

Agentic RL: Optimizing for More Than Just Performance

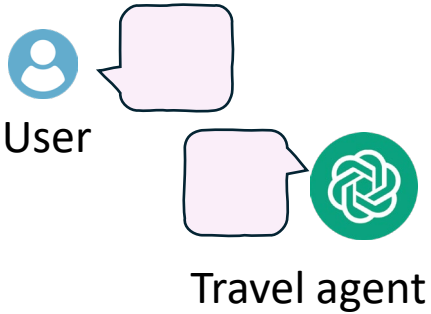
Laixi Shi

Johns Hopkins University

Department of Electrical and Computer Engineering & DSAI Institute



Agentic RL: Optimizing for More Than Just Performance



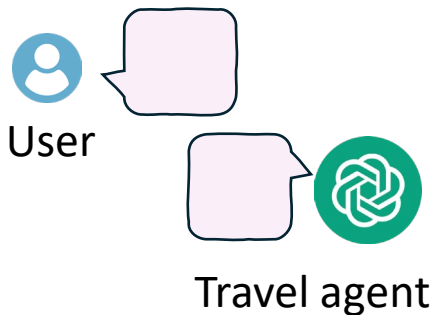
Task performance



User satisfaction

I: Agentic RL for multi-turn agent-user interaction:
behavior regularization → Pareto optimal

Agentic RL: Optimizing for More Than Just Performance

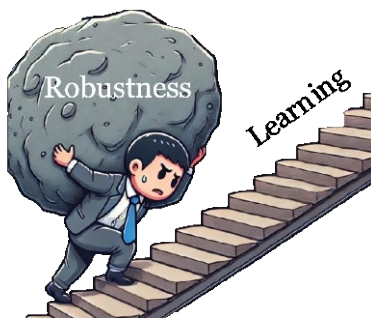


Task performance



User satisfaction

I: Agentic RL for multi-turn agent-user interaction:
behavior regularization → Pareto optimal



Robust performance



Data efficiency

II: Robust RL
without sacrificing data efficiency, even enhancing it

Multi-Turn Agent-User Interaction

Task: Book Travel Plan



User

Help me book a round trip from PIT to SFO from 04 April to 11 April, my credit card is XXX, ...

Yes, Apr 04 to Apr 11, 1 adult, Economy. Prefer morning flights if possible.

Yes, please book that.

Great, thank you!



Travel agent

1 Understand & collect details



Extract: PIT → SFO
Apr 04 – Apr 11
Round-trip, 1 traveler

Sure! I can book that for you.
Could you confirm:
• Exact dates: Apr 04 to Apr 11?
• Traveler: 1 adult?
• Cabin class: Economy?
• Any time preferences?

2 Search flights



Search flights
PIT → SFO (Apr 04)
SFO → PIT (Apr 11)

Here are the best options I found:
1) 6:00 AM PIT → 8:35 AM SFO
7:30 AM SFO → 3:55 PM PIT
\$358

Would you like me to book this one?

3 Hold & book



Hold seats → Book with
payment (card ending XXX)

Booking confirmed! 🎉

Round-trip: PIT → SFO (Apr 04)
SFO → PIT (Apr 11)
Confirmation: ABC123
Total charged: \$358
A receipt has been sent to your email.

4 Confirmation



Send confirmation email
& receipt

You're welcome! Safe travels! ✈️
Let me know if you need anything else.

Multi-Turn Agent-User Interaction

Personalized Task: Book Travel Plan



User

Help me book a round trip from PIT to SFO from 04 April to 11 April, my credit card is XXX, ...

Yes, Apr 04 to Apr 11,
1 adult, Economy.

Prefer morning flights if possible.

Yes, please book that.

Great, thank you!



Travel agent

1 Understand & collect details



Extract: PIT → SFO
Apr 04 – Apr 11
Round-trip, 1 traveler

Sure! I can book that for you.
Could you confirm:
• Exact dates: Apr 04 to Apr 11?
• Traveler: 1 adult?
• Cabin class: Economy?
• Any time preferences?

2 Search flights



Search flights
PIT → SFO (Apr 04)
SFO → PIT (Apr 11)

Here are the best options I found:
1) 6:00 AM PIT → 8:35 AM SFO
7:30 AM SFO → 3:55 PM PIT
\$358

Would you like me to book this one?

3 Hold & book



Hold seats → Book with
payment (card ending XXX)

Booking confirmed! 🎉

Round-trip: PIT → SFO (Apr 04)
SFO → PIT (Apr 11)
Confirmation: ABC123
Total charged: \$358
A receipt has been sent to your email.

4 Confirmation



Send confirmation email
& receipt

You're welcome! Safe travels! ✈️
Let me know if you need anything else.

Multi-Turn Agent-User Interaction

Personalized Task: Book Travel Plan

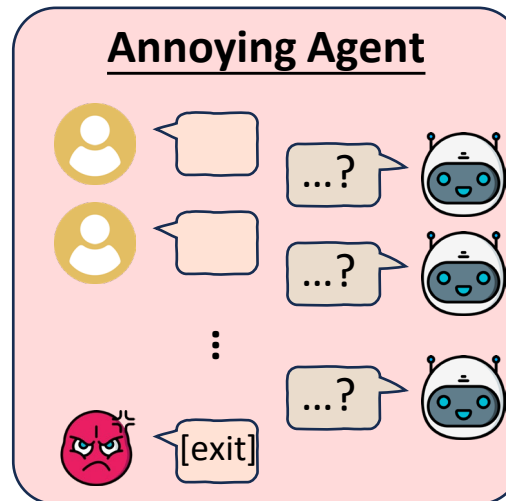
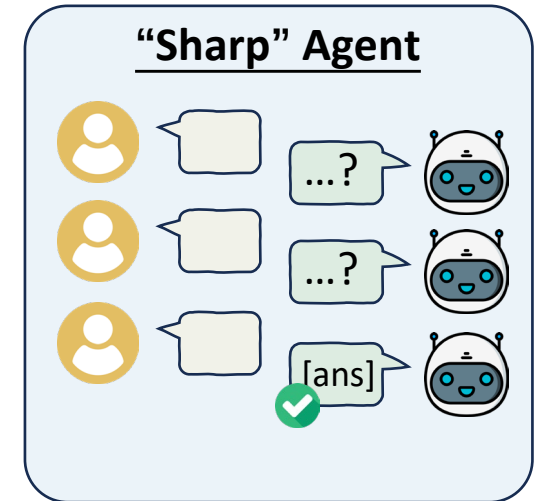
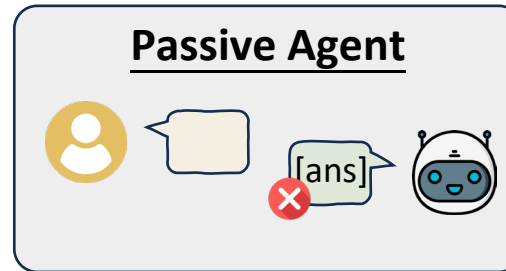
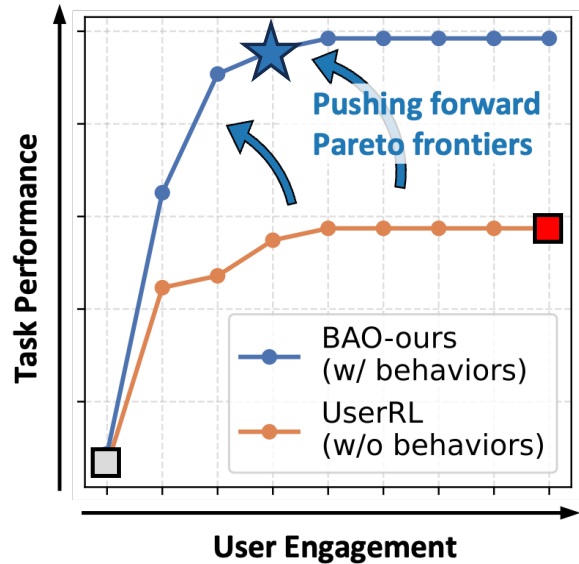
Need User-Preference-Tailored Decision Making

Proactive Agent vs. Passive Agent



Smart agent should proactively gather personal information from the users.

Challenge: Balancing Performance and User Engagement



Proactive agent has the risk of reducing user satisfaction

Problem Formulation: Multi-Objective Contextual MDP

Contextual MDP: $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, \boxed{c}, \mu_0\}$

Context c : User profile/persona, underlying intension

Transition dynamics: $s_{t+1} \sim P(\cdot \mid s_t, a_t, \boxed{c})$

Reward function: $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \boxed{c} \rightarrow \mathbb{R}$

Action Space: $\mathcal{A} := \mathcal{A}_u \cup \mathcal{A}_e$
User-involved actions + Environment-involved actions

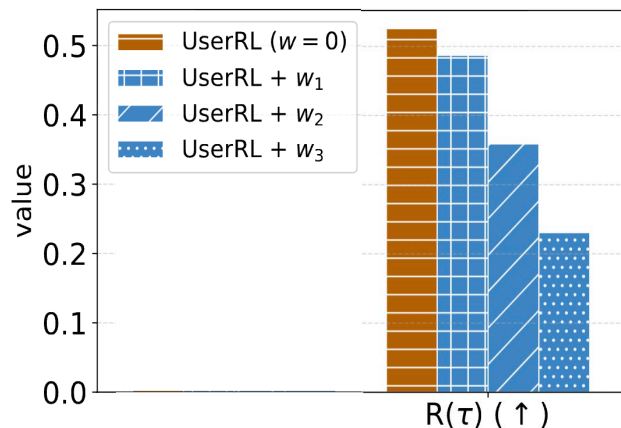
Goal: Pareto optimal of $R(\tau)$ and $U(\tau)$

Task Performance: $R(\tau) := \sum r_t$

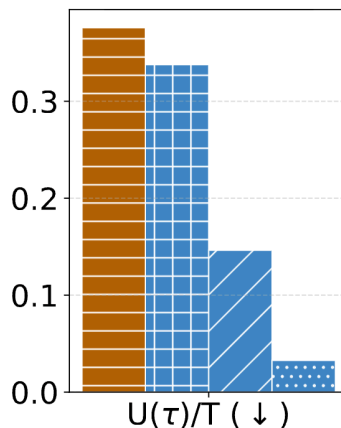
User Engagement: $U(\tau) = |\{t : a_t \in \mathcal{A}_u\}|$

Natural Multi-Objective Optimization Does Not Work

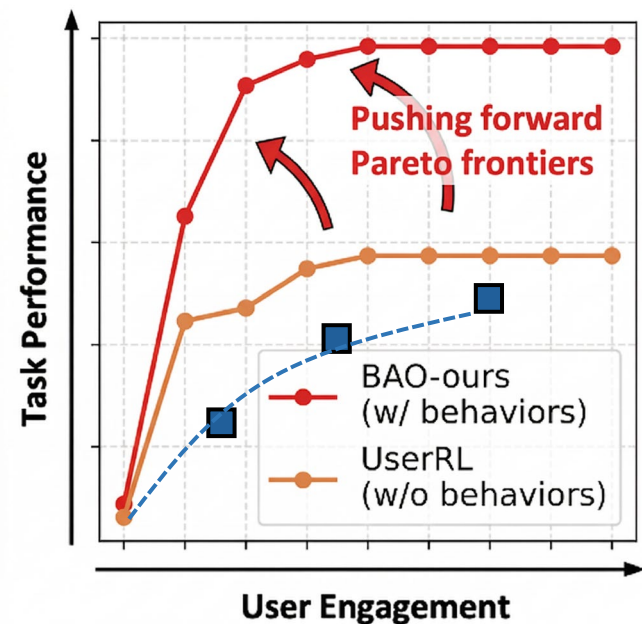
$$\max_{\pi_{\theta}} \mathbb{E}[R(\tau) - \boxed{w}U(\tau)] \quad \text{s.t. } |\tau| \leq T$$



Task performance



User engagement



Directly weighting two objective functions will not work

This Work: Behavior Regularization

Useful Inter-Turn Behaviors

- Retrospective Reasoning: memorizing past

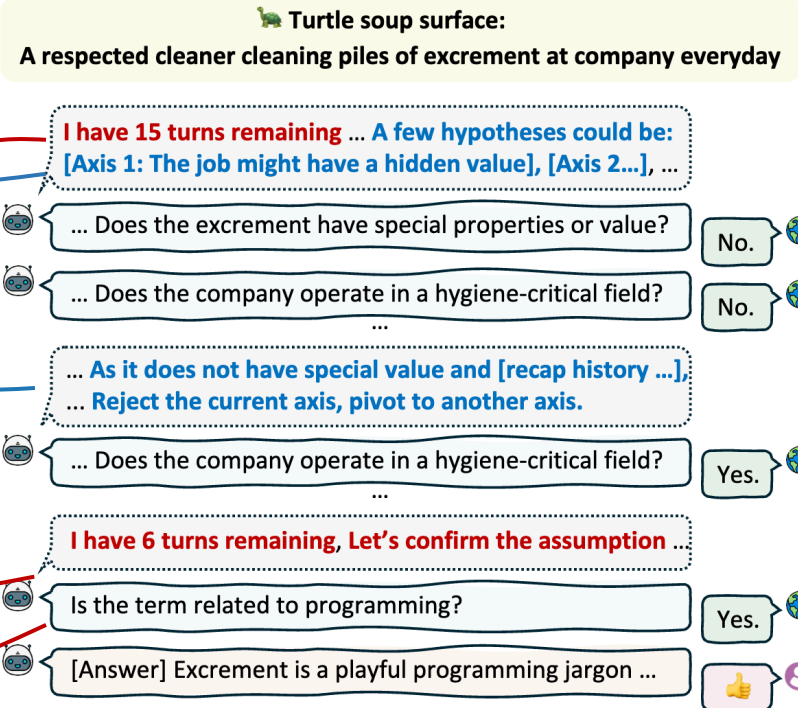
(1) Hypothesis Refinement

(2) Memory management

- Prospective Planning: envisioning future

(1) Dynamic Scheduling

(2) Strategical Querying



This Work: Behavior Regularization

Behavior regularization through reward shaping

- Hypothesis Refinement: avoid redundant reasoning without sufficient user engagements
 - : the budget of interaction turns
 - : the actual length of the failure trajectory

