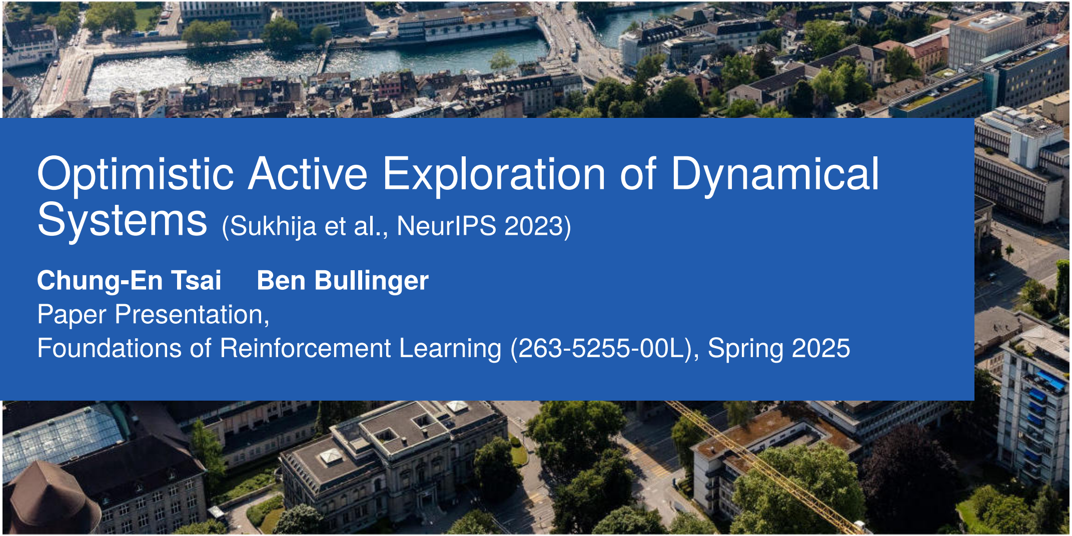
ETH zürich

ETH Zürich
Department of Computer Science



Optimistic Active Exploration of Dynamical Systems (Sukhija et al., NeurIPS 2023)

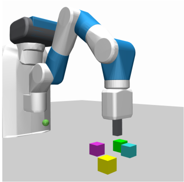
Chung-En Tsai **Ben Bullinger**

Paper Presentation,
Foundations of Reinforcement Learning (263-5255-00L), Spring 2025

Outline

Outline

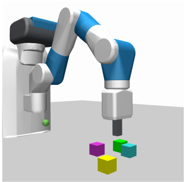
Motivation: Beyond Task-Specific Models



Traditional Model-Based RL:

- Learn a model for one specific reward
- Biased exploration of the state-action space
- May not generalize to new tasks

Motivation: Beyond Task-Specific Models



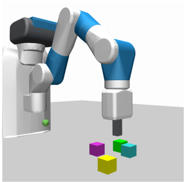
Traditional Model-Based RL:

- Learn a model for one specific reward
- Biased exploration of the state-action space
- May not generalize to new tasks

A More Ambitious Goal:

- Learn the underlying **dynamical system**
- Global exploration
- **Zero-shot** transfer to new tasks

Motivation: Beyond Task-Specific Models



Traditional Model-Based RL:

- Learn a model for one specific reward
- Biased exploration of the state-action space
- May not generalize to new tasks

A More Ambitious Goal:

- Learn the underlying **dynamical system**
- Global exploration
- **Zero-shot** transfer to new tasks

Can we ***efficiently*** *learn accurate models that generalize to* ***any*** *downstream task?*

Outline

Problem Formulation (1/2)

Given an unknown discrete-time dynamical system $\boldsymbol{f}^\star : \mathcal{S} \times \mathcal{A} \to \mathcal{S} \subseteq \mathbb{R}^d$:

$$\boldsymbol{s}_{t+1} = \boldsymbol{f}^\star(\boldsymbol{s}_t, \boldsymbol{a}_t) + \boldsymbol{w}_t, \quad \boldsymbol{w}_t \sim \mathcal{N}(0, I).$$

Goal: Design algorithms for learning $\hat{\boldsymbol{f}} \approx \boldsymbol{f}^\star$ in episodic setting that is good for any reward functions.

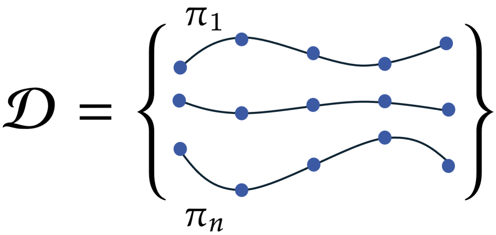
Introduction video by Bhavya Sukhija. NeurIPS 2023.

