence** setting

$$\inf_{\tau,\pi,I} \mathbb{E}^{\pi}[\tau] \quad \text{s.t.} \quad \mathbb{P}^{\pi}(I_\tau(\mathcal{D}_\tau) = H^\star) \geq 1-\delta.$$

- τ indicates the number of samples... formally, it's a *stopping rule.* **Informally,** this rule tells you if you should **stop** or **continue** given the data $\mathcal{D}_t$.
- **First simplification**: we define a stopping action a_{stop} and set $\mathcal{A} \leftarrow \mathcal{A} \cup \{a_{\text{stop}}\}$, $\tau = \inf\{t \in \mathbb{N} : a_t = a_{\text{stop}}\} \Rightarrow$ *the policy π decides when to stop.*
- **Second simplification**: we study the dual problem!.

The Fixed Confidence Setting

Recall the **fixed confidence** setting

$$\inf_{\tau,\pi,I} \mathbb{E}^{\pi}[\tau] \quad \text{s.t.} \quad \mathbb{P}^{\pi}\left(I_{\tau}(\mathcal{D}_{\tau}) = H^{\star}\right) \geq 1-\delta.$$

- τ indicates the number of samples... formally, it's a *stopping rule.* **Informally,** this rule tells you if you should **stop** or **continue** given the data $\mathcal{D}_t$.
- **First simplification**: we define a stopping action a_{stop} and set $\mathcal{A} \leftarrow \mathcal{A} \cup \{a_{\text{stop}}\}$, $\tau = \inf\{t \in \mathbb{N} : a_t = a_{\text{stop}}\}$ $\Rightarrow$ *the policy π decides when to stop.*
- **Second simplification**: we study the dual problem!.

The Fixed Confidence Setting

Recall the **fixed confidence** setting

$$\inf_{\tau,\pi,I} \mathbb{E}^\pi[\tau] \quad \text{s.t.} \quad \mathbb{P}^\pi\left(I_\tau(\mathcal{D}_\tau) = H^\star\right) \geq 1-\delta.$$

- τ indicates the number of samples... formally, it's a *stopping rule.* **Informally,** this rule tells you if you should **stop** or **continue** given the data $\mathcal{D}_t$.
- **First simplification**: we define a stopping action a_{stop} and set $\mathcal{A} \leftarrow \mathcal{A} \cup \{a_{\text{stop}}\}$, $\tau = \inf\{t \in \mathbb{N} : a_t = a_{\text{stop}}\} \Rightarrow$ *the policy* π *decides when to stop.*
- **Second simplification**: we study the dual problem!.

The Fixed Confidence Setting

Recall the **fixed confidence** setting

$$\inf_{\tau,\pi,I} \mathbb{E}^\pi[\tau] \quad \text{s.t.} \quad \mathbb{P}^\pi(I_\tau(\mathcal{D}_\tau) = H^\star) \geq 1-\delta.$$

- τ indicates the number of samples... formally, it's a *stopping rule.* **Informally,** this rule tells you if you should **stop** or **continue** given the data $\mathcal{D}_t$.
- **First simplification**: we define a stopping action a_{stop} and set $\mathcal{A} \leftarrow \mathcal{A} \cup \{a_{\text{stop}}\}$, $\tau = \inf\{t \in \mathbb{N} : a_t = a_{\text{stop}}\} \Rightarrow$ *the policy* π *decides when to stop.*
- **Second simplification**: we study the dual problem!.

The Fixed Confidence Setting

We then study the dual problem

$$\inf_{\lambda \geq 0} \sup_{\pi, I} V_\lambda(\pi, I), \quad \text{where} \quad V_\lambda(\pi, I) := -\mathbb{E}^\pi[\tau] + \lambda \left[\mathbb{P}^\pi\left(I(\mathcal{D}_\tau) = H^\star\right) - 1 + \delta\right].$$

- Can we use the optimal inference rule result? Yes.
- Can it be simplified to an RL problem? Yes.
- Do we get the δ-probably correct guarantees? Not immediately.

The Fixed Confidence Setting

We then study the dual problem

$$\inf_{\lambda \geq 0} \sup_{\pi, I} V_\lambda(\pi, I), \quad \text{where} \quad V_\lambda(\pi, I) := -\mathbb{E}^\pi[\tau] + \lambda \left[\mathbb{P}^\pi\left(I(\mathcal{D}_\tau) = H^\star\right) - 1 + \delta\right].$$

- Can we use the optimal inference rule result? Yes.
- Can it be simplified to an RL problem? Yes.
- Do we get the δ-probably correct guarantees? Not immediately.

The Fixed Confidence Setting

We then study the dual problem

$$\inf_{\lambda \geq 0} \sup_{\pi, I} V_\lambda(\pi, I), \quad \text{where} \quad V_\lambda(\pi, I) := -\mathbb{E}^\pi[\tau] + \lambda \left[\mathbb{P}^\pi\left(I(\mathcal{D}_\tau) = H^\star\right) - 1 + \delta\right].$$

- Can we use the optimal inference rule result? Yes.
- Can it be simplified to an RL problem? Yes.
- Do we get the δ-probably correct guarantees? Not immediately.

We then study the dual problem

$$\inf_{\lambda \geq 0} \sup_{\pi, I} V_\lambda(\pi, I), \quad \text{where } V_\lambda(\pi, I) := -\mathbb{E}^\pi[\tau] + \lambda \left[\mathbb{P}^\pi\left(I(\mathcal{D}_\tau) = H^\star\right) - 1 + \delta\right].$$

- Can we use the optimal inference rule result? Yes.
- Can it be simplified to an RL problem? Yes.
- Do we get the δ-probably correct guarantees? Not immediately.

The Fixed Confidence Setting

We then study the dual problem

$$\inf_{\lambda \geq 0} \sup_{\pi, I} V_\lambda(\pi, I), \quad \text{where } V_\lambda(\pi, I) := -\mathbb{E}^\pi[\tau] + \lambda \left[\mathbb{P}^\pi \left(I(\mathcal{D}_\tau) = H^\star\right) - 1 + \delta\right].$$

- Can we use the optimal inference rule result? Yes.
- Can it be simplified to an RL problem? Yes.
- Do we get the δ-probably correct guarantees? Not immediately.