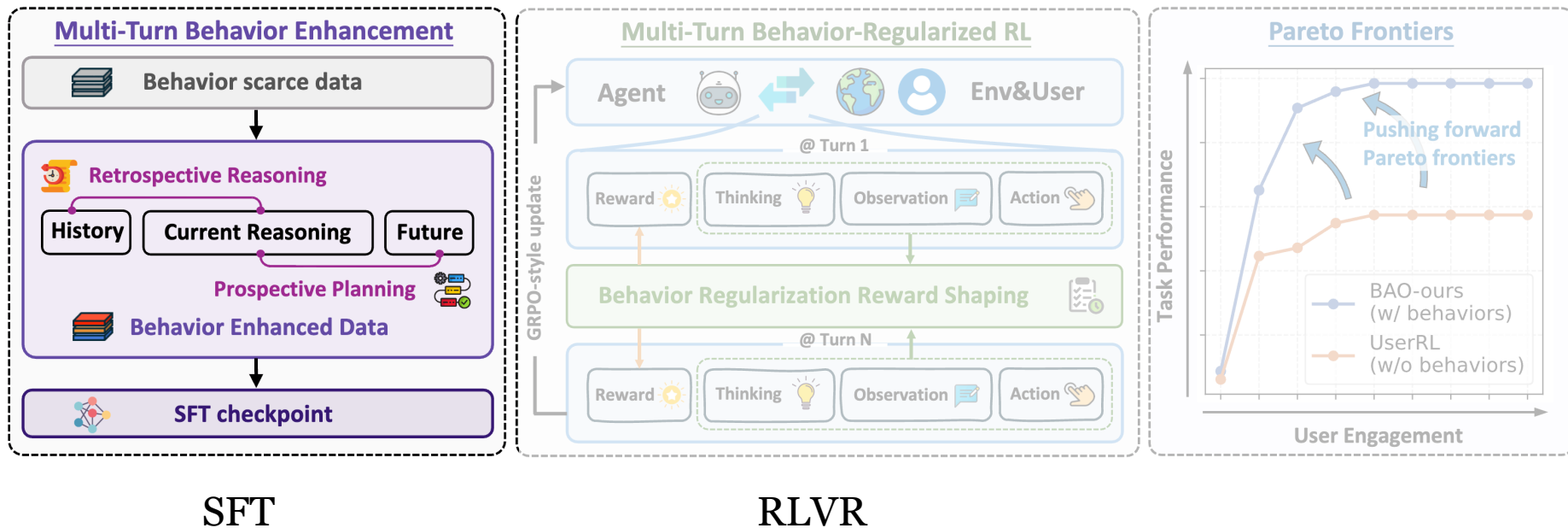
This Work: Behavior Regularization

Behavior regularization through reward shaping

- Hypothesis Refinement: avoid redundant reasoning without sufficient user engagements
 - : the budget of interaction turns
 - : the actual length of the failure trajectory
- Strategical querying: stop consecutive request of user engagement

$$r_t \leftarrow r_t - \lambda_{\text{ans}}, \text{ if } a_t, a_{t-1} \in \mathcal{A}_u$$

Training Pipeline: Behavior Regularization

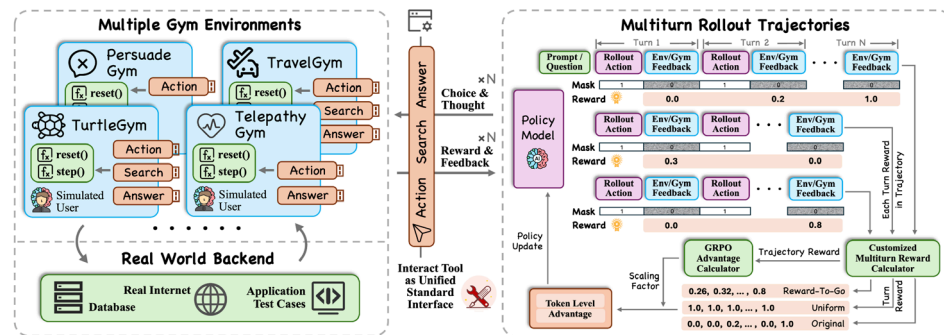


- SFT: use behavior enhanced data with teacher model's help
- RLVR: use behavior regularized reward with GRPO

Experimental Settings

■ Evaluation benchmark: user-centric tasks in UserRL

- Multi-turn interaction (~15)
- Tool-use integrated
- Simulated users
- Rule-based RM and LLM-as-judges



■ Baselines

- Commercial and Open-source models,
- RL models (UserRL): same SFT seed data, teacher model, and RL data

■ Metrics

1. **Task performance:** Pass@U-k: Pass rate with k user engagement actions per trajectory
2. **User engagement Ratio (UR):** ratio of user-involved actions

Cheng Qian, et al. "Userrl: Training interactive user-centric agent via reinforcement learning." arXiv preprint arXiv:2509.19736 (2025).

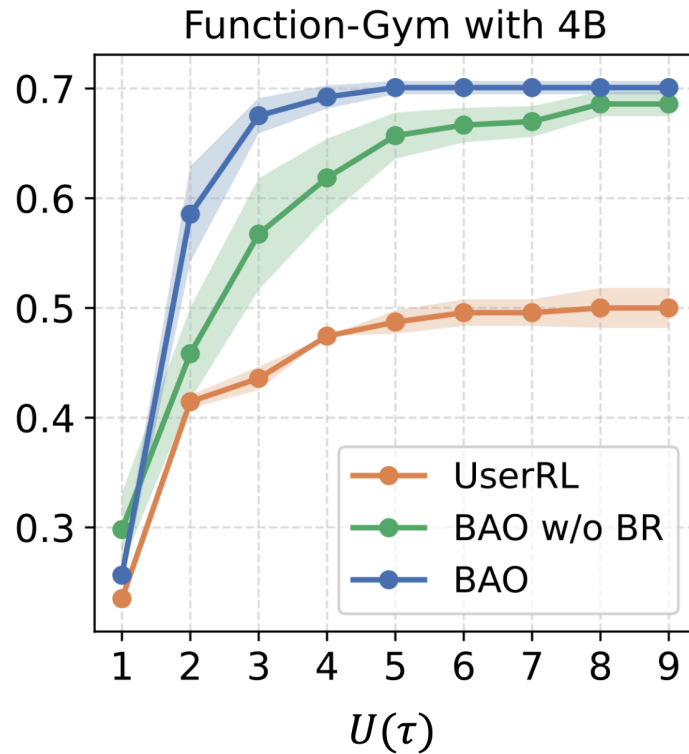
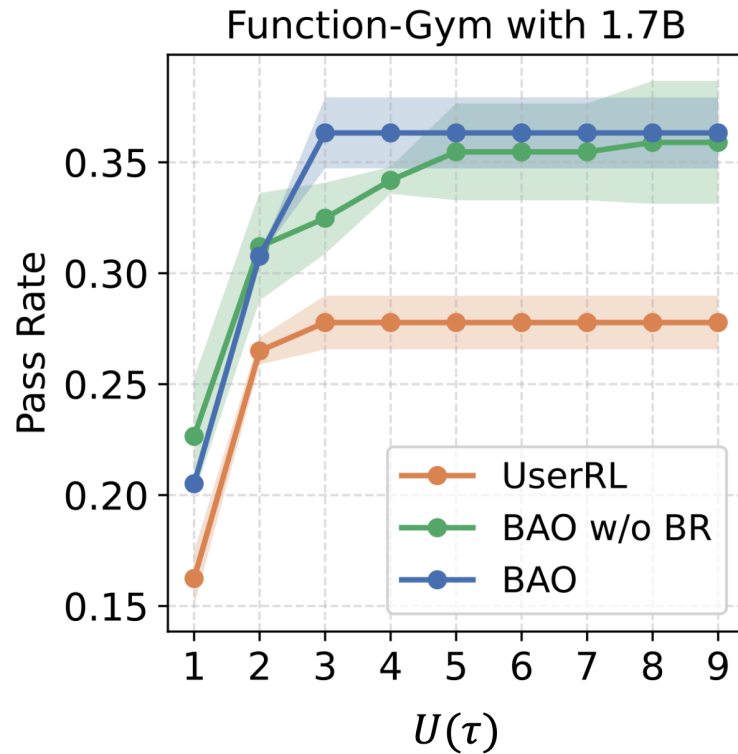
Main Results

	Function-Gym				Telepathy-Gym				Turtle-Gym			
	Pass@U-1(↑)	Pass@U-2(↑)	Score(↑)	UR(↓)	Pass@U-1(↑)	Pass@U-2(↑)	Score(↑)	UR(↓)	Pass@U-1(↑)	Pass@U-2(↑)	Score(↑)	UR(↓)
<i>Commercial Models</i>												
Gemini-3-flash	0.4103	0.4231	0.4231	0.0440	0.8049	0.8293	0.8537	0.1414	0.1823	0.2073	0.2073	0.0613
Gemini-2.5-Pro	0.3205	0.4359	0.4359	0.1003	0.5854	0.6585	0.7317	0.2330	0.3156	0.3469	0.3844	0.2250
Gemini-2.5-Flash	0.2308	0.3333	0.3718	0.1290	0.5853	0.6585	0.7073	0.1625	0.1604	0.1844	0.2198	0.2155
GPT-5-0807	0.5000	0.6923	0.7436	0.1365	0.7317	0.8293	0.8293	0.1270	0.3500	0.3521	0.3521	0.0897
GPT-4o	0.1282	0.1282	0.1410	0.2248	0.4634	0.5854	0.6342	0.1747	0.1719	0.2333	0.2750	0.2553
GPT-4o-mini	0.0256	0.0256	0.0385	0.1767	0.4634	0.5610	0.6216	0.1183	0.0598	0.0732	0.0793	0.3323
<i>Open-Source (Raw Models)</i>												
Qwen3-32B (Raw)	0.1667	0.1795	0.1795	0.0379	0.5366	0.5610	0.5854	0.1913	0.1000	0.1500	0.1688	0.3082
Qwen3-14B (Raw)	0.1154	0.1154	0.1282	0.0389	0.3902	0.4634	0.4634	0.2532	0.1281	0.1406	0.1500	0.5264
Qwen3-8B (Raw)	0.0513	0.0513	0.0513	0.0559	0.3902	0.4390	0.4878	0.1699	0.0792	0.0938	0.1135	0.4778
<i>Fine-tuned Models</i>												
UserRL-4B	0.2436	0.4103	0.5256	0.3758	0.4878	0.5122	0.6585	0.4251	0.0417	0.0563	0.0594	0.7617
BAO (ours)-4B	0.2692	0.5354	0.6923	0.2148	0.5123	0.6585	0.6585	0.1870	0.1063	0.1125	0.1125	0.2917
BAO (ours) w/o BE-4B	0.2436	0.3718	0.4744	0.2483	0.4878	0.5366	0.5610	0.1452	0.0802	0.0979	0.1146	0.3359
BAO (ours) w/o BR-4B	0.3333	0.4615	0.7179	0.3755	0.3902	0.4390	0.5610	0.5290	0.0729	0.0750	0.0813	0.8698
UserRL-1.7B	0.1538	0.2692	0.2949	0.3193	0.2683	0.2683	0.4390	0.6543	0.0479	0.0510	0.0583	0.7721
BAO (ours)-1.7B	0.1923	0.3205	0.3590	0.2133	0.5610	0.5854	0.6097	0.1601	0.0563	0.0667	0.0667	0.2976
BAO (ours) w/o BE-1.7B	0.1410	0.2179	0.2436	0.2404	0.4390	0.4878	0.4878	0.1559	0.0615	0.0660	0.0750	0.3776
BAO (ours) w/o BR-1.7B	0.2308	0.2949	0.3974	0.4284	0.1707	0.1951	0.4390	0.6911	0.0594	0.0625	0.0771	0.9064

Prior RL baseline struggle with proactive agent tasks:
low performance and high user engagement

The proposed BAO is efficient in proactive agent tasks

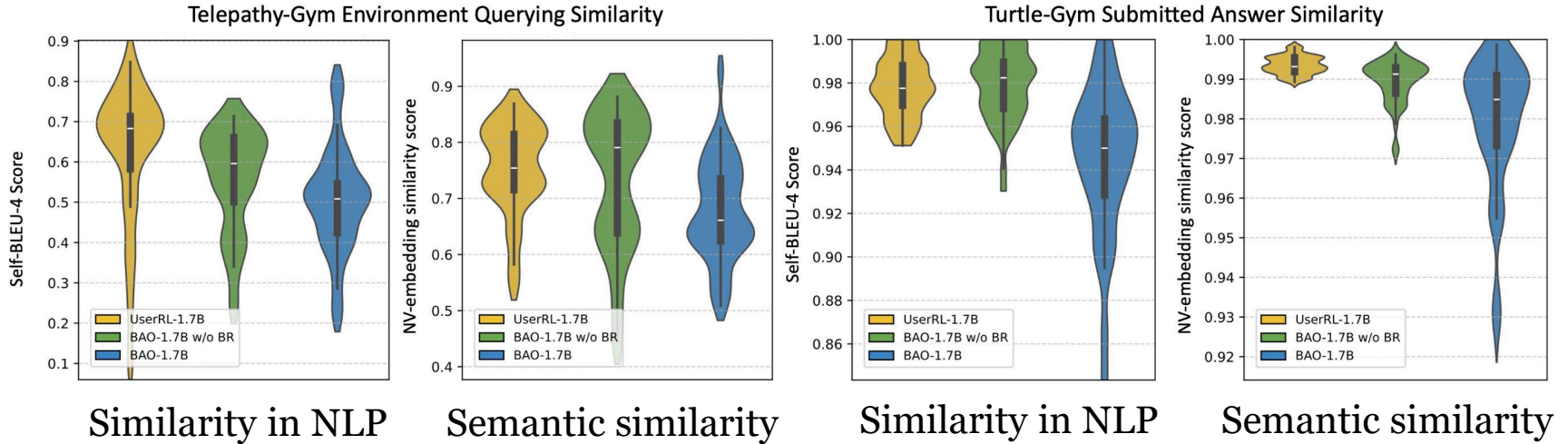
Main Results



Behavior Enhancement via SFT and Behavior Regularization via RLVR
Jointly Push Forward the Pareto Frontier