Problem Formulation (1/2)

Given an unknown discrete-time dynamical system $\boldsymbol{f}^\star : \mathcal{S} \times \mathcal{A} \to \mathcal{S} \subseteq \mathbb{R}^d$:

$$\boldsymbol{s}_{t+1} = \boldsymbol{f}^\star(\boldsymbol{s}_t, \boldsymbol{a}_t) + \boldsymbol{w}_t, \quad \boldsymbol{w}_t \sim \mathcal{N}(0, I).$$

Goal: Design algorithms for learning $\hat{f} \approx \boldsymbol{f}^\star$ in episodic setting that is good for any reward functions.



Introduction video by Bhavya Sukhija. NeurIPS 2023.

Problem Formulation (2/2)

Our problem can be divided into two parts:

- Exploration (our focus!): Collecting data by executing policies.
- Estimation: Building model from data.

Problem Formulation (2/2)

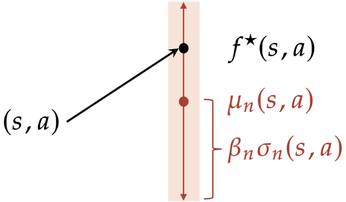
Our problem can be divided into two parts:

- Exploration (our focus!): Collecting data by executing policies.
- Estimation: Building model from data.

Well-Calibration Assumption

Given n data, we can construct confidence intervals:

$$|\boldsymbol{f}_i^\star(\boldsymbol{s}, \boldsymbol{a}) - \boldsymbol{\mu}_{n,i}(\boldsymbol{s}, \boldsymbol{a})| \leqslant \beta_n \boldsymbol{\sigma}_{n,i}(\boldsymbol{s}, \boldsymbol{a}) \quad \forall i \in [d], \text{ w.h.p.,}$$



Outline

Related Work: Active Learning

	Bandit Optimization	Our Problem
Goal	$\max_x f^\star(x)$	Find $\hat{f} \approx f^\star$
$f^\star$	Objective	Dynamics
State space	No	Continuous
Planning	No	Yes

N. Srinivas et al. "Gaussian process optimization in the bandit setting: No regret and experimental design." ICML 2010. (The 1st presentation today!)

Related Work: Active Learning

	Bandit Optimization	Our Problem
Goal	$\max_x f^\star(x)$	Find $\hat{f} \approx f^\star$
$f^\star$	Objective	Dynamics
State space	No	Continuous
Planning	No	Yes
Technique	Optimism in the face of uncertainty	

N. Srinivas et al. “Gaussian process optimization in the bandit setting: No regret and experimental design.” ICML 2010. (The 1st presentation today!)

Related Work: Maximum Entropy RL

What about collecting data by maximum entropy or entropy-regularized RL?

$$\max_{\boldsymbol{\pi}} \mathsf{H}\left(d_{\mu}^{\boldsymbol{\pi}}\right), \quad \max_{\boldsymbol{\pi}} \mathsf{E}\left[\sum_{t=0}^{\infty} \gamma^{t}\left(r(\boldsymbol{s}_t, \boldsymbol{a}_t) - \beta \log \pi(\boldsymbol{a}_t|\boldsymbol{s}_t)\right) \middle| \boldsymbol{s}_0 \sim \mu, \boldsymbol{\pi}\right].$$

- No guarantee for arbitrary reward functions in downstream tasks.

E. Hazan et al. "Provably efficient maximum entropy exploration." ICML 2019. (Previous presentation.)
B. Eysenbach and S. Levine. "Maximum entropy RL (provably) solves some robust RL problems." ICLR 2022.

Related Work: Reward-Free RL

- Study this problem in tabular settings or impose structural assumptions on values function and transitions.
- No practical algorithms.

C. Jin et al. "Reward-free exploration for reinforcement learning." ICML 2020.
R. Wang et al. "On reward-free reinforcement learning with linear function approximation." NeurIPS 2020.
J. Chen et al. "On the statistical efficiency of reward-free exploration in non-linear RL." NeurIPS 2022.

Contributions

- Propose OpAX, a practical algorithm for learning non-linear dynamics with continuous state-action spaces and without structural assumptions.
- In this setting, OpAX is the first algorithm to have a theoretical guarantee for a rich family of nonlinear dynamics.