The Fixed Confidence Setting

We then study the dual problem

$$\inf_{\lambda \geq 0} \sup_{\pi, I} V_\lambda(\pi, I), \quad \text{where} \quad V_\lambda(\pi, I) := -\mathbb{E}^\pi[\tau] + \lambda \left[\mathbb{P}^\pi\left(I(\mathcal{D}_\tau) = H^\star\right) - 1 + \delta\right].$$

- Can we use the optimal inference rule result? Yes.
- Can it be simplified to an RL problem? Yes.
- Do we get the δ-probably correct guarantees? Not immediately.

The Fixed Confidence Setting

We then study the dual problem

$$\inf_{\lambda \geq 0} \sup_{\pi, I} V_\lambda(\pi, I), \quad \text{where } V_\lambda(\pi, I) := -\mathbb{E}^\pi[\tau] + \lambda \left[\mathbb{P}^\pi\left(I(\mathcal{D}_\tau) = H^\star\right) - 1 + \delta\right].$$

- Can we use the optimal inference rule result? Yes.
- Can it be simplified to an RL problem? Yes.
- Do we get the δ-probably correct guarantees? Not immediately.

The Fixed Confidence Setting: Optimal Exploration Policy

$$\inf_{\lambda\geq 0}\sup_{\pi,I} V_\lambda(\pi,I), \quad \text{where} \quad V_\lambda(\pi,I) := -\mathbb{E}^\pi[\tau] + \lambda\left[\mathbb{P}^\pi\left(I(\mathcal{D}_\tau) = H^\star\right) - 1 + \delta\right].$$

Define the reward

$$r_\lambda(\mathcal{D}_t, a) = -\mathbf{1}_{\{a\neq a_{\text{stop}}\}} + \lambda \mathbf{1}_{\{a=a_{\text{stop}}\}} \max_H \mathbb{P}(H^\star = H|\mathcal{D}_t),$$

and the Q-function

$$Q_\lambda(\mathcal{D}_t, a) = r_\lambda(\mathcal{D}_t, a) + \mathbf{1}_{\{a\neq a_{\text{stop}}\}}\mathbb{E}_{x_{t+1}|(\mathcal{D}_t,a)}\left[\max_{a'} Q_{t+1,\lambda}\left((\mathcal{D}_t, a, x_{t+1}), a'\right)\right].$$

where $x_{t+1}|(\mathcal{D}_t, a)$ indicates the posterior distribution of x_{t+1} given $(\mathcal{D}_t, a)$.

The Fixed Confidence Setting: Optimal Exploration Policy

$$\inf_{\lambda \geq 0} \sup_{\pi, I} V_\lambda(\pi, I), \quad \text{where} \quad V_\lambda(\pi, I) := -\mathbb{E}^\pi[\tau] + \lambda \left[\mathbb{P}^\pi\left(I(\mathcal{D}_\tau) = H^\star\right) - 1 + \delta\right].$$

Define the reward

$$r_\lambda(\mathcal{D}_t, a) = -\mathbf{1}_{\{a \neq a_{\text{stop}}\}} + \lambda \mathbf{1}_{\{a = a_{\text{stop}}\}} \max_H \mathbb{P}(H^\star = H | \mathcal{D}_t),$$

and the Q-function

$$Q_\lambda(\mathcal{D}_t, a) = r_\lambda(\mathcal{D}_t, a) + \mathbf{1}_{\{a \neq a_{\text{stop}}\}} \mathbb{E}_{x_{t+1}|(\mathcal{D}_t, a)} \left[\max_{a'} Q_{t+1,\lambda}\left((\mathcal{D}_t, a, x_{t+1}), a'\right)\right].$$

where $x_{t+1}|(\mathcal{D}_t, a)$ indicates the posterior distribution of x_{t+1} given $(\mathcal{D}_t, a)$.

The Fixed Confidence Setting: Optimal Exploration Policy

$$\inf_{\lambda\geq 0}\sup_{\pi,I} V_\lambda(\pi,I), \quad \text{where} \quad V_\lambda(\pi,I) := -\mathbb{E}^\pi[\tau] + \lambda\left[\mathbb{P}^\pi\left(I(\mathcal{D}_\tau) = H^\star\right) - 1 + \delta\right].$$

Define the reward

$$r_\lambda(\mathcal{D}_t, a) = -\mathbf{1}_{\{a\neq a_{\text{stop}}\}} + \lambda\mathbf{1}_{\{a=a_{\text{stop}}\}}\max_H \mathbb{P}(H^\star = H|\mathcal{D}_t),$$

and the Q-function

$$Q_\lambda(\mathcal{D}_t, a) = r_\lambda(\mathcal{D}_t, a) + \mathbf{1}_{\{a\neq a_{\text{stop}}\}}\mathbb{E}_{x_{t+1}|(\mathcal{D}_t,a)}\left[\max_{a'} Q_{t+1,\lambda}\left((\mathcal{D}_t, a, x_{t+1}), a'\right)\right].$$

where $x_{t+1}|(\mathcal{D}_t, a)$ indicates the posterior distribution of x_{t+1} given $(\mathcal{D}_t, a)$.

The Fixed Confidence Setting: Optimal Exploration Policy

$$r_\lambda(\mathcal{D}_t, a) = -\mathbf{1}_{\{a \neq a_{\text{stop}}\}} + \lambda \mathbf{1}_{\{a = a_{\text{stop}}\}} \max_H \mathbb{P}(H^\star = H | \mathcal{D}_t),$$

$$Q_\lambda(\mathcal{D}_t, a) = r_\lambda(\mathcal{D}_t, a) + \mathbf{1}_{\{a \neq a_{\text{stop}}\}} \mathbb{E}_{x_{t+1}|(\mathcal{D}_t, a)} \left[\max_{a'} Q_{t+1,\lambda}\left((\mathcal{D}_t, a, x_{t+1}), a' \right) \right].$$

Proposition

Let $\pi_\lambda^\star(\mathcal{D}_t) = \arg\max_{a \in \mathcal{A}} Q_\lambda(\mathcal{D}_t, a)$. Then, for $\lambda \geq 0$ the pair $(I^\star, \pi_\lambda^\star)$, (with $I^\star$ as before), is an optimal solution of $\sup_{\pi, I} V_\lambda(\pi, I)$. Furthermore, under suitable identifiability conditions[5], any maximizer $\lambda^\star$ guarantees that $\pi_{\lambda^\star}^\star$ satisfies the δ-correctness criterion.

We can do RL! Rejoice!

[5] See the appendix for more details.