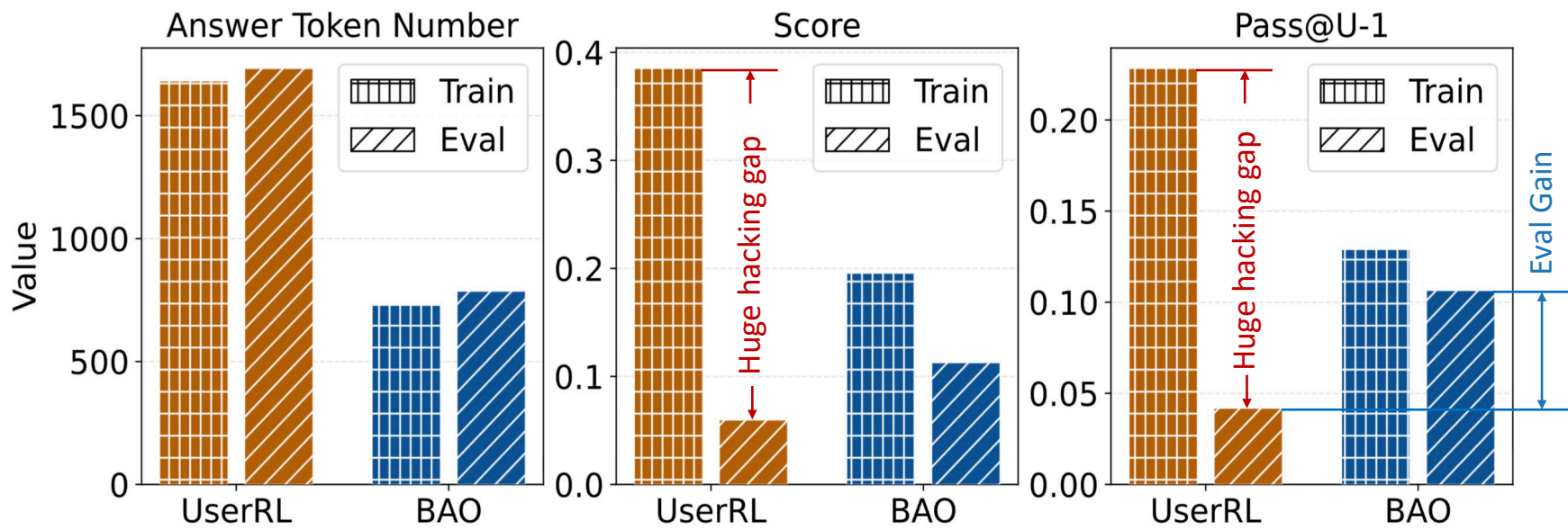
Optimizing Two Objectives without An Explicit Weighted Sum:
Implicit Behavior Regularization

Behavior Injection Helps Exploration



Behavior Injection Promotes Answer Diversity

Behavior Injection Reduce Reward Hacking

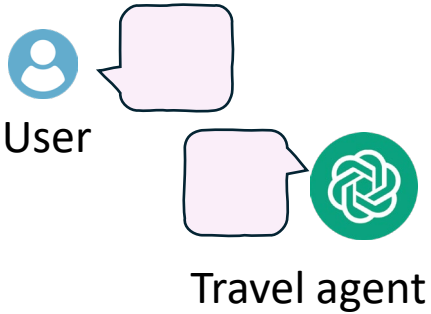


Train: Qwen3-8B as judges

Eval: GPT-4o as judges

Failure cases: repeatedly generating long answers to confuse LLM as judges

Agentic RL: Optimizing for More Than Just Performance



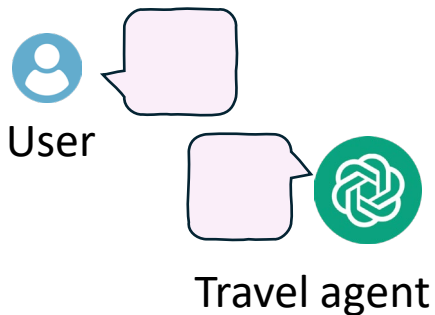
Task performance



User satisfaction

I: Agentic RL for multi-turn agent-user interaction:
behavior regularization → Pareto optimal

Agentic RL: Optimizing for More Than Just Performance

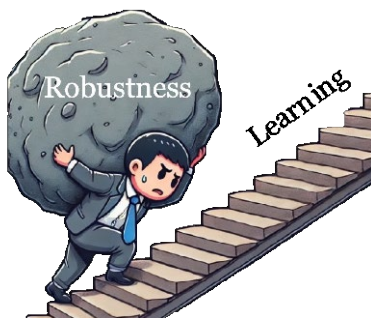


Task performance



User satisfaction

I: Agentic RL for multi-turn agent-user interaction:
behavior regularization → Pareto optimal



Robust performance



Data efficiency

II: Robust RL
without sacrificing data efficiency, even enhancing it

Pipeline of Decision-Making AI

Training

Deployment

Pipeline of Decision-Making AI

Training

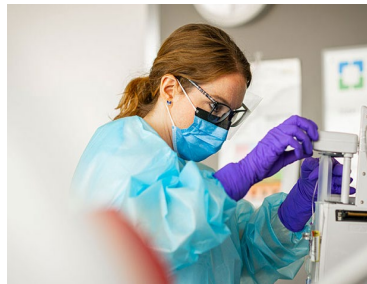
Accessible
Environments

Deployment



Pick up boxes

Robotics¹



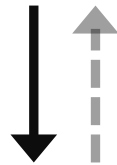
Clinical treatments²

Pipeline of Decision-Making AI

Training

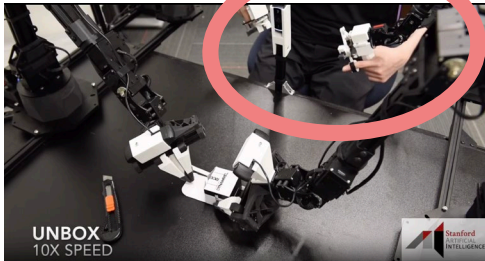
Accessible
Environments

Data &
Information

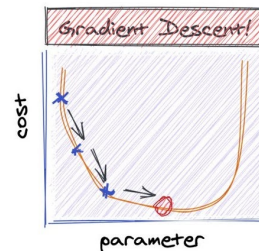


Learning
Algorithm

Deployment



Human Demonstrations¹



Optimization

[1] Zhao, Tony Z., Vikash Kumar, Sergey Levine, and Chelsea Finn. "Learning fine-grained bimanual manipulation with low-cost hardware." *arXiv preprint arXiv:2304.13705* (2023).

Pipeline of Decision-Making AI

Training

Accessible
Environments

Data &
Information



Learning
Algorithm

Deployment

Real-world
Environments

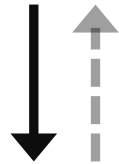


Pipeline of Decision-Making AI

Training

Accessible
Environments

Data &
Information



Learning
Algorithm

Deployment

Real-world
Environments



Is AI Ready for Real-World Applications?

✗ I. Data shortage

✗ III. Not robust to uncertainty

Is AI Ready for Real-World Applications?

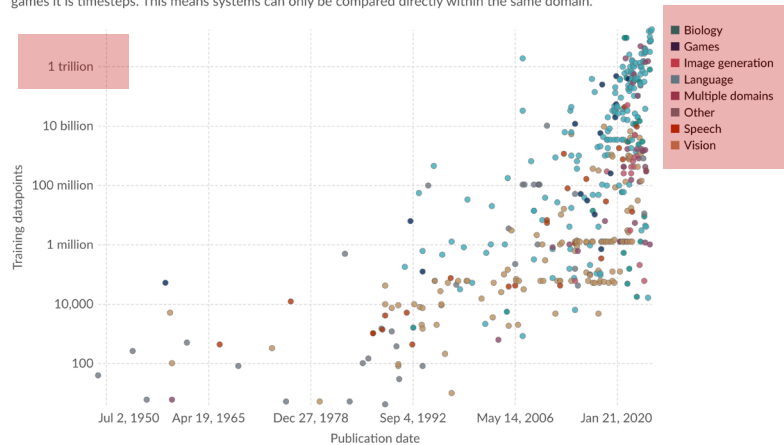


I. Data shortage

Datapoints used to train notable artificial intelligence systems

Each domain has a specific data point unit; for example, for vision it is images, for language it is words, and for games it is timesteps. This means systems can only be compared directly within the same domain.

Our World
in Data



ChatGPT training dataset size

ChatGPT-1

117 million
parameters

ChatGPT-2

1.5 billion
parameters

ChatGPT-3

175 billion
parameters

ChatGPT-4

1 trillion
parameters*

*Estimated

Dataset sizes for the first four versions of ChatGPT (Source: [datacamp](#) and [Wired](#).)

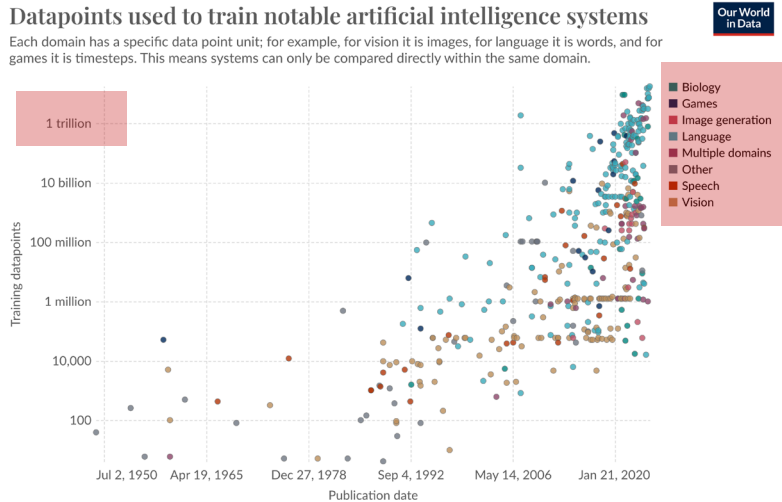
Is AI Ready for Real-World Applications?



I. Data shortage

Datapoints used to train notable artificial intelligence systems

Each domain has a specific data point unit; for example, for vision it is images, for language it is words, and for games it is timesteps. This means systems can only be compared directly within the same domain.



ChatGPT training dataset size

ChatGPT-1

117 million
parameters

ChatGPT-2

1.5 billion
parameters

ChatGPT-3

175 billion
parameters

ChatGPT-4

1 trillion
parameters*

*Estimated

A.I. Companies Are Running Out of Training Data: Study

By [Alexandra Tremayne-Pengelly](#) • 07/19/24 1:06pm

4 Data Hurdles for AI in Healthcare



CloudFactory

Feb 16, 2023

- ✓ Privacy and Regulatory Compliance
- ✓ Data Availability and Collection
- ✓ AI Bias
- ✓ Industry Expertise and Quality Assurance

Dataset sizes for the first four versions of ChatGPT (Source: [datacamp](#) and [Wired](#).)

Is AI Ready for Real-World Applications?

✗ I. Data shortage

10 confused Cruise robotaxis
create an autonomous traffic
jam in San Francisco

BEN HSU • AUG 15, 2023

f X reddit Start Conversation



✗ III. Not robust to uncertainty

AI for medicine is overhyped



DAN ELTON

MAR 29, 2022

Robustness woes

Most failings of AI in radiology can be traced to the fact that deep learning models are not robust to distributional shift.

 I. Data shortage

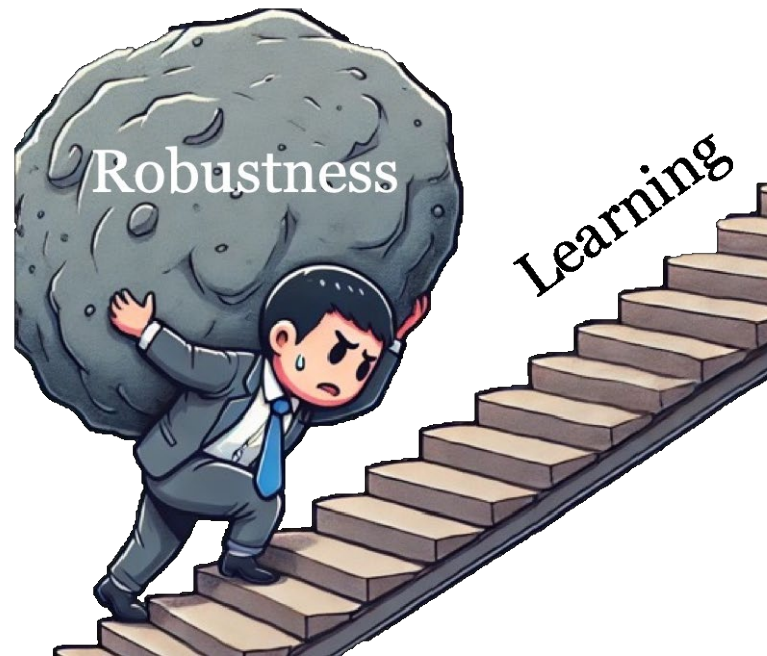
Calls for Data Efficiency

 III. Not robust to uncertainty

Calls for Robust Performance
against Risk & Uncertainty

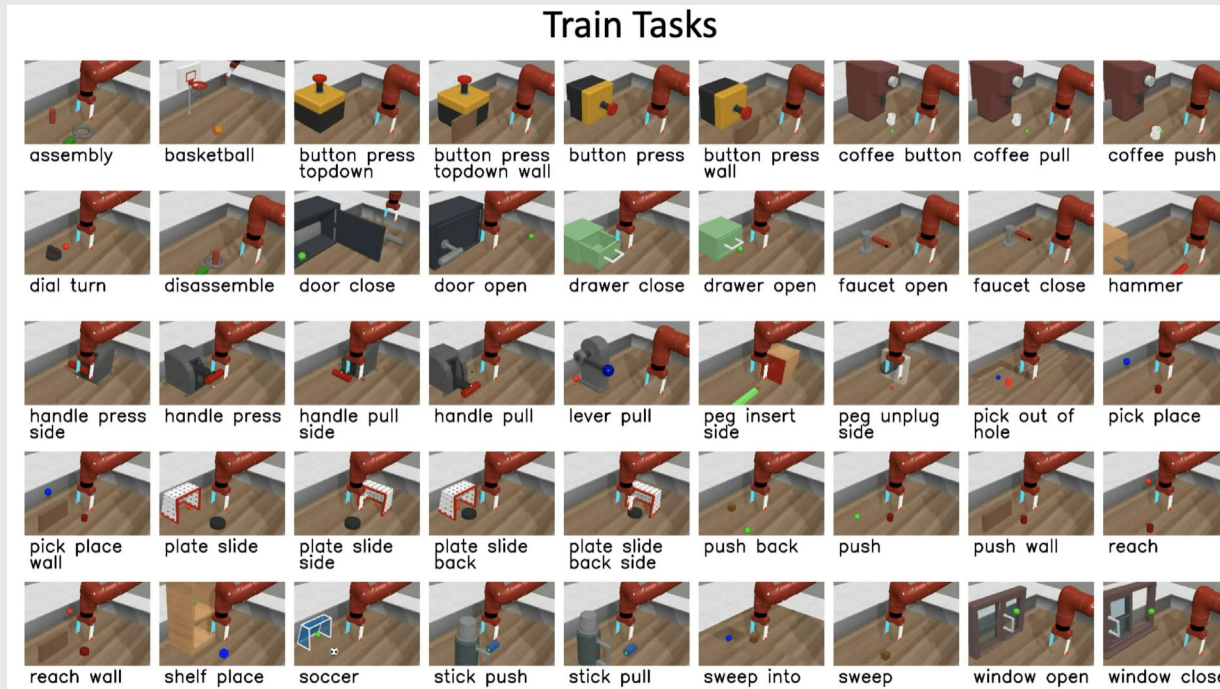
Can we design data-efficient & robust AI?