Outline

OpAX: Overview

*Learn dynamics by **maximizing information gain**, not rewards*

OpAX: Overview

*Learn dynamics by **maximizing information gain**, not rewards*

	Traditional Model-Based RL	OpAX
Objective	$\max_\pi \mathbb{E}[R(\pi)]$	$\max_\pi I(f^*; \mathcal{D})$
Exploration	Reward-driven	Uncertainty-driven
Convergence to $f^\star$	No guarantees	$\boldsymbol{\sigma}_n(\boldsymbol{s}, \boldsymbol{a}) \to 0$ (GP)
Generalization	No guarantees	Zero-shot (GP)

OpAX (1/4): Information-Theoretic Objective

Ideal Goal: Maximize mutual information between dynamics $\boldsymbol{f}^{\star}$ and collected data

OpAX (1/4): Information-Theoretic Objective

Ideal Goal: Maximize mutual information between dynamics $\boldsymbol{f}^\star$ and collected data

Greedy Episodic Approximation: At episode n, choose policy maximizing expected information gain:

$$\boldsymbol{\pi}_n^* = \underset{\boldsymbol{\pi} \in \Pi}{\arg\max}\, \mathsf{E}_{\tau^{\boldsymbol{\pi}}} \left[\mathrm{I}\left(\boldsymbol{f}^\star_{\tau^{\boldsymbol{\pi}}} ; \tau^{\boldsymbol{\pi}} \mid \mathcal{D}_{1:n-1} \right) \right]$$

where:

- $\boldsymbol{f}^\star_{\tau^{\boldsymbol{\pi}}} = (\boldsymbol{f}^\star(\boldsymbol{s}_0, \boldsymbol{a}_0), \ldots, \boldsymbol{f}^\star(\boldsymbol{s}_{T-1}, \boldsymbol{a}_{T-1}))$: true dynamics along trajectory
- $\mathcal{D}_{1:n-1}$: data from previous episodes

OpAX (1/4): Information-Theoretic Objective

Ideal Goal: Maximize mutual information between dynamics $\boldsymbol{f}^\star$ and collected data

Greedy Episodic Approximation: At episode n, choose policy maximizing expected information gain:

$$\boldsymbol{\pi}_n^* = \underset{\boldsymbol{\pi}\in\Pi}{\arg\max}\, \mathrm{E}_{\tau^{\boldsymbol{\pi}}} \left[\mathrm{I}\left(\boldsymbol{f}^\star_{\tau^{\boldsymbol{\pi}}} ; \tau^{\boldsymbol{\pi}} \mid \mathcal{D}_{1:n-1} \right) \right]$$

where:

- $\boldsymbol{f}^\star_{\tau^{\boldsymbol{\pi}}} = (\boldsymbol{f}^\star(\boldsymbol{s}_0, \boldsymbol{a}_0), \ldots, \boldsymbol{f}^\star(\boldsymbol{s}_{T-1}, \boldsymbol{a}_{T-1}))$: true dynamics along trajectory
- $\mathcal{D}_{1:n-1}$: data from previous episodes

Challenge

This objective is intractable in general. Need tractable approximation.

OpAX (2/4): Tractable Upper Bound

Lemma (Information Gain Upper Bound)

For Gaussian noise $\boldsymbol{w}_t \sim \mathcal{N}(0, I)$ *and epistemic uncertainty* $\boldsymbol{\sigma}_{n-1}$*:*

$$I\left(\boldsymbol{f}^{\star}_{\boldsymbol{\tau}^{\pi}}; \boldsymbol{\tau}^{\pi} \mid \mathcal{D}_{1:n-1}\right) \leqslant \frac{1}{2} \sum_{t=0}^{T-1} \sum_{j=1}^{d} \log\left(1 + \boldsymbol{\sigma}^2_{n-1,j}(\boldsymbol{s}_t, \boldsymbol{a}_t)\right)$$

OpAX (2/4): Tractable Upper Bound

Lemma (Information Gain Upper Bound)

For Gaussian noise $\boldsymbol{w}_t \sim \mathcal{N}(0, I)$ and epistemic uncertainty $\boldsymbol{\sigma}_{n-1}$:

$$I\left(\boldsymbol{f}^\star_{\boldsymbol{\tau}^\pi}; \boldsymbol{\tau}^\pi \mid \mathcal{D}_{1:n-1}\right) \leqslant \frac{1}{2} \sum_{t=0}^{T-1} \sum_{j=1}^{d} \log\left(1 + \boldsymbol{\sigma}^2_{n-1,j}(\boldsymbol{s}_t, \boldsymbol{a}_t)\right)$$