The Fixed Confidence Setting: Optimal Exploration Policy

$$r_\lambda(\mathcal{D}_t, a) = -\mathbf{1}_{\{a \neq a_{\text{stop}}\}} + \lambda \mathbf{1}_{\{a = a_{\text{stop}}\}} \max_H \mathbb{P}(H^\star = H | \mathcal{D}_t),$$

$$Q_\lambda(\mathcal{D}_t, a) = r_\lambda(\mathcal{D}_t, a) + \mathbf{1}_{\{a \neq a_{\text{stop}}\}} \mathbb{E}_{x_{t+1}|(\mathcal{D}_t, a)} \left[\max_{a'} Q_{t+1,\lambda}\left((\mathcal{D}_t, a, x_{t+1}), a' \right) \right].$$

Proposition

Let $\pi^\star_\lambda(\mathcal{D}_t) = \arg\max_{a \in \mathcal{A}} Q_\lambda(\mathcal{D}_t, a)$. Then, for $\lambda \geq 0$ the pair $(I^\star, \pi^\star_\lambda)$, (with $I^\star$ as before), is an optimal solution of $\sup_{\pi, I} V_\lambda(\pi, I)$. Furthermore, under suitable identifiability conditions[5], any maximizer $\lambda^\star$ guarantees that $\pi^\star_{\lambda^\star}$ satisfies the δ-correctness criterion.

We can do RL! Rejoice!

[5] See the appendix for more details.

$$r_\lambda(\mathcal{D}_t, a) = -\mathbf{1}_{\{a \neq a_{\text{stop}}\}} + \lambda \mathbf{1}_{\{a = a_{\text{stop}}\}} \max_H \mathbb{P}(H^\star = H | \mathcal{D}_t),$$

$$Q_\lambda(\mathcal{D}_t, a) = r_\lambda(\mathcal{D}_t, a) + \mathbf{1}_{\{a \neq a_{\text{stop}}\}} \mathbb{E}_{x_{t+1}|(\mathcal{D}_t, a)} \left[\max_{a'} Q_{t+1,\lambda}\left((\mathcal{D}_t, a, x_{t+1}), a'\right) \right].$$

Proposition

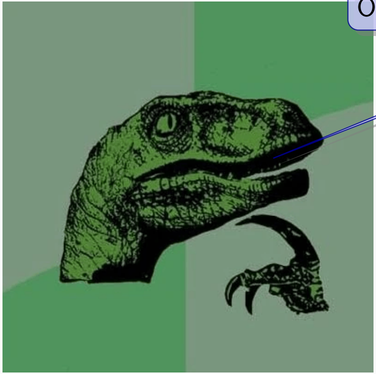
Let $\pi^\star_\lambda(\mathcal{D}_t) = \arg\max_{a \in \mathcal{A}} Q_\lambda(\mathcal{D}_t, a)$. Then, for $\lambda \geq 0$ the pair $(I^\star, \pi^\star_\lambda)$, (with $I^\star$ as before), is an optimal solution of $\sup_{\pi, I} V_\lambda(\pi, I)$. Furthermore, under suitable identifiability conditions[5], any maximizer $\lambda^\star$ guarantees that $\pi^\star_{\lambda^\star}$ satisfies the δ-correctness criterion.

We can do RL! Rejoice!

[5] See the appendix for more details.

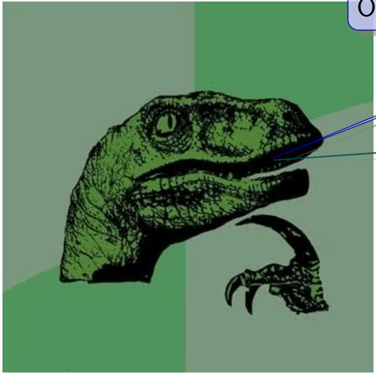
ICPE: Practical Design

Now What?



Ok we can use RL...What else do we have? A prior.

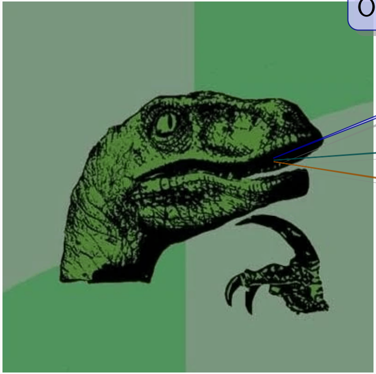
Now What?



Ok we can use RL...What else do we have? A prior.

If we have a simulator that can sample from this prior...

Now What?

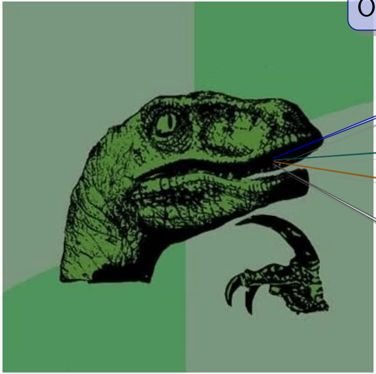


Ok we can use RL...What else do we have? A prior.

If we have a simulator that can sample from this prior...

...we could then learn π, I using data from this simulator...

Now What?



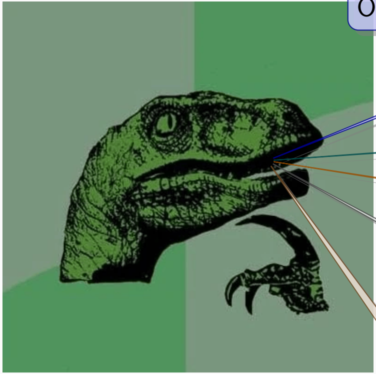
Ok we can use RL...What else do we have? A prior.

If we have a simulator that can sample from this prior...

...we could then learn π, I using data from this simulator...

...and then use π, I on the real problem.

Now What?



Ok we can use RL...What else do we have? A prior.

If we have a simulator that can sample from this prior...

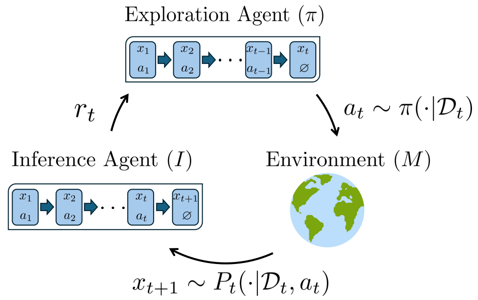
...we could then learn π, I using data from this simulator...

...and then use π, I on the real problem.