Current Methods Gain Robustness by **Sacrificing** Efficiency



Current Methods Gain Robustness by **Sacrificing** Efficiency

Enhance Dataset: larger and more diverse



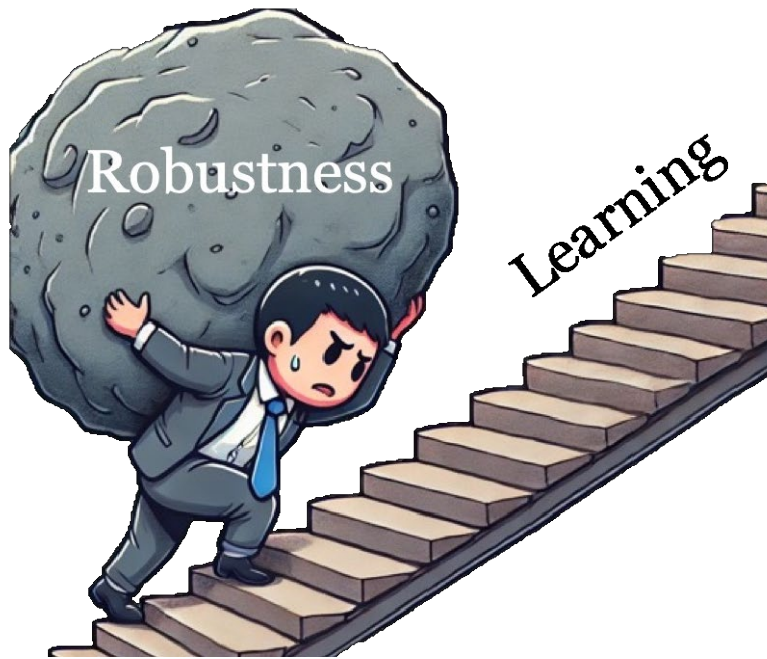
[Yu, et al., '21]

Sacrifices Data Efficiency

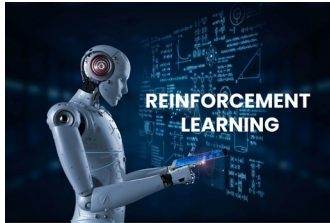
Is Sacrificing Data Efficiency Necessary for Robustness?



Is Sacrificing Data Efficiency Necessary for Robustness?



Robustness **Without Sacrificing** Data Efficiency

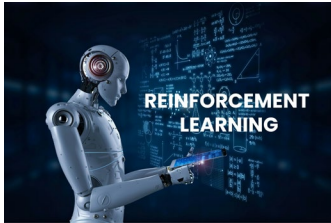


I. Data shortage



III. Not robust to uncertainty

Robustness **Without Sacrificing Data** Efficiency



Data efficiency

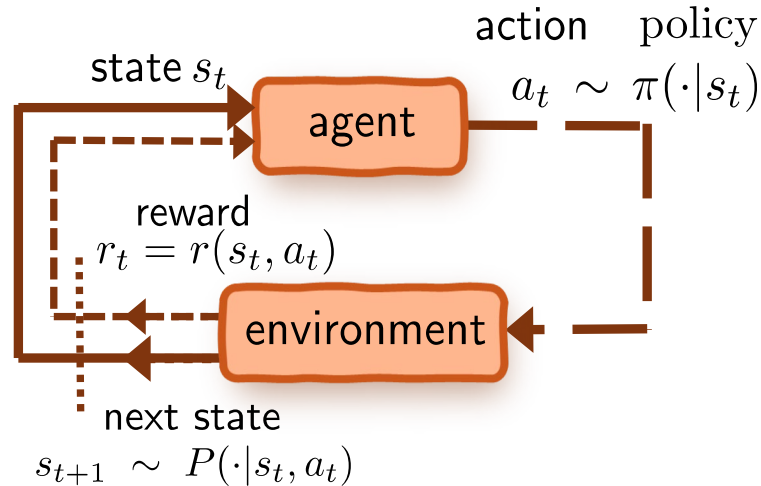


Robustness

I: Robust decision-making
without sacrificing **data efficiency**

RL: Markov Decision Process (MDP)

An agent learns by interacting sequentially with an environment

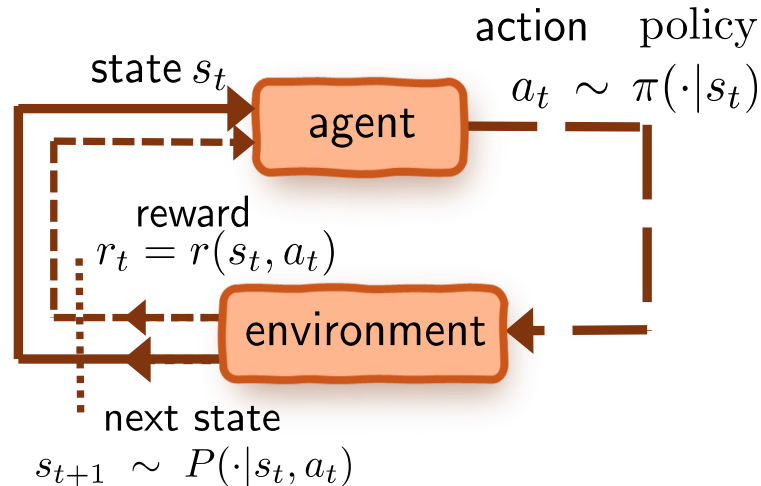


- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: **unknown** immediate reward
- $P(\cdot|s, a)$: **unknown** transition probabilities

} environment

RL: Markov Decision Process (MDP)

An agent learns by interacting sequentially with an environment



- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $r(s, a) \in [0, 1]$: **unknown** immediate reward
- $P(\cdot|s, a)$: **unknown** transition probabilities

} environment

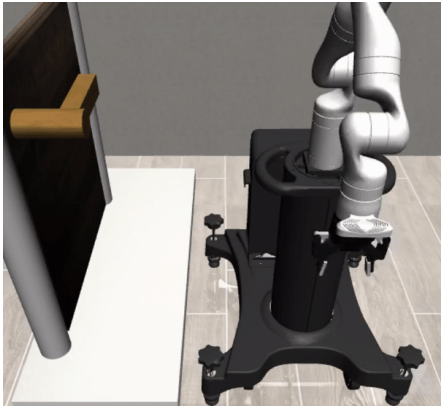
Goal: find a policy $\pi(\cdot|s)$ that $\max_{\pi} V^{\pi, P} := \mathbb{E}_{\pi, P} \left[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) \right]$

Value function: long-term return

RL is Not Robust to Environmental Shift

Control arm to open the door

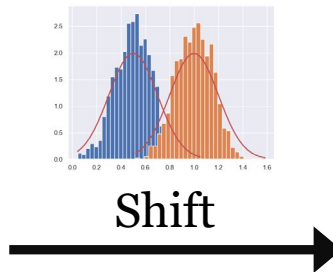
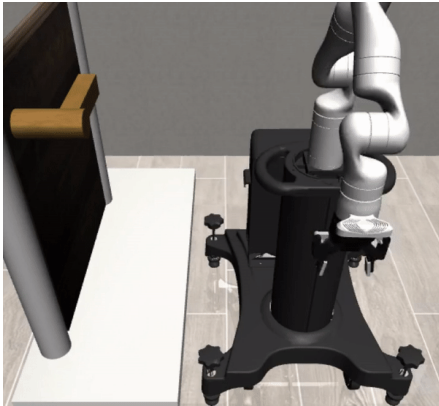
Training



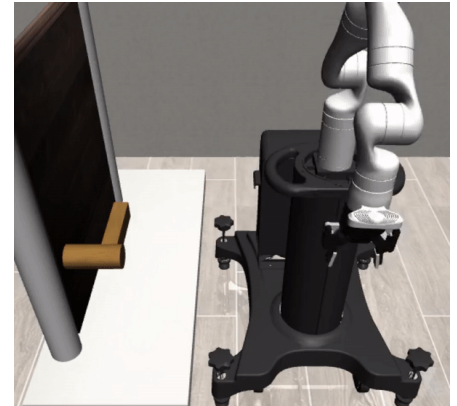
RL is Not Robust to Environmental Shift

Control arm to open the door

Training



Deployment

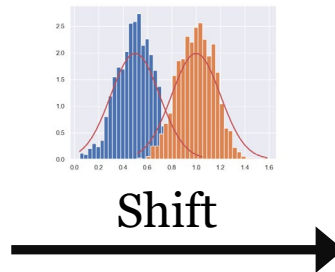
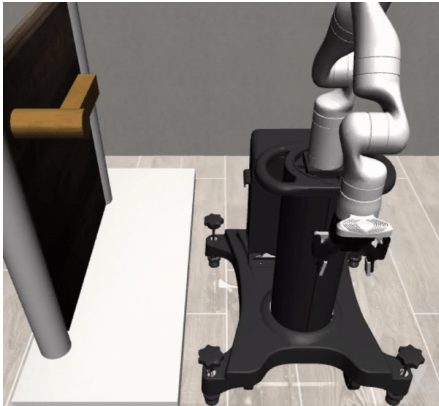


Standard RL Fails

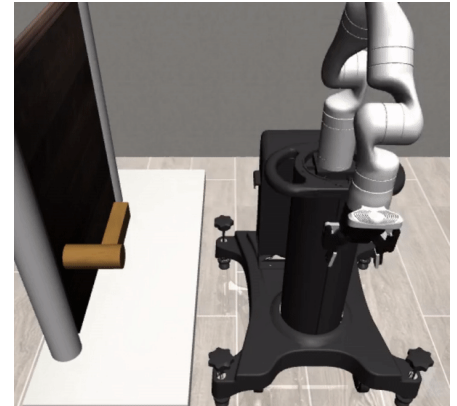
RL is Not Robust to Environmental Shift ☹️

Control arm to open the door

Training



Deployment



Standard RL Fails

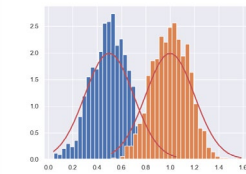
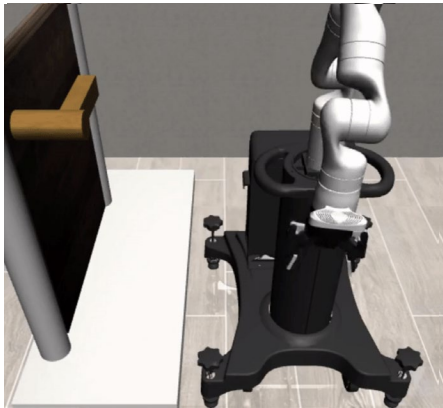
Environmental shift occurs frequently

- Sim-to-real gaps: robotics, self-driving cars...
- Disparities: supply chain, healthcare systems, computer vision...
- System noise: hardware, technician...
- Sudden events: wildfire, traffic accidents...

A Pressing Need:
Robustness against Environmental Shift

This Talk: Distributionally Robust RL

Training



Shift

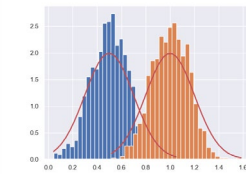
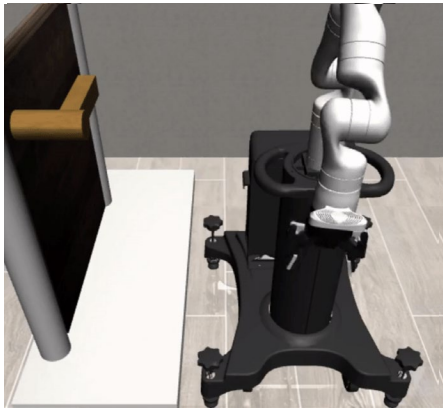


Deployment

Robust RL in training and shifted envs

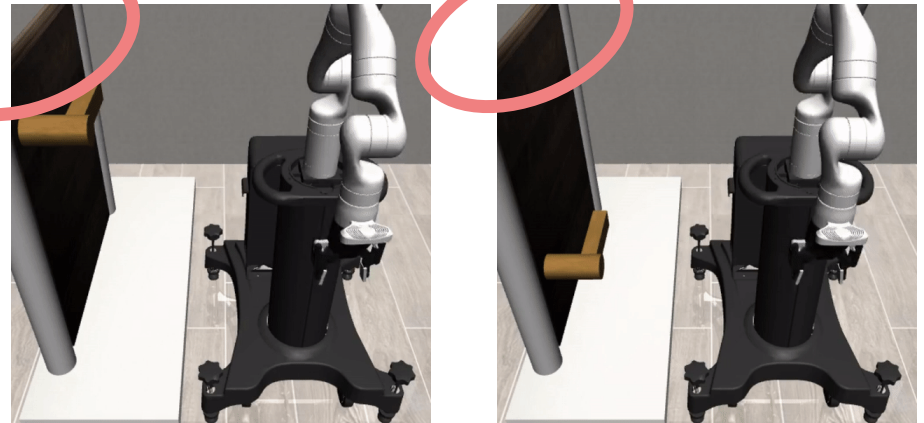
This Talk: Distributionally Robust RL

Training



Shift

Deployment



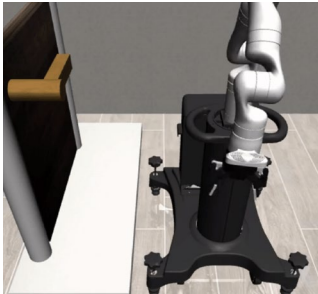
Robust RL in training and shifted envs

Robust RL fosters the emergence of smarter solutions

Modeling Environment Uncertainty

Training (nominal)

P^o



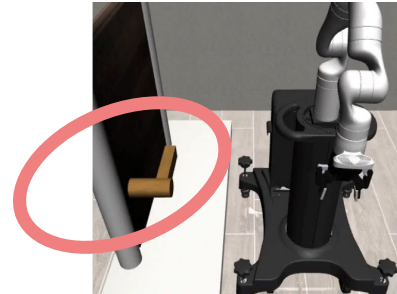
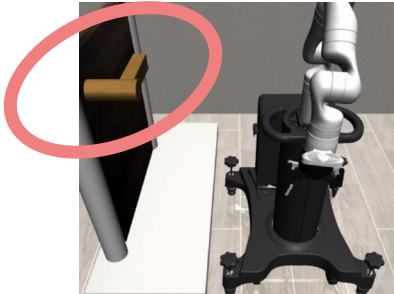
Modeling Environment Uncertainty

Training (nominal)

P^o

Deployment

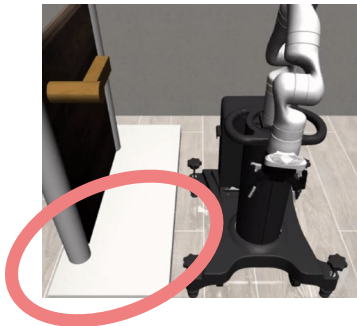
P^o
 P



Modeling Environment Uncertainty

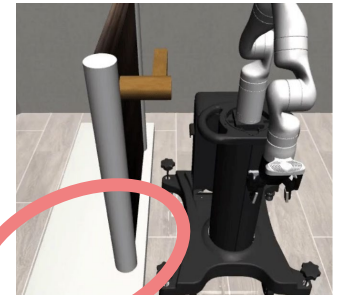
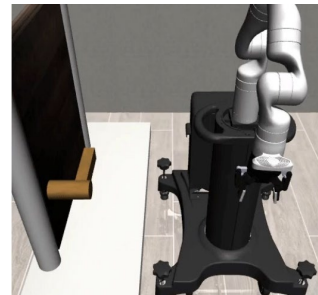
Training (nominal)

P^o



Deployment

P P^o



...