Tractable Exploration Objective: Maximize the upper bound

$$\boldsymbol{\pi}^*_n = \underset{\boldsymbol{\pi} \in \Pi}{\arg\max}\, \mathbb{E}_{\boldsymbol{\tau}^\pi}\left[\sum_{t=0}^{T-1} \sum_{j=1}^{d} \log\left(1 + \boldsymbol{\sigma}^2_{n-1,j}(\boldsymbol{s}_t, \boldsymbol{a}_t)\right)\right]$$

subject to: $\boldsymbol{s}_{t+1} = \boldsymbol{f}^\star(\boldsymbol{s}_t, \boldsymbol{a}_t) + \boldsymbol{w}_t$.

OpAX (2/4): Tractable Upper Bound

Lemma (Information Gain Upper Bound)

For Gaussian noise $\boldsymbol{w}_t \sim \mathcal{N}(0, I)$ *and epistemic uncertainty* $\boldsymbol{\sigma}_{n-1}$*:*

$$I\left(\boldsymbol{f}^{\star}_{\boldsymbol{\tau}^{\pi}}; \boldsymbol{\tau}^{\pi} \mid \mathcal{D}_{1:n-1}\right) \leqslant \frac{1}{2} \sum_{t=0}^{T-1} \sum_{j=1}^{d} \log\left(1 + \boldsymbol{\sigma}^2_{n-1,j}(\boldsymbol{s}_t, \boldsymbol{a}_t)\right)$$

Tractable Exploration Objective: Maximize the upper bound

$$\boldsymbol{\pi}^*_n = \operatorname*{arg\,max}_{\boldsymbol{\pi} \in \Pi} \mathrm{E}_{\boldsymbol{\tau}^{\pi}} \left[\sum_{t=0}^{T-1} \sum_{j=1}^{d} \log\left(1 + \boldsymbol{\sigma}^2_{n-1,j}(\boldsymbol{s}_t, \boldsymbol{a}_t)\right) \right]$$

subject to: $\boldsymbol{s}_{t+1} = \boldsymbol{f}^{\star}(\boldsymbol{s}_t, \boldsymbol{a}_t) + \boldsymbol{w}_t$.

Collect data where model is most uncertain $\Rightarrow$ maximal information gain

OpAX (3/4): Hallucination Policy

Problem: The planning objective requires knowing $\boldsymbol{f}^{\star}$, but $\boldsymbol{f}^{\star}$ is unknown!

[1] K. Chua et al. "Deep reinforcement learning in a handful of trials using probabilistic dynamics models." NeurIPS 2018.
[2] M. Simchowitz et al. "Naive exploration is optimal for online lqr." ICML 2020.

OpAX (3/4): Hallucination Policy

Problem: The planning objective requires knowing $\boldsymbol{f}^\star$, but $\boldsymbol{f}^\star$ is unknown!

Naive Approach: Use mean estimate $\boldsymbol{\mu}_{n-1}$ instead of $\boldsymbol{f}^\star$

[1] K. Chua et al. "Deep reinforcement learning in a handful of trials using probabilistic dynamics models." NeurIPS 2018.

[2] M. Simchowitz et al. "Naive exploration is optimal for online lqr." ICML 2020.

OpAX (3/4): Hallucination Policy

Problem: The planning objective requires knowing $\boldsymbol{f}^\star$, but $\boldsymbol{f}^\star$ is unknown!

Naive Approach: Use mean estimate $\boldsymbol{\mu}_{n-1}$ instead of $\boldsymbol{f}^\star$

- ✗ Susceptible to model biases[1]
- ✗ Provably optimal only for linear systems[2]

[1] K. Chua et al. "Deep reinforcement learning in a handful of trials using probabilistic dynamics models." NeurIPS 2018.

[2] M. Simchowitz et al. "Naive exploration is optimal for online lqr." ICML 2020.

OpAX (3/4): Hallucination Policy

Problem: The planning objective requires knowing $\boldsymbol{f}^\star$, but $\boldsymbol{f}^\star$ is unknown!