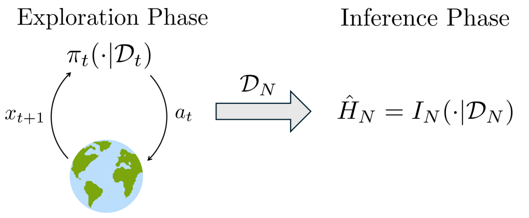
Anything else? Yes, we use Transformers to handle the sequence of data $\mathcal{D}_t$.

Training vs Inference

Training time



Inference time



- At training time ICPE interacts with a simulator: each episode draws an instance $M \sim \mathcal{P}$ and generates a trajectory.
- We maintain a buffer $\mathcal{B}$ to store the training data.

Training phase

Let us consider the **fixed confidence setting**. What do we have?

$$
\begin{aligned}
I^\star(\mathcal{D}_t) &= \arg\max_{H \in \mathcal{H}} \mathbb{P}(H^\star = H | \mathcal{D}_t), \\
r_\lambda(\mathcal{D}_t, a) &= -\mathbf{1}_{\{a \neq a_{\text{stop}}\}} + \lambda \mathbf{1}_{\{a = a_{\text{stop}}\}} \max_H \mathbb{P}(H^\star = H | \mathcal{D}_t), \\
Q_\lambda(\mathcal{D}_t, a) &= r_\lambda(\mathcal{D}_t, a) + \mathbf{1}_{\{a \neq a_{\text{stop}}\}} \mathbb{E}_{x_{t+1} | (\mathcal{D}_t, a)} \left[\max_{a'} Q_{t+1,\lambda}\left((\mathcal{D}_t, a, x_{t+1}), a' \right) \right].
\end{aligned}
$$

Training phase

Let us consider the **fixed confidence setting**. What do we have?

$$I^\star(\mathcal{D}_t) = \arg\max_{H \in \mathcal{H}} \mathbb{P}(H^\star = H | \mathcal{D}_t) \Rightarrow \text{learn } I(H|\mathcal{D}_t) = \mathbb{P}(H^\star = H|\mathcal{D}_t) \; ,$$

$$r_\lambda(\mathcal{D}_t, a) = -\mathbf{1}_{\{a \neq a_{\text{stop}}\}} + \lambda \mathbf{1}_{\{a = a_{\text{stop}}\}} \max_H \mathbb{P}(H^\star = H | \mathcal{D}_t),$$

$$Q_\lambda(\mathcal{D}_t, a) = r_\lambda(\mathcal{D}_t, a) + \mathbf{1}_{\{a \neq a_{\text{stop}}\}} \mathbb{E}_{x_{t+1}|(\mathcal{D}_t, a)} \left[\max_{a'} Q_{t+1,\lambda}\left((\mathcal{D}_t, a, x_{t+1}), a'\right) \right].$$

Training phase

Let us consider the **fixed confidence setting**. What do we have?

$$I^\star(\mathcal{D}_t) = \arg\max_{H \in \mathcal{H}} \mathbb{P}(H^\star = H | \mathcal{D}_t) \Rightarrow \text{learn } I(H|\mathcal{D}_t) = \mathbb{P}(H^\star = H|\mathcal{D}_t) \ ,$$

$$r_\lambda(\mathcal{D}_t, a) = -\mathbf{1}_{\{a \neq a_{\text{stop}}\}} + \lambda \mathbf{1}_{\{a = a_{\text{stop}}\}} \max_H I(H|\mathcal{D}_t),$$

$$Q_\lambda(\mathcal{D}_t, a) = r_\lambda(\mathcal{D}_t, a) + \mathbf{1}_{\{a \neq a_{\text{stop}}\}} \mathbb{E}_{x_{t+1}|(\mathcal{D}_t, a)} \left[\max_{a'} Q_{t+1,\lambda}\left((\mathcal{D}_t, a, x_{t+1}), a'\right) \right].$$

General idea: cross-entropy loss to learn I and your favorite off-policy Deep-RL technique to learn π.

Training phase

Let us consider the **fixed confidence setting**. What do we have?

$$I^\star(\mathcal{D}_t) = \arg\max_{H \in \mathcal{H}} \mathbb{P}(H^\star = H | \mathcal{D}_t) \Rightarrow \text{learn } I(H|\mathcal{D}_t) = \mathbb{P}(H^\star = H|\mathcal{D}_t) \ ,$$

$$r_\lambda(\mathcal{D}_t, a) = -\mathbf{1}_{\{a \neq a_{\text{stop}}\}} + \lambda \mathbf{1}_{\{a = a_{\text{stop}}\}} \max_H I(H|\mathcal{D}_t),$$

$$Q_\lambda(\mathcal{D}_t, a) = r_\lambda(\mathcal{D}_t, a) + \mathbf{1}_{\{a \neq a_{\text{stop}}\}} \mathbb{E}_{x_{t+1}|(\mathcal{D}_t, a)} \left[\max_{a'} Q_{t+1,\lambda}\left((\mathcal{D}_t, a, x_{t+1}), a'\right) \right].$$

General idea: cross-entropy loss to learn I and your favorite off-policy Deep-RL technique to learn π.

Training phase: inference rule

Inference rule: parametrize I by ϕ. We train it with the loss[6]

$$\mathcal{L}_{\text{inf}}(\phi) = -\frac{1}{|B|} \sum_{(\mathcal{D}_t, a_t, x_{t+1}, H^\star) \in B} \log I_\phi(H^\star | \mathcal{D}_{t+1}). \tag{3}$$

where $B \sim \mathcal{B}$ is a batch of data from the buffer.

[6] In expectation this is (up to an additive constant) equivalent to minimizing the KL-divergence between $\mathbb{P}(H^\star = H|\mathcal{D})$ and $I_\phi(H|\mathcal{D})$.

Training phase: exploration policy

Exploration policy:

$$r(\mathcal{D}_t, a) = -\mathbf{1}_{\{a \neq a_{\text{stop}}\}} + \lambda \mathbf{1}_{\{a = a_{\text{stop}}\}} \max_H \underbrace{I_{\bar{\phi}}}_{\text{target network}} (H | \mathcal{D}_t),$$

$$Q_\theta(\mathcal{D}_t, a) = r(\mathcal{D}_t, a) + \mathbf{1}_{\{a \neq a_{\text{stop}}\}} \mathbb{E}_{x_{t+1} | (\mathcal{D}_t, a)} \left[\max_{a'} \underbrace{Q_{\bar{\theta}}}_{\text{target network}} ((\mathcal{D}_t, a, x_{t+1}), a') \right].$$

- We use target parameters $\bar{\phi}$ and $\bar{\theta}$ to stabilize training[7]. [8]

[7] Target networks were introduced in DQN [Mnih et al., 2013]

[8] We also use a cost variable c instead of λ to avoid the product $\lambda \cdot I$ (see appendix for details).