Modeling Environment Uncertainty

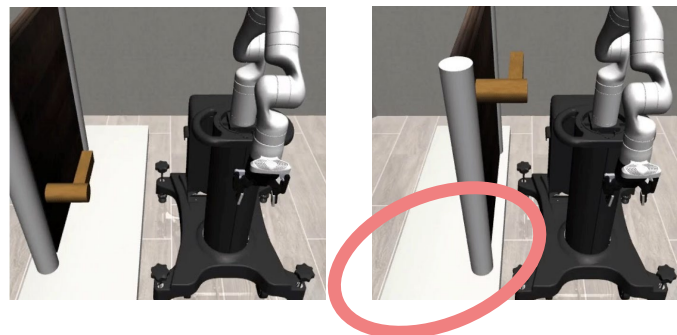
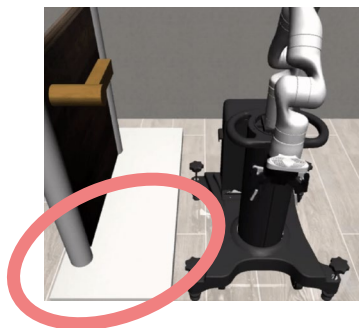
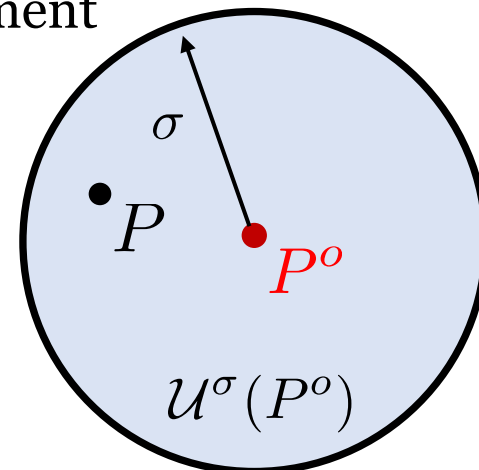
Uncertainty set of the deploying environment around P^o :

$$\mathcal{U}^\sigma(P^o) = \{P : \rho(P, P^o) \leq \sigma\}$$

Training (nominal)

P^o

Deployment



Modeling Environment Uncertainty

Uncertainty set of the deploying environment around P^o :

$$\mathcal{U}^\sigma(P^o) = \{P : \rho(P, P^o) \leq \sigma\}$$

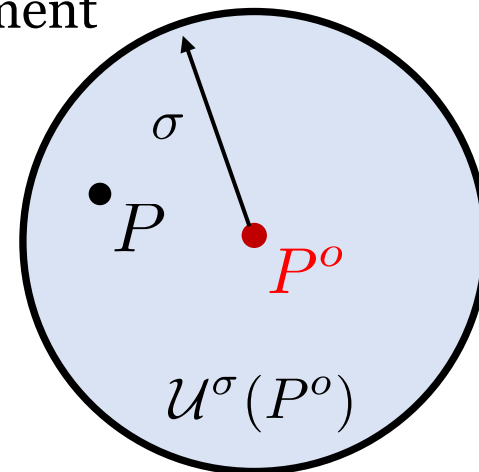
Measures the divergence

Training (nominal)



P^o

Deployment



- Examples of ρ : f-divergence (TV, χ^2 , KL...), Wasserstein distance, learning-based

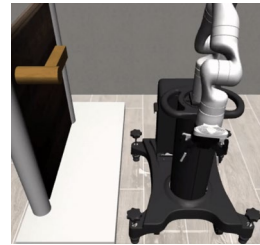
Data Collection from a Generative Model

— [Kearns and Singh, '1999]

Data from
training
(**unknown**
nominal P^o)

=

Generative model



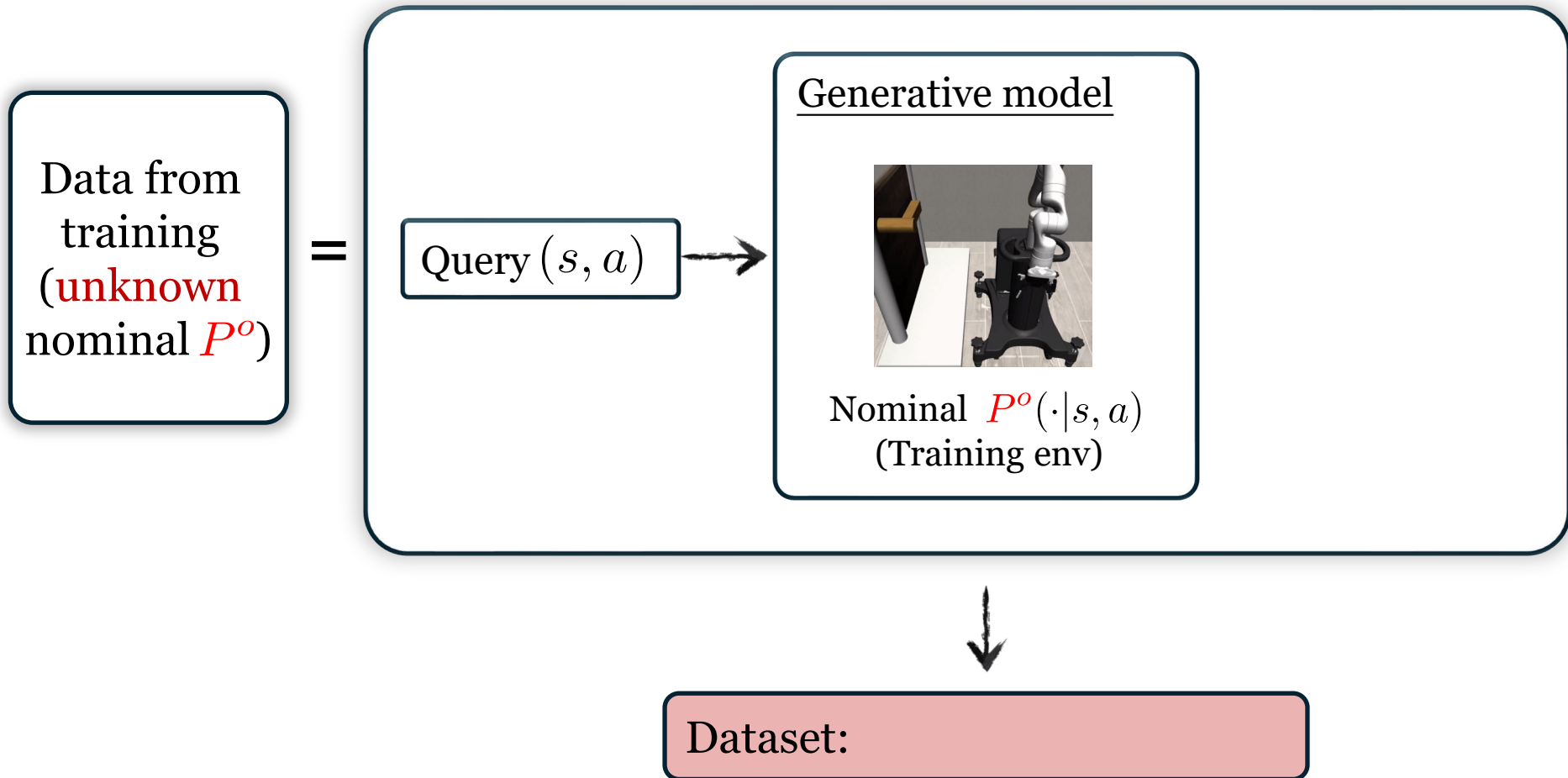
Nominal $P^o(\cdot|s, a)$
(Training env)



Dataset:

Data Collection from a Generative Model

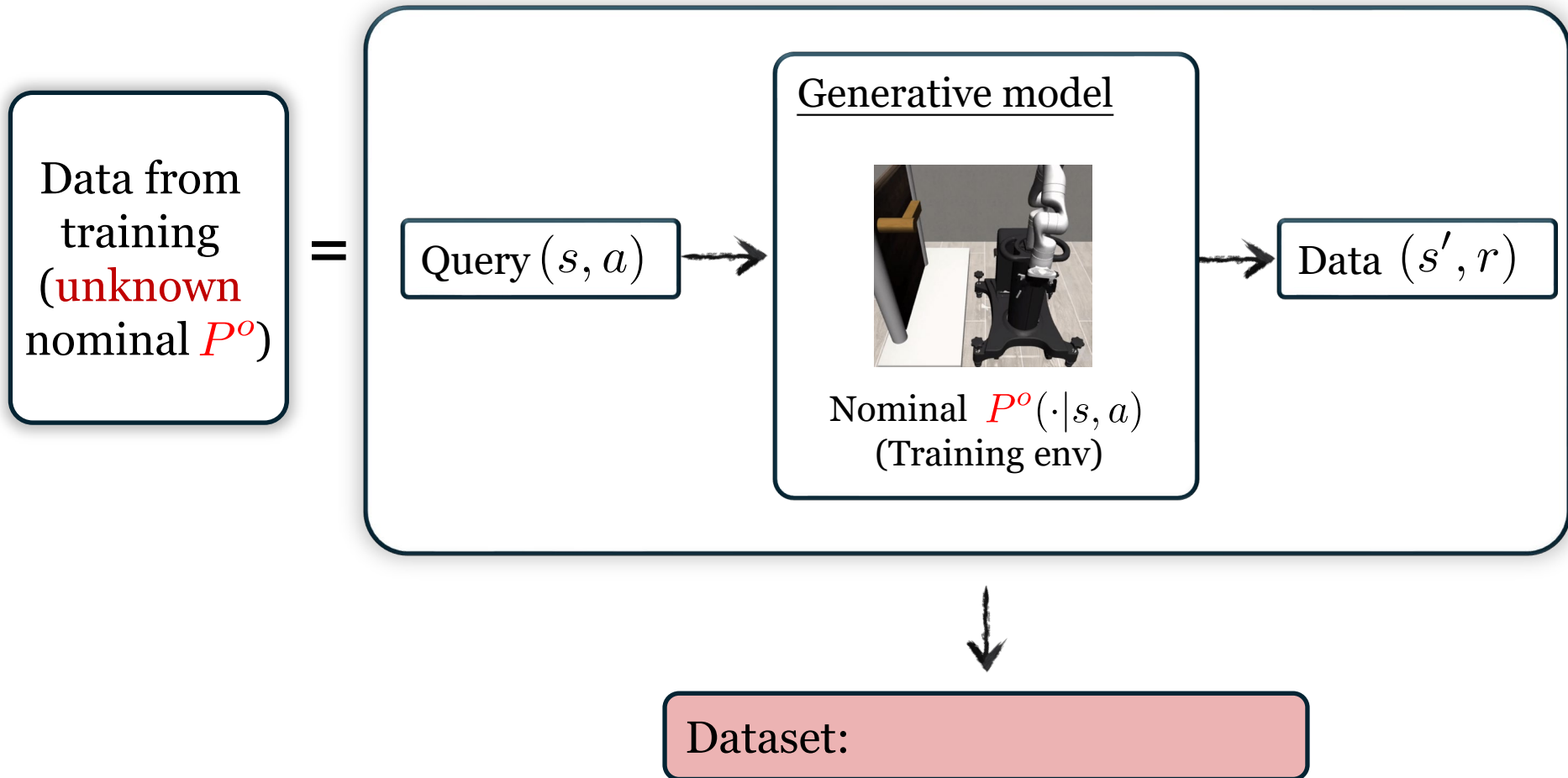
— [Kearns and Singh, '1999]



- $\mathcal{S}$: state space, $s \in \mathcal{S}$
- $\mathcal{A}$: action space, $a \in \mathcal{A}$