Naive Approach: Use mean estimate $\boldsymbol{\mu}_{n-1}$ instead of $\boldsymbol{f}^\star$

- ✗ Susceptible to model biases[1]
- ✗ Provably optimal only for linear systems[2]

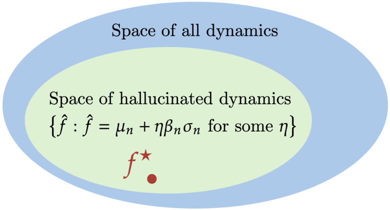
OpAX Solution:

Optimism in the face of uncertainty

Use a **hallucination policy** $\boldsymbol{\eta}$ to select optimistic dynamics

$$\hat{\boldsymbol{f}}_j(\boldsymbol{s}, \boldsymbol{a}) = \boldsymbol{\mu}_{n-1,j}(\boldsymbol{s}, \boldsymbol{a}) + \beta_{n-1}\boldsymbol{\sigma}_{n-1,j}(\boldsymbol{s}, \boldsymbol{a}) \cdot \boldsymbol{\eta}_j(\boldsymbol{s})$$

where $\boldsymbol{\eta} : \mathcal{S} \to [-1, 1]^d$.



[1] K. Chua et al. "Deep reinforcement learning in a handful of trials using probabilistic dynamics models." NeurIPS 2018.

[2] M. Simchowitz et al. "Naive exploration is optimal for online lqr." ICML 2020.

OpAX (4/4): Optimistic Planning

OpAX Optimistic Objective

$$\boldsymbol{\pi}_n, \boldsymbol{\eta}_n = \underset{\boldsymbol{\pi}\in\Pi, \boldsymbol{\eta}\in\Xi}{\arg\max}\, \mathbb{E}_{\boldsymbol{\tau}^{\boldsymbol{\pi},\boldsymbol{\eta}}}\left[\sum_{t=0}^{T-1}\sum_{j=1}^{d}\log\left(1+\boldsymbol{\sigma}^2_{n-1,j}(\hat{\boldsymbol{s}}_t, \boldsymbol{\pi}(\hat{\boldsymbol{s}}_t))\right)\right]$$

subject to:

$$\hat{\boldsymbol{s}}_{t+1} = \boldsymbol{\mu}_{n-1}(\hat{\boldsymbol{s}}_t, \boldsymbol{\pi}(\hat{\boldsymbol{s}}_t)) + \beta_{n-1}\boldsymbol{\sigma}_{n-1}(\hat{\boldsymbol{s}}_t, \boldsymbol{\pi}(\hat{\boldsymbol{s}}_t)) \cdot \boldsymbol{\eta}(\hat{\boldsymbol{s}}_t) + \boldsymbol{w}_t$$

where $\Xi = \{\boldsymbol{\eta} : \mathcal{S} \to [-1, 1]^d\}$ is the space of **hallucination policies**.

- $\eta(\hat{\boldsymbol{s}}_t)$ picks the most optimistic transition within confidence bounds

OpAX (4/4): Optimistic Planning

OpAX Optimistic Objective

$$\boldsymbol{\pi}_n, \boldsymbol{\eta}_n = \underset{\boldsymbol{\pi}\in\Pi, \boldsymbol{\eta}\in\Xi}{\arg\max}\, \mathrm{E}_{\boldsymbol{\tau}^{\boldsymbol{\pi},\boldsymbol{\eta}}}\left[\sum_{t=0}^{T-1}\sum_{j=1}^{d}\log\left(1+\boldsymbol{\sigma}^2_{n-1,j}(\hat{\boldsymbol{s}}_t, \boldsymbol{\pi}(\hat{\boldsymbol{s}}_t))\right)\right]$$

subject to:

$$\hat{\boldsymbol{s}}_{t+1} = \boldsymbol{\mu}_{n-1}(\hat{\boldsymbol{s}}_t, \boldsymbol{\pi}(\hat{\boldsymbol{s}}_t)) + \beta_{n-1}\boldsymbol{\sigma}_{n-1}(\hat{\boldsymbol{s}}_t, \boldsymbol{\pi}(\hat{\boldsymbol{s}}_t)) \cdot \boldsymbol{\eta}(\hat{\boldsymbol{s}}_t) + \boldsymbol{w}_t$$

where $\Xi = \{\boldsymbol{\eta} : \mathcal{S} \to [-1, 1]^d\}$ is the space of **hallucination policies**.

- $\eta(\hat{\boldsymbol{s}}_t)$ picks the most optimistic transition within confidence bounds

Implementation

Solve as optimal control with extended action space $(\boldsymbol{\pi}, \boldsymbol{\eta})$.