Training phase: exploration policy

$$r(\mathcal{D}_t, a) = -\mathbf{1}_{\{a \neq a_{\text{stop}}\}} + \lambda \mathbf{1}_{\{a = a_{\text{stop}}\}} \max_H I_{\bar{\phi}}(H|\mathcal{D}_t),$$

$$Q_\theta(\mathcal{D}_t, a) = r(\mathcal{D}_t, a) + \mathbf{1}_{\{a \neq a_{\text{stop}}\}} \mathbb{E}_{x_{t+1}|(\mathcal{D}_t, a)} \left[\max_{a'} Q_{\bar{\theta}}\left((\mathcal{D}_t, a, x_{t+1}), a'\right) \right].$$

We use this DQN-like loss

$$\mathcal{L}_{\text{policy}}(B;\theta) = \frac{1}{|B|} \sum_{(\mathcal{D}_t, a_t, x_{t+1}) \in B} \Bigg[\mathbf{1}_{\{a_t \neq a_{\text{stop}}\}} \cdot \left(-1 + \max_a Q_{\bar{\theta}}(\mathcal{D}_{t+1}, a) - Q_\theta(\mathcal{D}_t, a_t) \right)^2 \tag{4}$$

$$+ \left(\lambda \max_H I_{\bar{\phi}}(H|\mathcal{D}_t) - Q_\theta(\mathcal{D}_t, a_{\text{stop}}) \right)^2 \Bigg], \tag{5}$$

9

[9] The Q-value of a_{stop} can be updated at any time, allowing retrospective evaluation of stopping.

Training phase: Lagrangian variable

Last, but not least, we need to update λ!

$$\inf_{\lambda \geq 0} \sup_{\pi, I} V_\lambda(\pi, I), \quad \text{where} \quad V_\lambda(\pi, I) := -\mathbb{E}^\pi[\tau] + \lambda \left[\mathbb{P}^\pi\left(I(\mathcal{D}_\tau) = H^\star\right) - 1 + \delta\right].$$

We learn λ using a gradient descent update:

$$\lambda \leftarrow \max\left[0, \lambda - \beta\left(\hat{p} - 1 + \delta\right)\right], \quad \text{where} \quad \hat{p} = \frac{1}{K} \sum_{i=1}^{K} \mathbf{1}_{\{\arg\max_H I_\phi(H|\mathcal{D}_\tau^{(i)})\} = H_i^\star\}}. \qquad (6)$$

using K i.i.d. trajectories $\{(\mathcal{D}_\tau^{(i)}, H_i^\star)\}_{i=1}^K$ with fixed (θ, ϕ).

Training phase: Lagrangian variable

Last, but not least, we need to update λ!

$$\inf_{\lambda \geq 0} \sup_{\pi, I} V_\lambda(\pi, I), \quad \text{where} \quad V_\lambda(\pi, I) := -\mathbb{E}^\pi[\tau] + \lambda \left[\mathbb{P}^\pi\left(I(\mathcal{D}_\tau) = H^\star\right) - 1 + \delta\right].$$

We learn λ using a gradient descent update:

$$\lambda \leftarrow \max\left[0, \lambda - \beta\left(\hat{p} - 1 + \delta\right)\right], \quad \text{where} \quad \hat{p} = \frac{1}{K}\sum_{i=1}^{K} \mathbf{1}_{\{\arg\max_H I_\phi(H|\mathcal{D}_\tau^{(i)})\} = H_i^\star\}}. \tag{6}$$

using K i.i.d. trajectories $\{(\mathcal{D}_\tau^{(i)}, H_i^\star)\}_{i=1}^K$ with fixed (θ, ϕ).

Full algorithm

Algorithm 1 ICPE (In-Context Pure Exploration)

Input: Tasks distribution $\mathcal{P}$; confidence δ; horizon N; initial λ and hyper-parameter T_ϕ, T_θ.
// Training phase
Initialize buffer $\mathcal{B}$, networks Q_θ, I_ϕ and set $\bar{\theta} \leftarrow \theta, \bar{\phi} \leftarrow \phi$.
while Training is not over **do**
  Sample environment $M \sim \mathcal{P}$ with hypothesis $H^\star$, observe $x_1 \sim \rho$ and set $t \leftarrow 1$.
  repeat
    Execute action $a_t = \arg\max_a Q_\theta(\mathcal{D}_t, a)$ in M and observe x_{t+1}.
    Add partial trajectory $(\mathcal{D}_t, a_t, x_{t+1}, H^\star)$ to $\mathcal{B}$ and set $t \leftarrow t + 1$.
  until $a_{t-1} = a_{\text{stop}}$ or $t > N$.
  In the fixed confidence, update λ according to eq. (11).
  Sample batch $B \sim \mathcal{B}$ and update θ, ϕ using $\mathcal{L}_{\text{inf}}(B;\phi)$ (eq. (7)) and $\mathcal{L}_{\text{policy}}(B;\theta)$ (eq. (8) or eq. (9)).
  Every T_ϕ steps set $\bar{\phi} \leftarrow \phi$ (similarly, every T_θ steps set $\bar{\theta} \leftarrow \theta$).
end while

// Inference phase
Sample unknown environment $M \sim \mathcal{P}$.
Collect a trajectory $\mathcal{D}_N$ (or $\mathcal{D}_\tau$ in fixed confidence) according to a policy $\pi_t(\mathcal{D}_t) = \arg\max_a Q_\theta(\mathcal{D}_t, a)$, until $t = N$ (or $a_t = a_{\text{stop}}$).
Return $\hat{H}_N = \arg\max_H I_\phi(H|\mathcal{D}_N)$ (or $\hat{H}_\tau = \arg\max_H I_\phi(H|\mathcal{D}_\tau)$ in the fixed confidence)

Numerical Results

Numerical Results

We tested **ICPE** on a range of problems:

- Generalized search problems
- BAI-like problems in Bandit and MDPs (with structure, hidden information, etc...)