Data Collection from a Generative Model

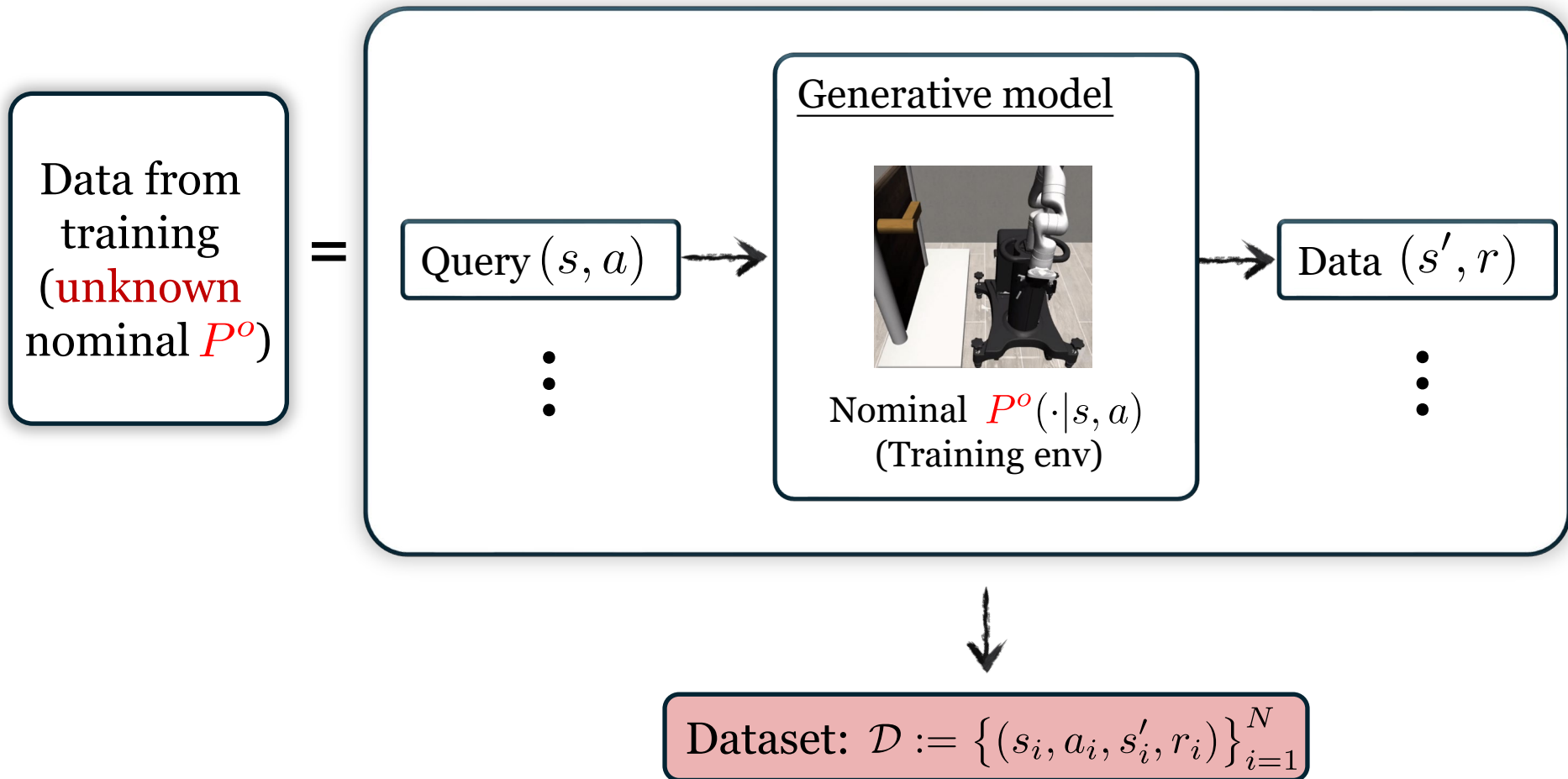
— [Kearns and Singh, '1999]



- $\mathcal{S}$: state space, $s \in \mathcal{S}$
- $\mathcal{A}$: action space, $a \in \mathcal{A}$

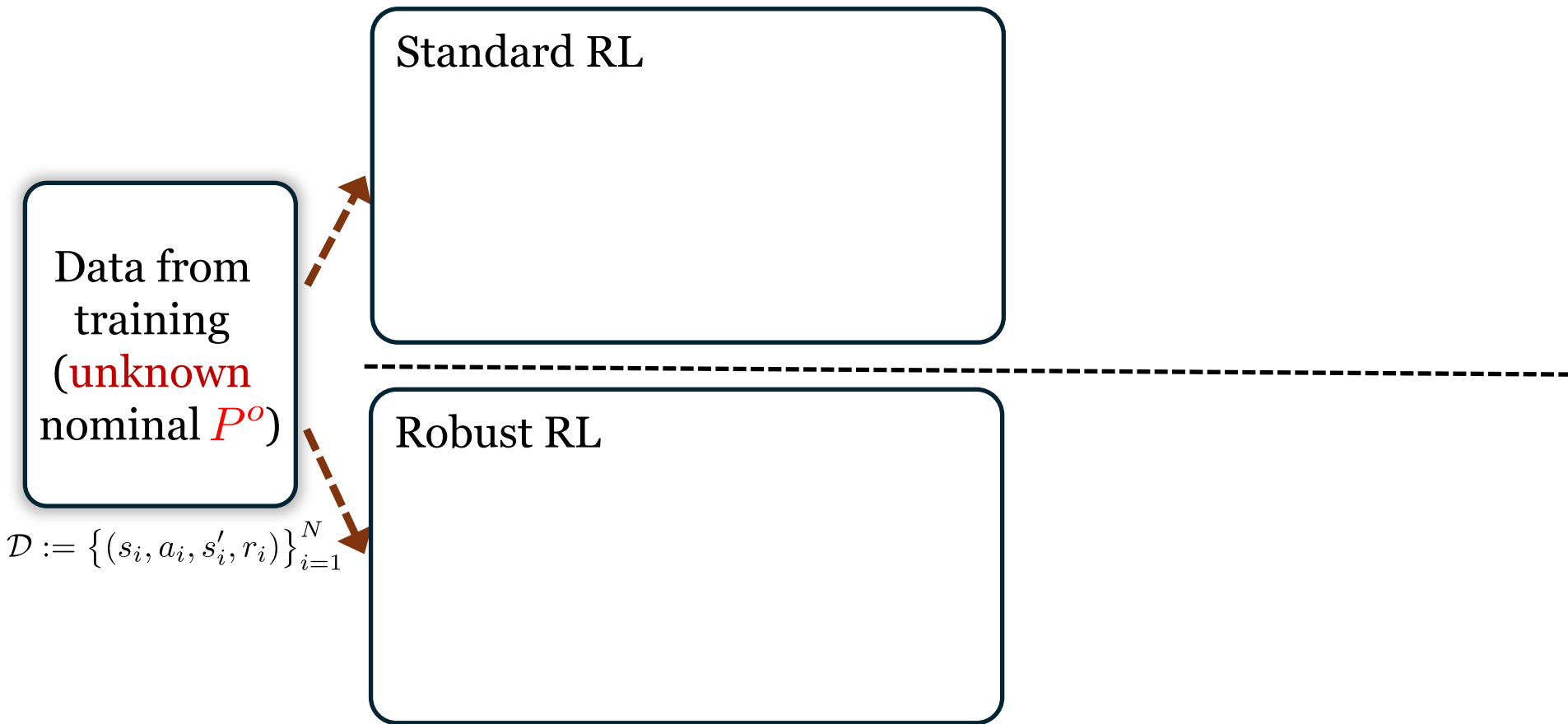
Data Collection from a Generative Model

— [Kearns and Singh, '1999]

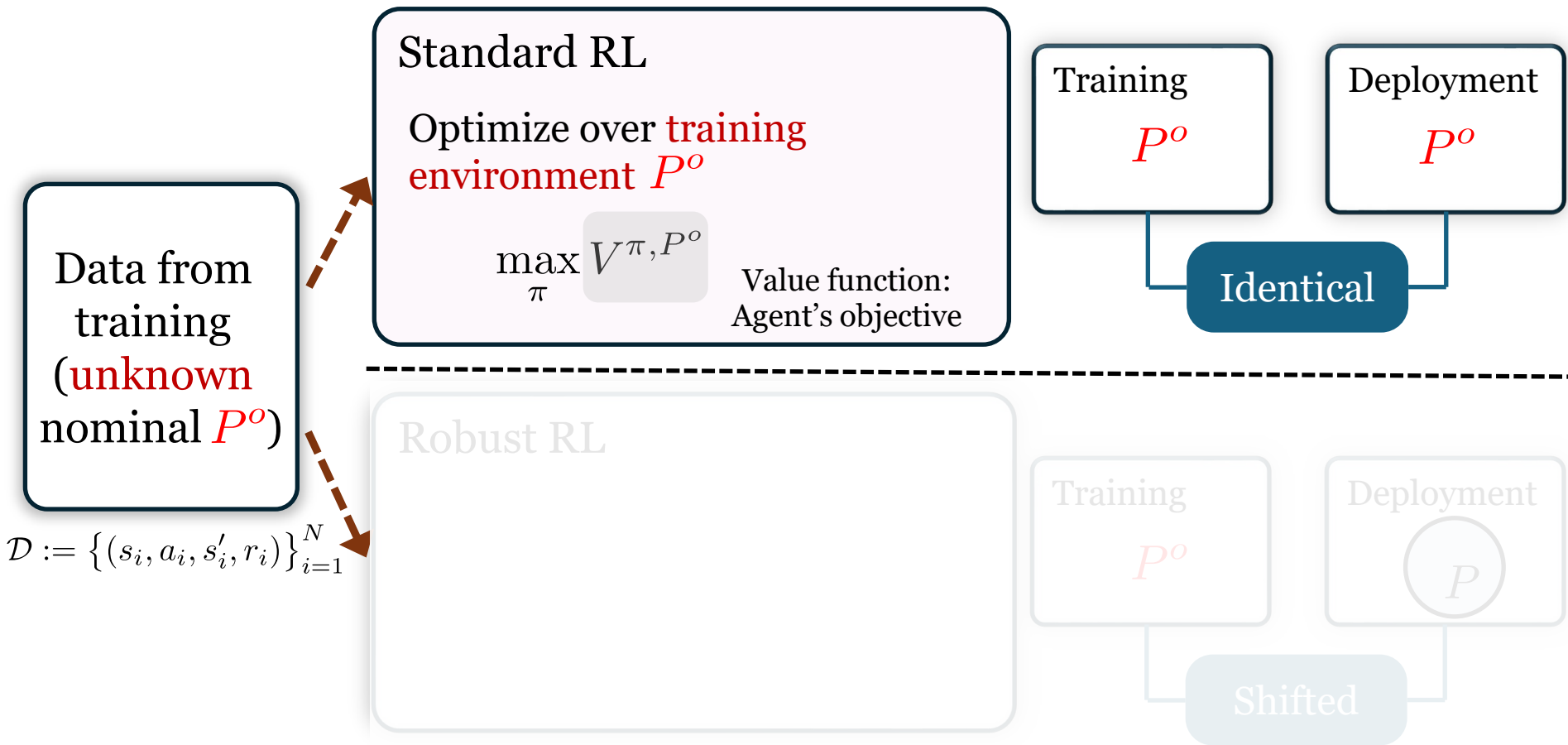


- $\mathcal{S}$: state space, $s \in \mathcal{S}$
- $\mathcal{A}$: action space, $a \in \mathcal{A}$

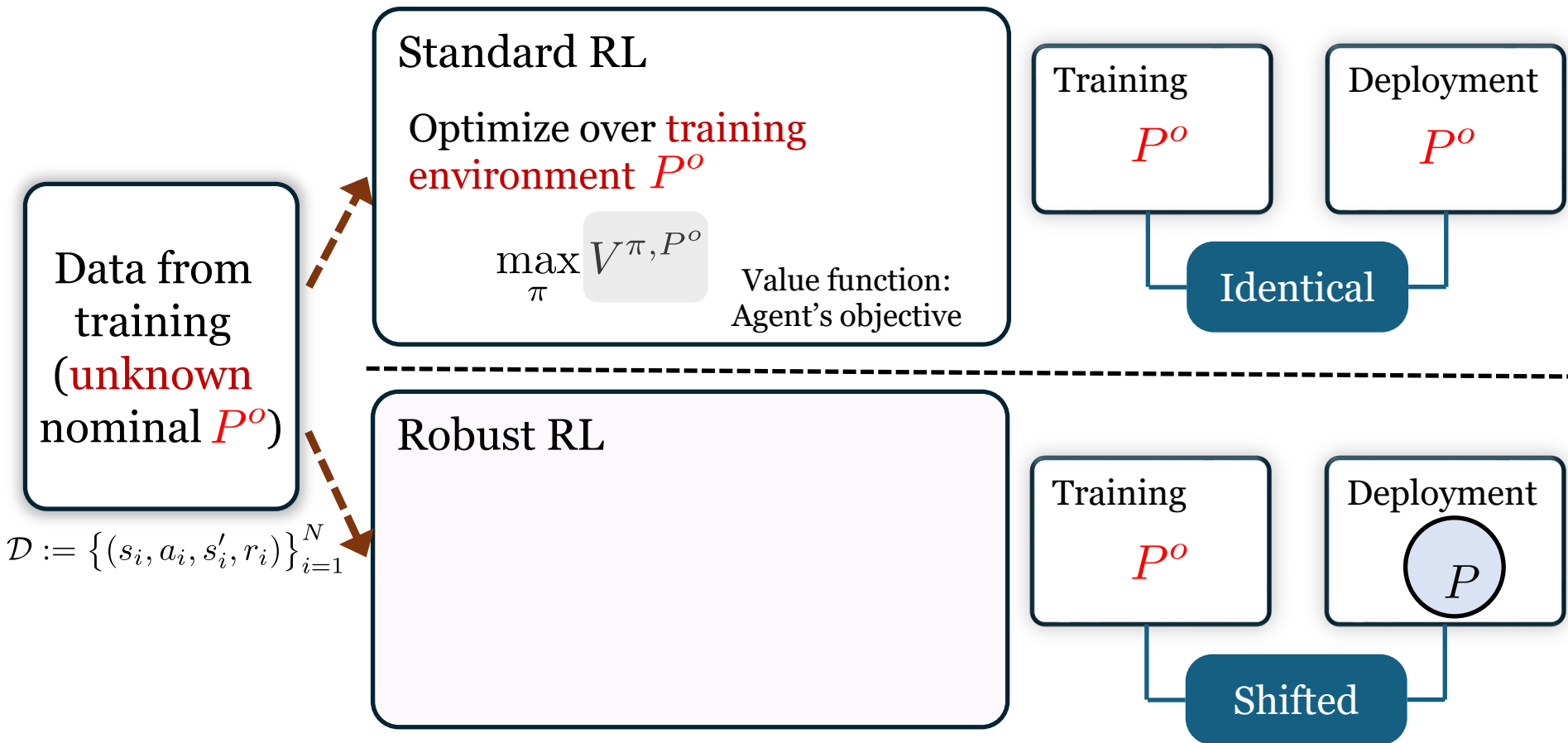
Robust RL (vs. Standard RL)