Algorithm

OpAX: Optimistic Active Exploration

Init: Statistical model $(\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0, \beta_0)$

for episode $n = 1, \ldots, N$ **do**

$$\boldsymbol{\pi}_n = \arg\max_{\pi \in \Pi} \max_{\eta \in \Xi} \mathbb{E}\left[\sum_{t=0}^{T-1} \sum_{j=1}^{d} \log\left(1 + \boldsymbol{\sigma}^2_{n-1,j}(\boldsymbol{s}_t, \boldsymbol{a}_t)\right)\right]$$ ➤ Prepare policy

$\mathcal{D}_n \leftarrow \textsc{Rollout}(\boldsymbol{\pi}_n)$ ➤ Measure

Update $(\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n, \beta_n) \leftarrow \mathcal{D}_{1:n}$ ➤ Update model

end for

Algorithm

OpAX: Optimistic Active Exploration

Init: Statistical model $(\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0, \beta_0)$

for episode $n = 1, \ldots, N$ **do**

$$\pi_n = \arg\max_{\pi \in \Pi} \max_{\eta \in \Xi} \mathbb{E}\left[\sum_{t=0}^{T-1} \sum_{j=1}^{d} \log\left(1 + \boldsymbol{\sigma}^2_{n-1,j}(\boldsymbol{s}_t, \boldsymbol{a}_t)\right)\right] \quad ➤ \text{Prepare policy}$$

$\mathcal{D}_n \leftarrow \text{Rollout}(\pi_n)$ ➤ Measure

Update $(\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n, \beta_n) \leftarrow \mathcal{D}_{1:n}$ ➤ Update model

end for

Key Components:

- **Information reward**: $\log(1 + \boldsymbol{\sigma}^2)$
- **Optimistic planning**: Max over η

Algorithm

OpAX: Optimistic Active Exploration

Init: Statistical model $(\boldsymbol{\mu}_0, \boldsymbol{\sigma}_0, \beta_0)$

for episode $n = 1, \ldots, N$ **do**

$$\boldsymbol{\pi}_n = \arg\max_{\boldsymbol{\pi} \in \Pi} \max_{\boldsymbol{\eta} \in \Xi} \mathbb{E}\left[\sum_{t=0}^{T-1} \sum_{j=1}^{d} \log\left(1 + \boldsymbol{\sigma}^2_{n-1,j}(\boldsymbol{s}_t, \boldsymbol{a}_t)\right)\right] \quad \text{➤ Prepare policy}$$

$$\mathcal{D}_n \leftarrow \textsc{Rollout}(\boldsymbol{\pi}_n) \quad \text{➤ Measure}$$

$$\text{Update } (\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n, \beta_n) \leftarrow \mathcal{D}_{1:n} \quad \text{➤ Update model}$$

end for

Key Components:

- **Information reward**: $\log(1 + \boldsymbol{\sigma}^2)$
- **Optimistic planning**: Max over $\boldsymbol{\eta}$

Implementation:

- MPC for short horizons
- SAC for long horizons

Outline

Assumptions

Regularity of the Dynamics

$f^\star$ lies in a RKHS with kernel k and $\|f^\star\| \leqslant 1$.

N. Srinivas et al. "Gaussian process optimization in the bandit setting: No regret and experimental design." ICML 2010. (The 1st presentation today!)

Assumptions

Regularity of the Dynamics

$f^\star$ lies in a RKHS with kernel k and $\|f^\star\| \leqslant 1$.

Lemma

GP$(0, k)$ *is well-calibrated.*

Recall

$$\gamma_n(k) = \max_{\mathcal{D}_1, \dots, \mathcal{D}_n} \log\det(I + K_n).$$

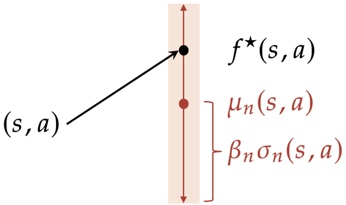
N. Srinivas et al. "Gaussian process optimization in the bandit setting: No regret and experimental design." ICML 2010. (The 1st presentation today!)

Main Guarantee

Theorem (Informal)

With high probability,

$$\max_{\pi} \mathbb{E}_{\pi}\left[\underbrace{\max_{t} \|\boldsymbol{\sigma}_n(\boldsymbol{s}_t, \boldsymbol{a}_t)\|_2^2}_{\text{maximum uncertainty along the trajectory}}\right] \leqslant \textit{poly}(T) \cdot \frac{\gamma_n(k)}{\sqrt{n}}, \quad \forall n \in \mathbb{N}.$$

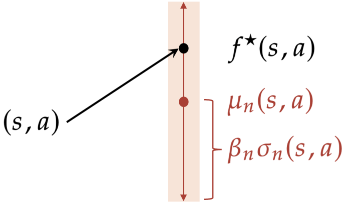


Main Guarantee

Theorem (Informal)