We look at **3** problems:

1. Can ICPE meta-learn binary search?
2. Can ICPE learn pixel-sampling for classification?
3. Can ICPE discover, and exploit, hidden information in BAI?

Numerical Results

We tested **ICPE** on a range of problems:

- Generalized search problems
- BAI-like problems in Bandit and MDPs (with structure, hidden information, etc...)

We look at **3** problems:

1. Can ICPE meta-learn binary search?
2. Can ICPE learn pixel-sampling for classification?
3. Can ICPE discover, and exploit, hidden information in BAI?

Numerical Results

We tested **ICPE** on a range of problems:

- Generalized search problems
- BAI-like problems in Bandit and MDPs (with structure, hidden information, etc...)

We look at **3** problems:

1. Can ICPE meta-learn binary search?
2. Can ICPE learn pixel-sampling for classification?
3. Can ICPE discover, and exploit, hidden information in BAI?

Numerical Results

We tested **ICPE** on a range of problems:

- Generalized search problems
- BAI-like problems in Bandit and MDPs (with structure, hidden information, etc...)

We look at **3** problems:

1. Can ICPE meta-learn binary search?
2. Can ICPE learn pixel-sampling for classification?
3. Can ICPE discover, and exploit, hidden information in BAI?

Numerical Results

We tested **ICPE** on a range of problems:

- Generalized search problems
- BAI-like problems in Bandit and MDPs (with structure, hidden information, etc...)

We look at **3** problems:

1. Can ICPE meta-learn binary search?
2. Can ICPE learn pixel-sampling for classification?
3. Can ICPE discover, and exploit, hidden information in BAI?

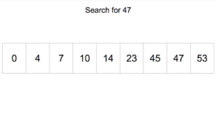
Binary Search

Can ICPE meta-learn binary search?

- Vector with K elements; need to find $H^\star \in \{1, \ldots, K\}$.
- Selecting $a \in \{1, \ldots, K\}$ yields a observation $x_t = -1$ or $x_t = +1$ (depending if $a < H^\star$ or not.

K	Min Accuracy	Mean Stop Time	Max Stop Time	$\log_2 K$
8	1.00	2.13 ± 0.12	3	3
16	1.00	2.93 ± 0.12	4	4
32	1.00	3.71 ± 0.15	5	5
64	1.00	4.50 ± 0.21	6	6
128	1.00	5.49 ± 0.23	7	7
256	1.00	6.61 ± 0.26	8	8

Table 1: ICPE performance on the binary search task as K increases.



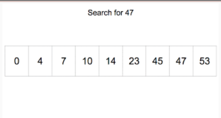
Can ICPE meta-learn binary search?

- Vector with K elements; need to find $H^\star \in \{1, \ldots, K\}$.
- Selecting $a \in \{1, \ldots, K\}$ yields a observation $x_t = -1$ or $x_t = +1$ (depending if $a < H^\star$ or not.

K	Min Accuracy	Mean Stop Time	Max Stop Time	$\log_2 K$
8	1.00	2.13 ± 0.12	3	3
16	1.00	2.93 ± 0.12	4	4
32	1.00	3.71 ± 0.15	5	5
64	1.00	4.50 ± 0.21	6	6
128	1.00	5.49 ± 0.23	7	7
256	1.00	6.61 ± 0.26	8	8

Table 1: ICPE performance on the binary search task as K increases.

Binary Search



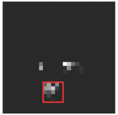
Can ICPE meta-learn binary search?

- Vector with K elements; need to find $H^\star \in \{1, \dots, K\}$.
- Selecting $a \in \{1, \dots, K\}$ yields a observation $x_t = -1$ or $x_t = +1$ (depending if $a < H^\star$ or not.

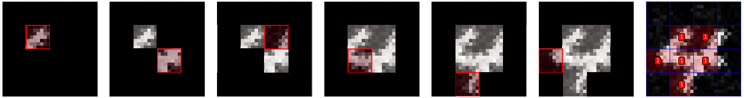
K	Min Accuracy	Mean Stop Time	Max Stop Time	$\log_2 K$
8	1.00	2.13 ± 0.12	3	3
16	1.00	2.93 ± 0.12	4	4
32	1.00	3.71 ± 0.15	5	5
64	1.00	4.50 ± 0.21	6	6
128	1.00	5.49 ± 0.23	7	7
256	1.00	6.61 ± 0.26	8	8

Table 1: ICPE performance on the binary search task as K increases.

Pixel Sampling as Generalized Search



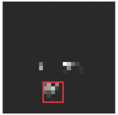
Can **ICPE** learn to select patch of pixels for classification?



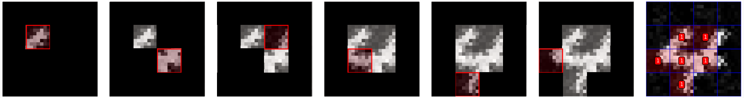
We already saw this example in the introduction.

But is **ICPE** really learning exploration strategies, or is it **just sampling at random**?

Pixel Sampling as Generalized Search



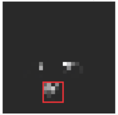
Can **ICPE** learn to select patch of pixels for classification?



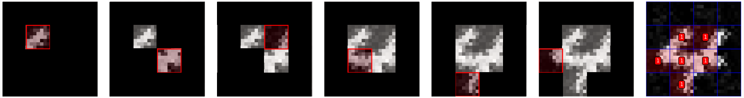
We already saw this example in the introduction.

But is **ICPE** really learning exploration strategies, or is it **just sampling at random**?

Pixel Sampling as Generalized Search

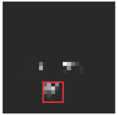


Can **ICPE** learn to select patch of pixels for classification?

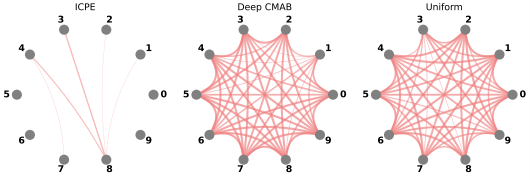


We already saw this example in the introduction.

But is **ICPE** really learning exploration strategies, or is it **just sampling at random**?

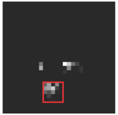


We compare **ICPE** with Deep Contextual Multi-Armed Bandit [Collier and Llorens, 2018].



- We compare region selection distributions across digit classes using pairwise chi-squared tests.
- A chord between two digits indicates that their distributions were not significantly different, with thicker chords representing higher p-values.