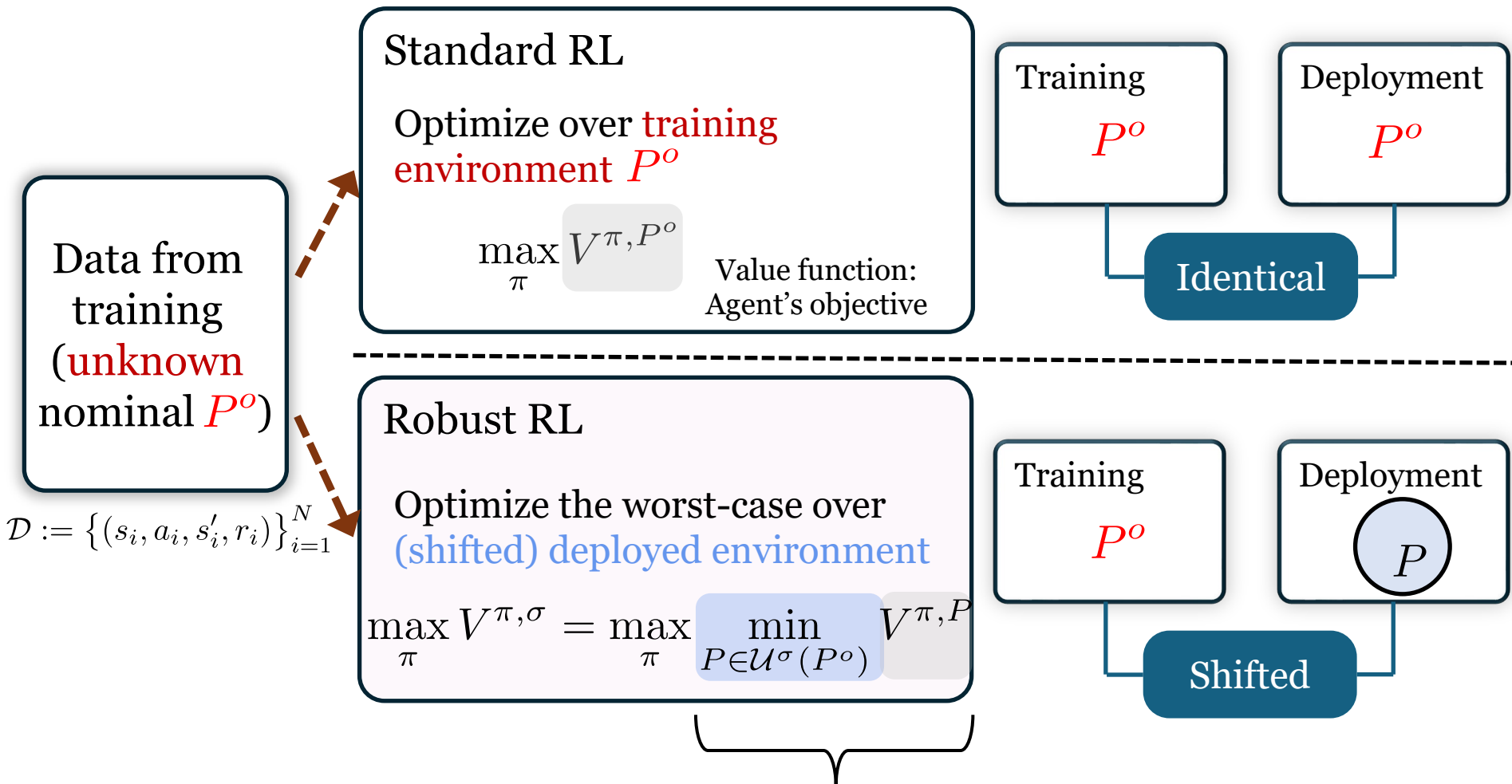
Robust RL (vs. Standard RL)



Robust RL (vs. Standard RL)

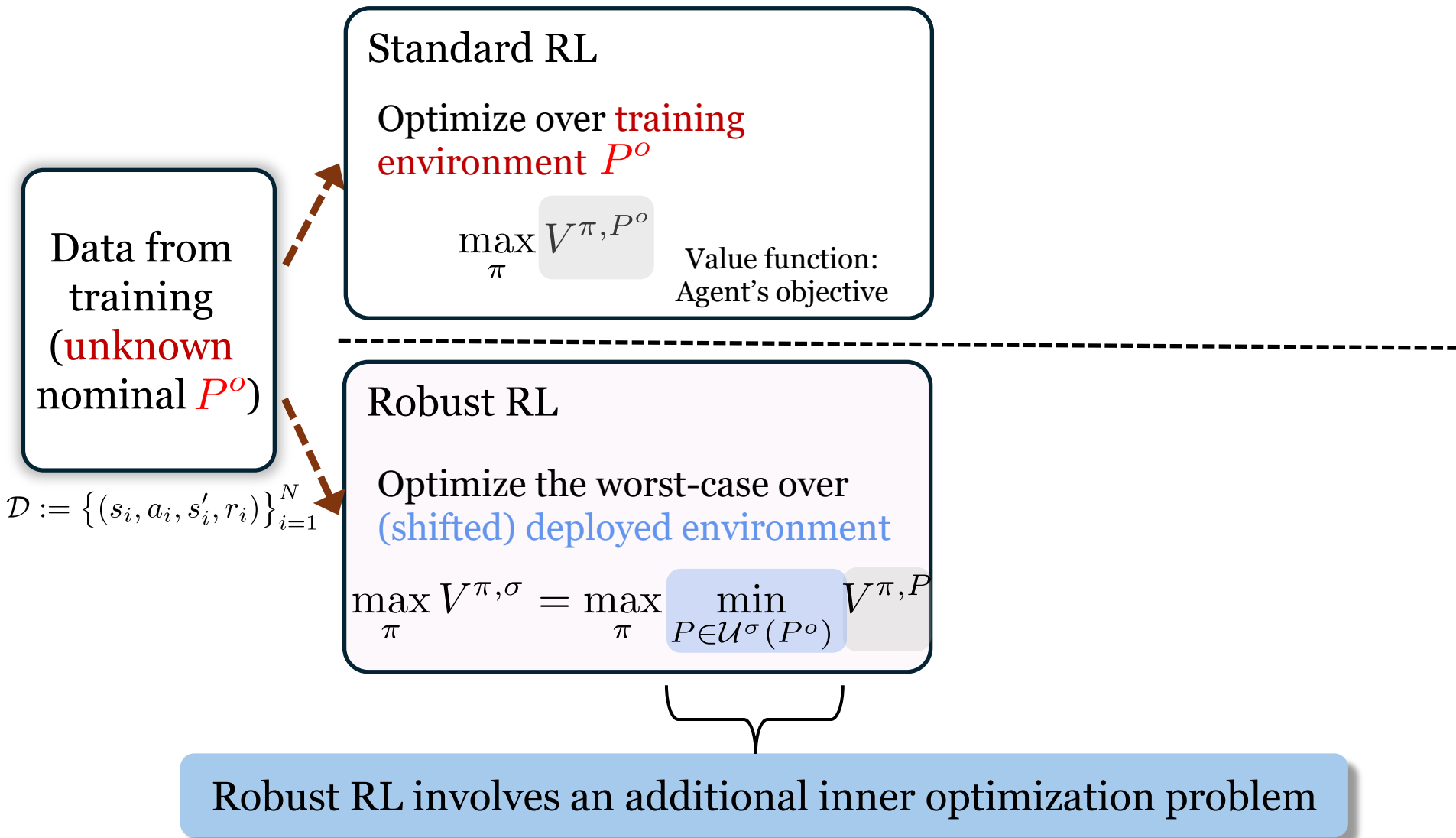


Robust RL (vs. Standard RL)

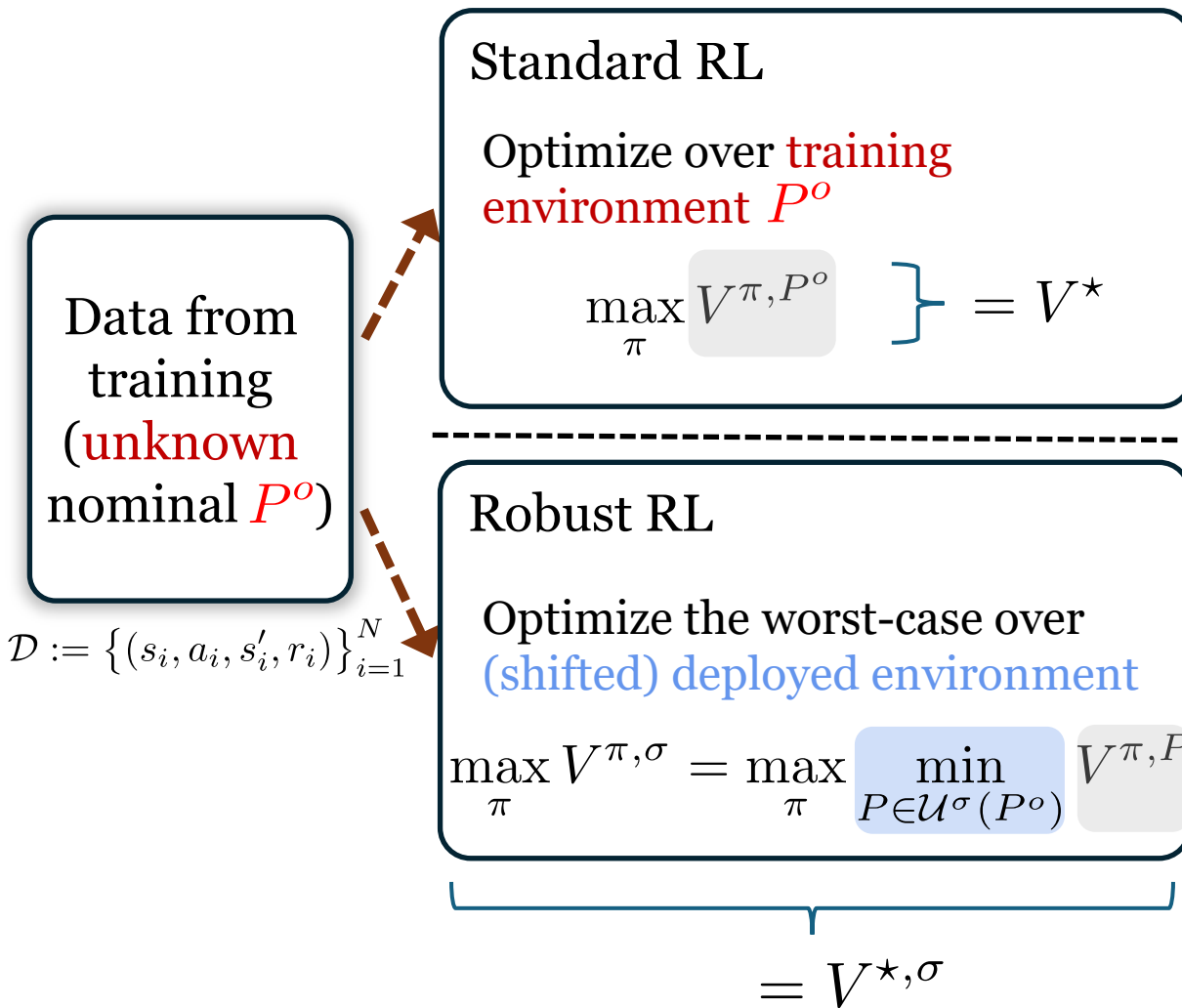


Measures the **worst-case** performance of the policy when the deployed environment $P \in \mathcal{U}^{\sigma}(P^o)$

Robust RL (vs. Standard RL)



Solving a RL problem: finding an ε -optimal policy



Solving a RL problem: finding an ε -optimal policy

Data from training
(**unknown**
nominal P^o)

Standard RL

Optimize over **training**
environment P^o

$$\max_{\pi} V^{\pi, P^o} \quad \} = V^*$$

Goal: find an ε -optimal $\hat{\pi}$:

$$V^* - V^{\hat{\pi}} \leq \varepsilon$$

Robust RL

Optimize the worst-case over
(shifted) deployed environment

$$\max_{\pi} V^{\pi, \sigma} = \max_{\pi} \min_{P \in \mathcal{U}^{\sigma}(P^o)} V^{\pi, P}$$

Goal: find an ε -optimal $\hat{\pi}$:

$$V^{*, \sigma} - V^{\hat{\pi}, \sigma} \leq \varepsilon$$

$$= V^{*, \sigma}$$

$$\mathcal{D} := \{(s_i, a_i, s'_i, r_i)\}_{i=1}^N$$

Data from
training
(**unknown**
nominal P^o)