With high probability,

$$\max_{\pi} \mathrm{E}_{\pi}\left[\underbrace{\max_{t}\|\boldsymbol{\sigma}_n(\boldsymbol{s}_t, \boldsymbol{a}_t)\|_2^2}_{\text{maximum uncertainty along the trajectory}}\right] \leqslant \textit{poly}(T) \cdot \frac{\gamma_n(k)}{\sqrt{n}}, \quad \forall n \in \mathbb{N}.$$



For linear and RBF kernels, $\gamma_n \approx \mathrm{polylog}(nT)$ and $\lim_{n\to\infty} \boldsymbol{\sigma}_n(s,a) \to 0$.

Zero-Shot Performance

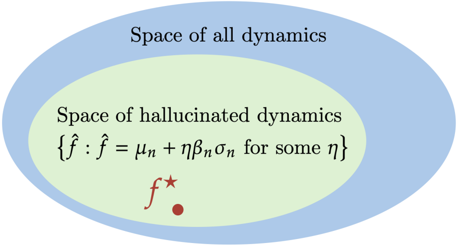
Theorem (Informal)

Consider an RL problem with a known bounded reward:

$$\arg\max_{\pi} R(\pi), \quad \boldsymbol{s}_{t+1} = \boldsymbol{f}^{\star}(\boldsymbol{s}_t, \boldsymbol{a}_t) + \boldsymbol{w}_t.$$

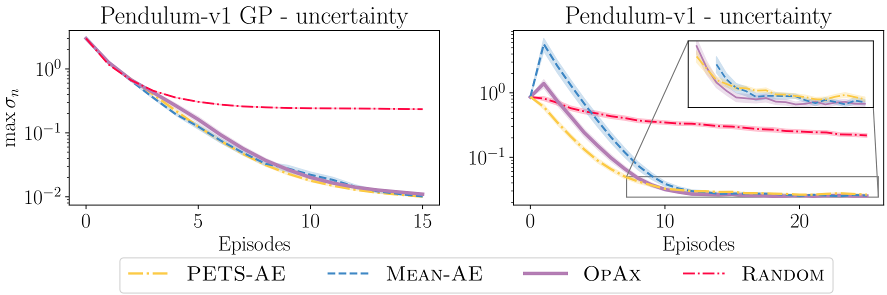
If $n \gtrsim \text{poly}(d, T, \varepsilon^{-1}, \gamma_n(k))$*, then* $\hat{\pi}$ *is* ε*-optimal, where*

$$\hat{\pi} \in \arg\max_{\boldsymbol{\pi}} \min_{\boldsymbol{\eta}} R(\boldsymbol{\pi}, \boldsymbol{\eta}), \quad \boldsymbol{s}_{t+1} = \boldsymbol{\mu}_n(\boldsymbol{s}_t, \boldsymbol{a}_t) + \boldsymbol{\eta}(s_t)\beta_n \boldsymbol{\sigma}_n(s_t, a_t) + \boldsymbol{w}_t.$$

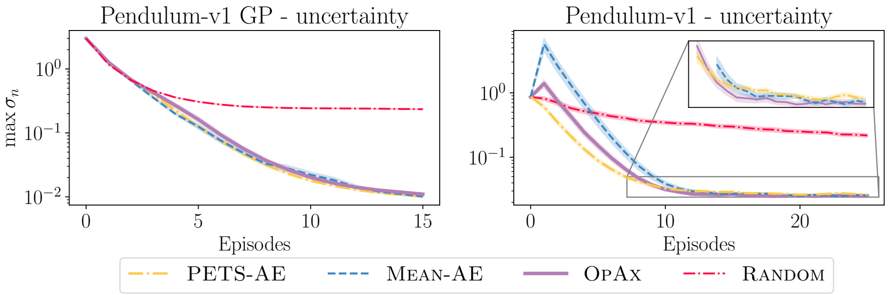


Outline

Experiments: Epistemic Uncertainty Reduction



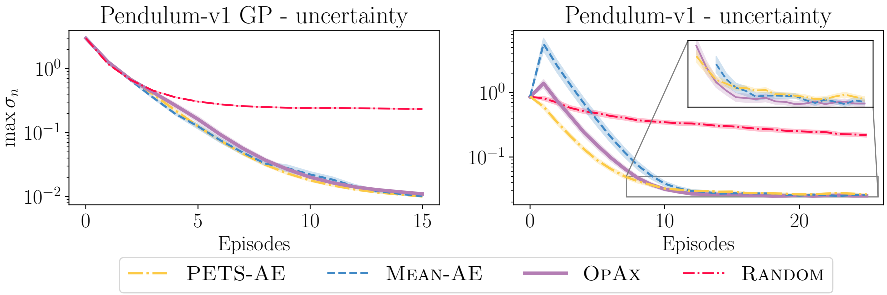
Experiments: Epistemic Uncertainty Reduction