Pixel Sampling as Generalized Search



Agent	Accuracy	Avg. Regions Used
ICPE	0.91 ± 0.03	10.09 ± 0.11
Deep CMAB	0.66 ± 0.04	7.90 ± 0.09
Uniform	0.25 ± 0.04	10.42 ± 0.09

Table 2: Accuracy and performance (mean $\pm$ 95% CI)

Bandit Problems with Hidden Structure: Magic Arm Problem

This is a bandit model with Gaussian rewards and a **twist**:

- **One** of the arms encodes information about the index of the best arm through its mean reward value. We call this arm the magic arm.
- Let's say the index of the magic arm is $m \in \{1, \ldots, K\}$, fixed. Define the mean reward as $\mu_m = \phi(a^\star)$, for some invertible mapping ϕ.
- For $a \neq m$ we let the rewards be distributed according to $\mathcal{N}(\mu_a, (1 - \sigma_m)^2)$ with $\sigma_m \in (0, 1)$ being the standard deviation of arm $m \Rightarrow$ the smaller σ_m, the more likely we should sample arm m.

This is a bandit model with Gaussian rewards and a **twist**:

- **One** of the arms encodes information about the index of the best arm through its mean reward value. We call this arm the magic arm.
- Let's say the index of the magic arm is $m \in \{1, \dots, K\}$, fixed. Define the mean reward as $\mu_m = \phi(a^\star)$, for some invertible mapping ϕ.
- For $a \neq m$ we let the rewards be distributed according to $\mathcal{N}(\mu_a, (1-\sigma_m)^2)$ with $\sigma_m \in (0,1)$ being the standard deviation of arm $m \Rightarrow$ the smaller σ_m, the more likely we should sample arm m.

Bandit Problems with Hidden Structure: Magic Arm Problem


This is a bandit model with Gaussian rewards and a **twist**:

- **One** of the arms encodes information about the index of the best arm through its mean reward value. We call this arm the magic arm.
- Let's say the index of the magic arm is $m \in \{1, \ldots, K\}$, fixed. Define the mean reward as $\mu_m = \phi(a^\star)$, for some invertible mapping ϕ.
- For $a \neq m$ we let the rewards be distributed according to $\mathcal{N}(\mu_a, (1 - \sigma_m)^2)$ with $\sigma_m \in (0, 1)$ being the standard deviation of arm $m \Rightarrow$ the smaller σ_m, the more likely we should sample arm m.

Bandit Problems with Hidden Structure: Magic Arm Problem



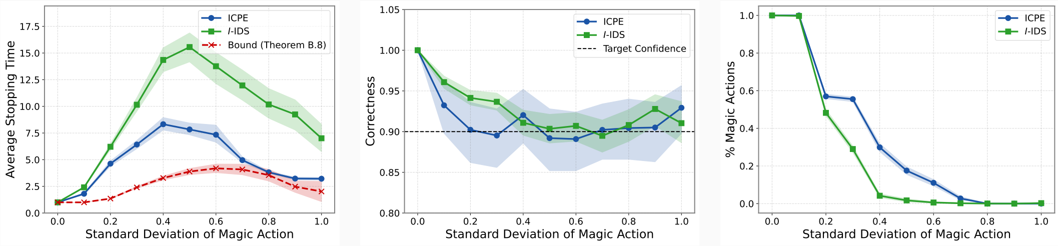


This is a bandit model with Gaussian rewards and a **twist**:

- **One** of the arms encodes information about the index of the best arm through its mean reward value. We call this arm the magic arm.
- Let's say the index of the magic arm is $m \in \{1, \ldots, K\}$, fixed. Define the mean reward as $\mu_m = \phi(a^\star)$, for some invertible mapping ϕ.
- For $a \neq m$ we let the rewards be distributed according to $\mathcal{N}(\mu_a, (1-\sigma_m)^2)$ with $\sigma_m \in (0,1)$ being the standard deviation of arm $m \Rightarrow$ the smaller σ_m, the more likely we should sample arm m.

Bandit Problems with Hidden Structure: Magic Arm Problem

We compared with the BAI version of Information Directed Sampling (IDS) [Russo and Van Roy, 2018] (based on posterior sampling, and we use the I-net of ICPE).

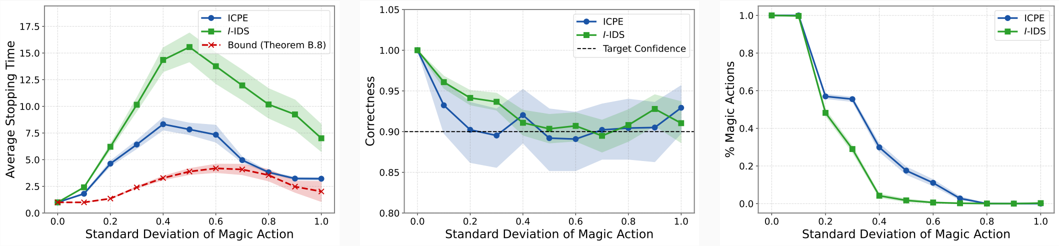


Left: average number of samples; **Middle**: average accuracy at the stopping time; **Right**: fraction of times the magic action was selected.

The distribution of the magic arm is $\mathcal{N}(\phi(a^\star), \sigma_m^2), \sigma_m \in (0,1)$. For $a \neq m$ is $\mathcal{N}(\mu_a, (1-\sigma_m)^2)$.

Bandit Problems with Hidden Structure: Magic Arm Problem

We compared with the BAI version of Information Directed Sampling (IDS) [Russo and Van Roy, 2018] (based on posterior sampling, and we use the I-net of ICPE).



Left: average number of samples; **Middle**: average accuracy at the stopping time; **Right**: fraction of times the magic action was selected.

The distribution of the magic arm is $\mathcal{N}(\phi(a^\star), \sigma_m^2), \sigma_m \in (0,1)$. For $a \neq m$ is $\mathcal{N}(\mu_a, (1-\sigma_m)^2)$.

Conclusions and Future Directions

Conclusions[10]

Thank you for reaching this point! **Plenty of questions are still open, and we look forward to collaborations:**

- What is a good neural architecture for sequential problems? And, how do we enable long horizons?
- We assumed access to a perfect simulator. What if there is some misspecification?
- Can we move from a Bayesian setting to a frequentist one? (i.e., adversarial).
- Plenty of theoretical questions still left unanswered (contact me for details!).

Thank you for listening!

[10] Credits to Flaticon.com for some of the logos used in this presentation.

Conclusions[10]

Thank you for reaching this point! **Plenty of questions are still open, and we look forward to collaborations:**

- What is a good neural architecture for sequential problems? And, how do we enable long horizons?
- We assumed access to a perfect simulator. What if there is some misspecification?
- Can we move from a Bayesian setting to a frequentist one? (i.e., adversarial).
- Plenty of theoretical questions still left unanswered (contact me for details!).

Thank you for listening!

[10] Credits to Flaticon.com for some of the logos used in this presentation.

Conclusions[10]

Thank you for reaching this point! **Plenty of questions are still open, and we look forward to collaborations:**

- What is a good neural architecture for sequential problems? And, how do we enable long horizons?
- We assumed access to a perfect simulator. What if there is some misspecification?
- Can we move from a Bayesian setting to a frequentist one? (i.e., adversarial).
- Plenty of theoretical questions still left unanswered (contact me for details!).

Thank you for listening!

[10] Credits to Flaticon.com for some of the logos used in this presentation.

Conclusions[10]

Thank you for reaching this point! **Plenty of questions are still open, and we look forward to collaborations:**

- What is a good neural architecture for sequential problems? And, how do we enable long horizons?
- We assumed access to a perfect simulator. What if there is some misspecification?
- Can we move from a Bayesian setting to a frequentist one? (i.e., adversarial).
- Plenty of theoretical questions still left unanswered (contact me for details!).

Thank you for listening!

[10] Credits to Flaticon.com for some of the logos used in this presentation.

Conclusions[10]

Thank you for reaching this point! **Plenty of questions are still open, and we look forward to collaborations:**

- What is a good neural architecture for sequential problems? And, how do we enable long horizons?
- We assumed access to a perfect simulator. What if there is some misspecification?
- Can we move from a Bayesian setting to a frequentist one? (i.e., adversarial).
- Plenty of theoretical questions still left unanswered (contact me for details!).

Thank you for listening!

[10] Credits to Flaticon.com for some of the logos used in this presentation.

Conclusions[10]

Thank you for reaching this point! **Plenty of questions are still open, and we look forward to collaborations:**

- What is a good neural architecture for sequential problems? And, how do we enable long horizons?
- We assumed access to a perfect simulator. What if there is some misspecification?
- Can we move from a Bayesian setting to a frequentist one? (i.e., adversarial).
- Plenty of theoretical questions still left unanswered (contact me for details!).

Thank you for listening!

[10] Credits to Flaticon.com for some of the logos used in this presentation.

Conclusions[10]

Thank you for reaching this point! **Plenty of questions are still open, and we look forward to collaborations:**

- What is a good neural architecture for sequential problems? And, how do we enable long horizons?
- We assumed access to a perfect simulator. What if there is some misspecification?
- Can we move from a Bayesian setting to a frequentist one? (i.e., adversarial).
- Plenty of theoretical questions still left unanswered (contact me for details!).

Thank you for listening!

[10] Credits to Flaticon.com for some of the logos used in this presentation.

References

Al Marjani, A., Garivier, A., and Proutiere, A. (2021).
Navigating to the best policy in markov decision processes.
Advances in Neural Information Processing Systems, 34:25852–25864.

Audibert, J.-Y. and Bubeck, S. (2010).
Best arm identification in multi-armed bandits.
In *COLT-23th Conference on learning theory-2010*, pages 13–p.

Chernoff, H. (1959).
Sequential design of experiments.
The Annals of Mathematical Statistics, 30(3):755–770.

Collier, M. and Llorens, H. U. (2018).
Deep contextual multi-armed bandits.
arXiv preprint arXiv:1807.09809.

References ii

Gan, K., Jia, S., and Li, A. (2021).
Greedy approximation algorithms for active sequential hypothesis testing.
Advances in Neural Information Processing Systems, 34:5012–5024.

Garivier, A. and Kaufmann, E. (2016).
Optimal best arm identification with fixed confidence.
In *Conference on Learning Theory*, pages 998–1027. PMLR.

Hero, A. O. and Cochran, D. (2011).
Sensor management: Past, present, and future.
IEEE Sensors Journal, 11(12):3064–3075.

Lai, T. L. and Robbins, H. (1985).
Asymptotically efficient adaptive allocation rules.
Advances in Applied Mathematics, 6(1):4–22.

Lattimore, T. and Szepesvári, C. (2020).
Bandit algorithms.
Cambridge University Press.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013).
Playing atari with deep reinforcement learning.
arXiv preprint arXiv:1312.5602.

Naghshvar, M. and Javidi, T. (2013).
Active sequential hypothesis testing.
The Annals of Statistics, 41(6):2703–2738.

Resnick, P. and Varian, H. R. (1997).
Recommender systems.
Communications of the ACM, 40(3):56–58.

References iv

Russo, A., Song, Y., and Pacchiano, A. (2025).
Pure exploration with feedback graphs.
In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR.

Russo, D. and Van Roy, B. (2018).
Learning to optimize via information-directed sampling.
Operations Research, 66(1):230–252.

Appendix

ICPE: Fixed Budget vs Fixed Confidence

These two objectives capture the main operational modes of pure exploration: "stop when certain" and "maximize accuracy over a fixed sampling budget".

- Note that we did not impose any restriction of the problem, except for the prior distribution. Compared to classical results, our setting generalizes MDP and Bandit problems.
- Classically the inference rule I is a maximum likelihood estimator. However, it's hard to compute for complex models. That's why we also optimize over inference rules.

ICPE: Fixed Budget vs Fixed Confidence

These two objectives capture the main operational modes of pure exploration: "stop when certain" and "maximize accuracy over a fixed sampling budget".

- Note that we did not impose any restriction of the problem, except for the prior distribution. Compared to classical results, our setting generalizes MDP and Bandit problems.
- Classically the inference rule I is a maximum likelihood estimator. However, it's hard to compute for complex models. That's why we also optimize over inference rules.

ICPE: Fixed Budget vs Fixed Confidence

These two objectives capture the main operational modes of pure exploration: "stop when certain" and "maximize accuracy over a fixed sampling budget".

- Note that we did not impose any restriction of the problem, except for the prior distribution. Compared to classical results, our setting generalizes MDP and Bandit problems.
- Classically the inference rule I is a maximum likelihood estimator. However, it's hard to compute for complex models. That's why we also optimize over inference rules.