Key Results:

- **Efficient exploration:** Active methods reduces uncertainty faster than random baseline
- **Optimism:** OpAX provides edge over mean planning

Experiments: Epistemic Uncertainty Reduction



Key Results:

- **Efficient exploration:** Active methods reduces uncertainty faster than random baseline
- **Optimism:** OpAX provides edge over mean planning

Experiments validate theoretical results

Experiments: Downstream Tasks

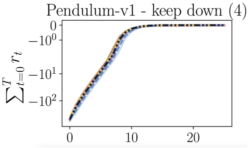
- `OpAX` performs **on par** with `H-UCRL`[1] on trained tasks
- `OpAX` **outperforms** `H-UCRL` on some unseen tasks

[1] S. Curi et al. "Efficient model-based reinforcement learning through optimistic policy search and planning." NeurIPS 2020.

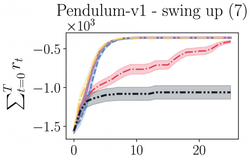
Experiments: Downstream Tasks

- `OpAX` performs **on par** with `H-UCRL`[1] on trained tasks
- `OpAX` **outperforms** `H-UCRL` on some unseen tasks

Evaluation (■: `H-UCRL`, ■: `OpAX`, ■: Random):



(a) Trained task



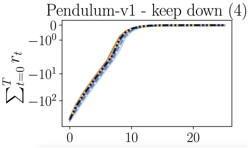
(b) Zero-shot task

[1] S. Curi et al. "Efficient model-based reinforcement learning through optimistic policy search and planning." NeurIPS 2020.

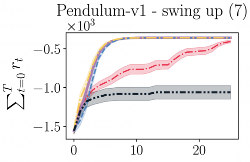
Experiments: Downstream Tasks

- `OpAX` performs **on par** with `H-UCRL`[1] on trained tasks
- `OpAX` **outperforms** `H-UCRL` on some unseen tasks

Evaluation (■: `H-UCRL`, ■: `OpAX`, ■: Random):



(a) Trained task



(b) Zero-shot task

Zero-shot swing up task:

(play)

(c) `H-UCRL`

(play)

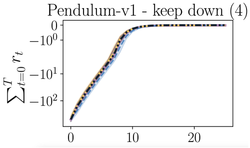
(d) `OpAx`

[1] S. Curi et al. "Efficient model-based reinforcement learning through optimistic policy search and planning." NeurIPS 2020.

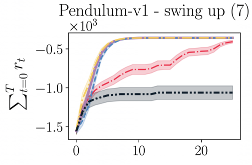
Experiments: Downstream Tasks

- `OpAX` performs **on par** with `H-UCRL`[1] on trained tasks
- `OpAX` **outperforms** `H-UCRL` on some unseen tasks

Evaluation (■: `H-UCRL`, ■: `OpAX`, ■: Random):



(a) Trained task



(b) Zero-shot task

Zero-shot swing up task:

(play)

(c) `H-UCRL`

(play)

(d) `OpAx`

More experiments in the paper.

[1] S. Curi et al. "Efficient model-based reinforcement learning through optimistic policy search and planning." NeurIPS 2020.

Outline

Conclusions

	Traditional Model-Based RL	**OpAX**
Objective	$\max_\pi \mathrm{E}[R(\pi)]$	$\max_\pi I(f^*; \mathcal{D})$
Exploration	Reward-driven	Uncertainty-driven
Convergence to $f^\star$	No guarantees	$\boldsymbol{\sigma}_n(\boldsymbol{s}, \boldsymbol{a}) \to 0$ (GP)
Generalization	No guarantees	Zero-shot (GP)

OpAX

Combines **active learning** with RL and achieves theoretical guarantees for learning many continuous nonlinear systems.

Open Challenges and Future Work

Key Limitations:

- **Computational cost:** Solving (π_n, η_n) is expensive
- **Model requirements:** Needs well-calibrated uncertainty
- **Known noise:** Guarantees only hold for (sub-)Gaussian noise.

Promising Directions:

- **Robust models:** Handle model misspecification
- **Beyond MDPs:** POMDPs, continuous time
- **Practical deployment:** Real-world robotics