Standard RL

Optimize over **training**
environment P^o

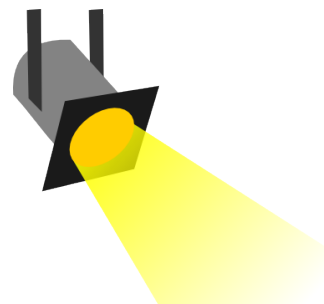
$$\max_{\pi} V^{\pi, P^o}$$

Robust RL

Optimize the worst-case over
(**shifted**) **deployed environment**

$$\max_{\pi} V^{\pi, \sigma} = \max_{\pi} \min_{P \in \mathcal{U}^{\sigma}(P^o)} V^{\pi, P}$$

$$\mathcal{D} := \{(s_i, a_i, s'_i, r_i)\}_{i=1}^N$$

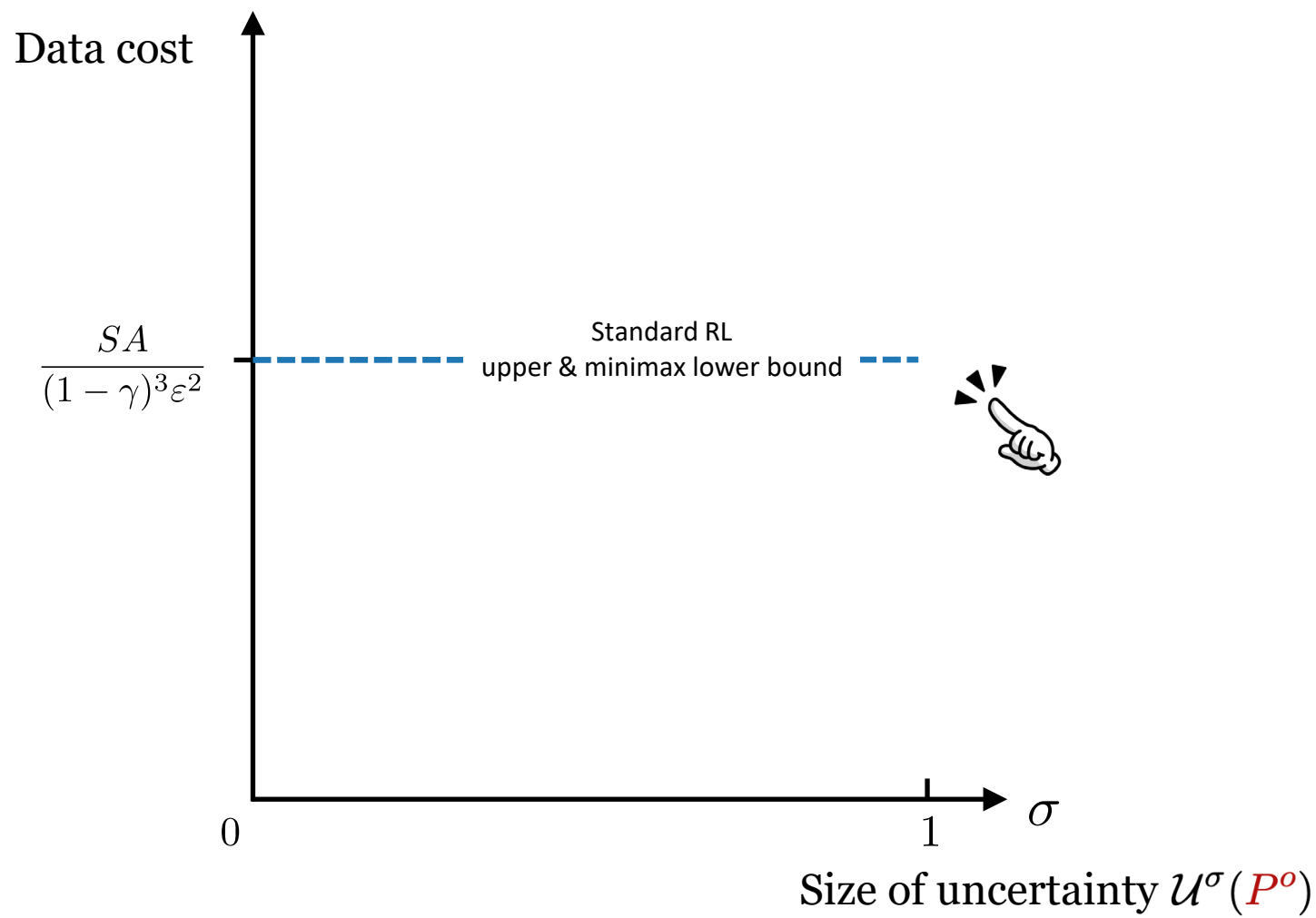


Does solving robust RL
need more data?

Robustness-Data trade-off? Is there a statistical premium
that one needs to pay in quest of additional robustness?

My Work:

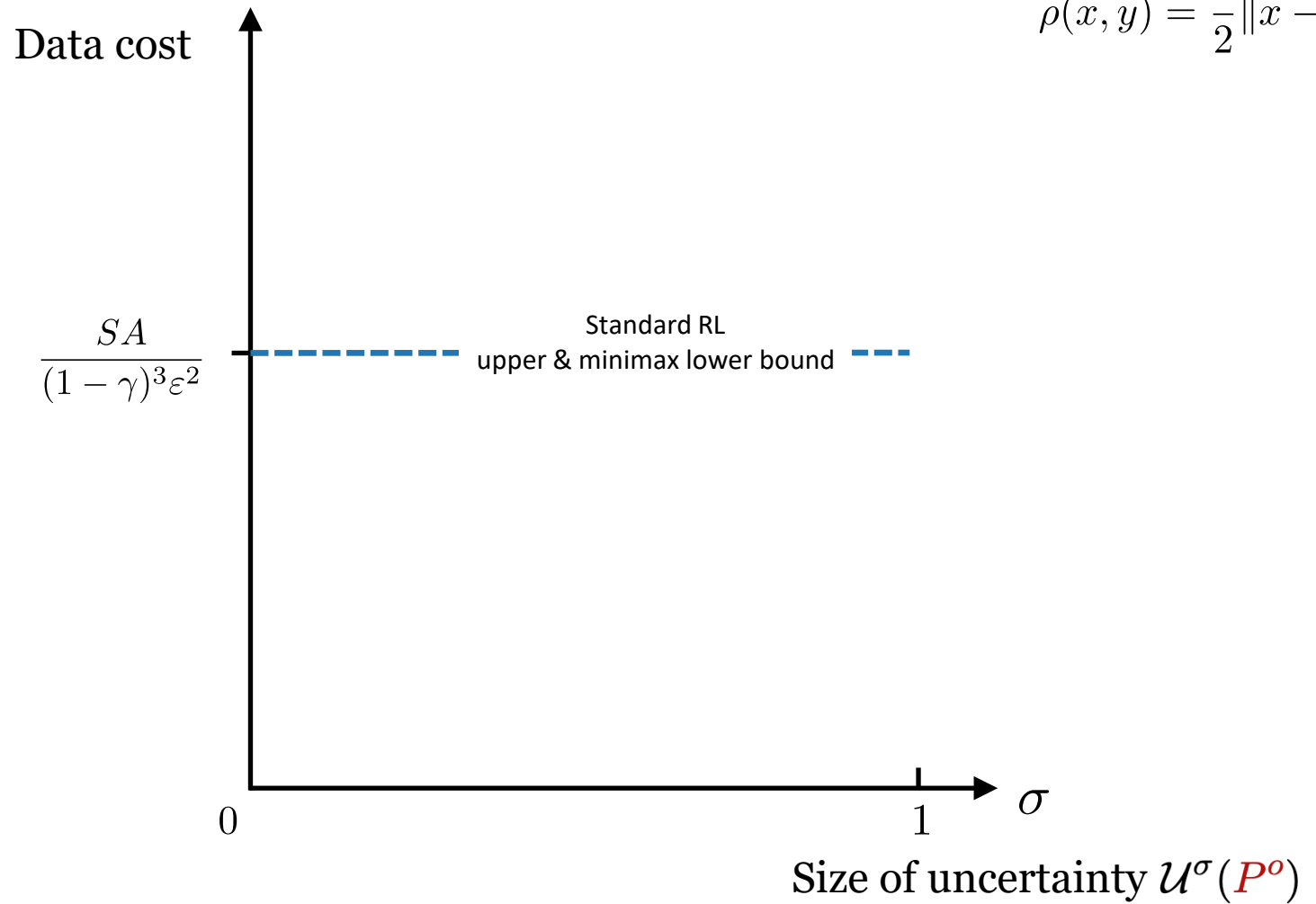
Robustness in RL Without Extra Data Cost



Robust RL with Total Variation

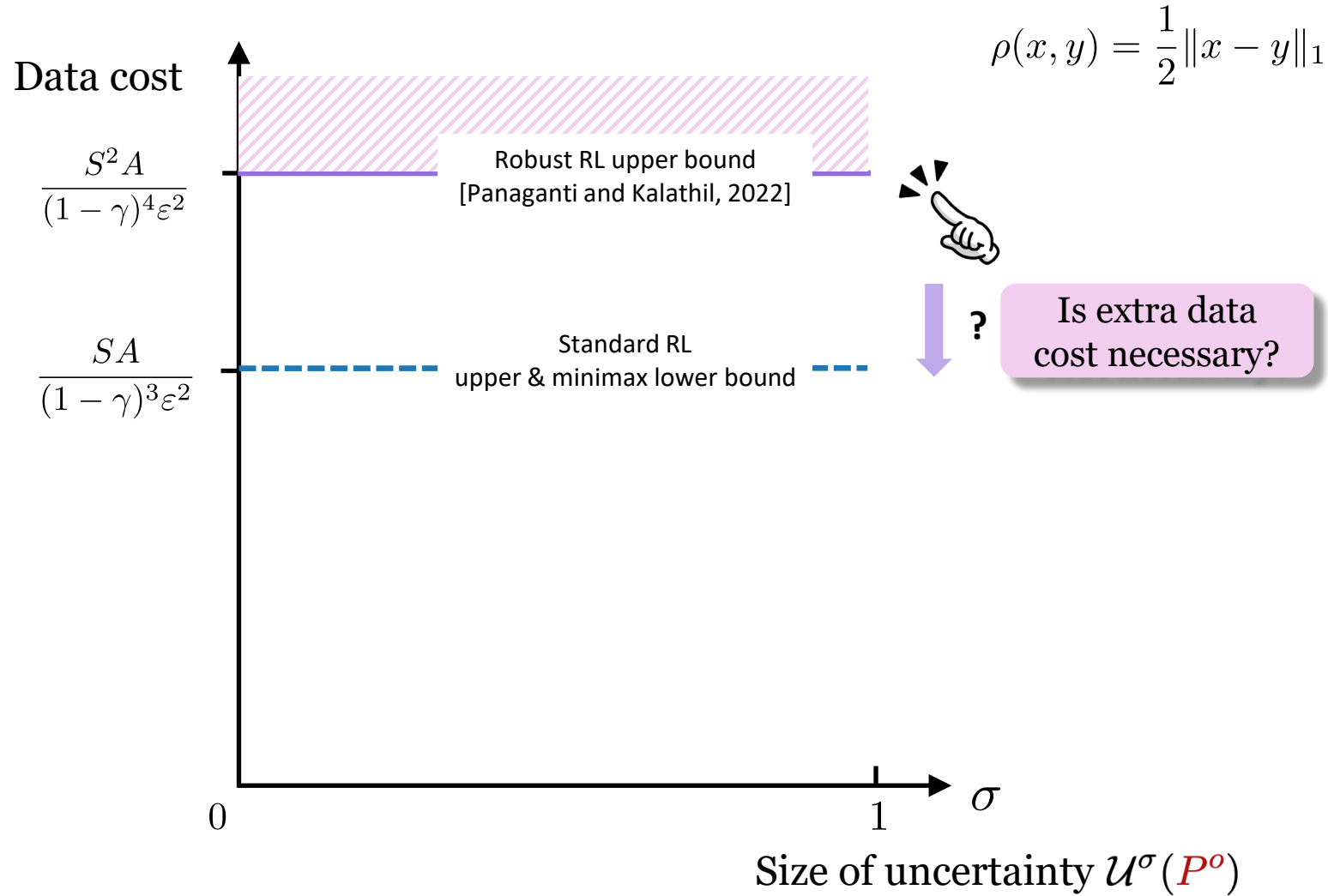
TV:

$$\rho(x, y) = \frac{1}{2} \|x - y\|_1$$



Robust RL with Total Variation

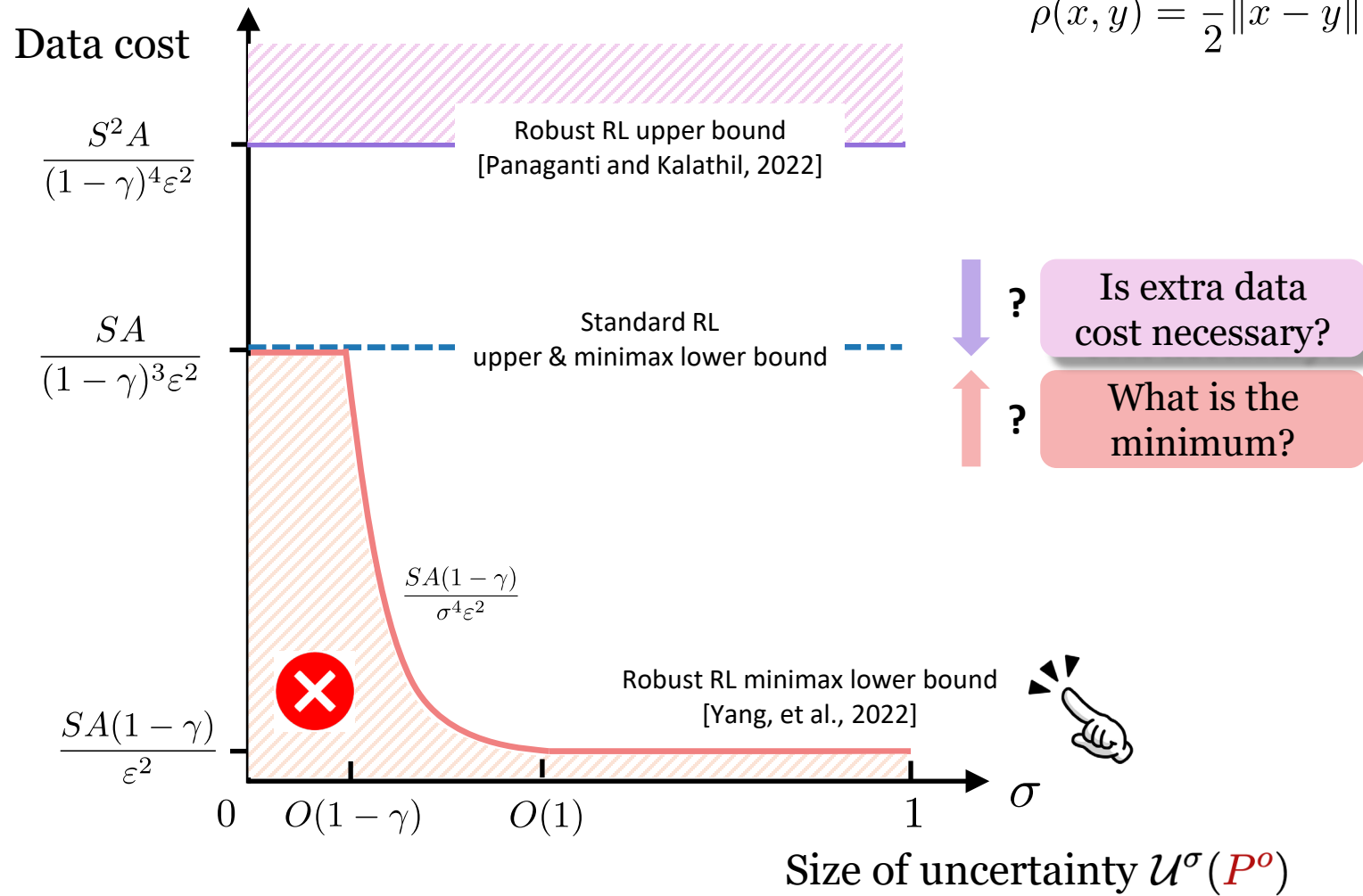
TV:



Robust RL with Total Variation

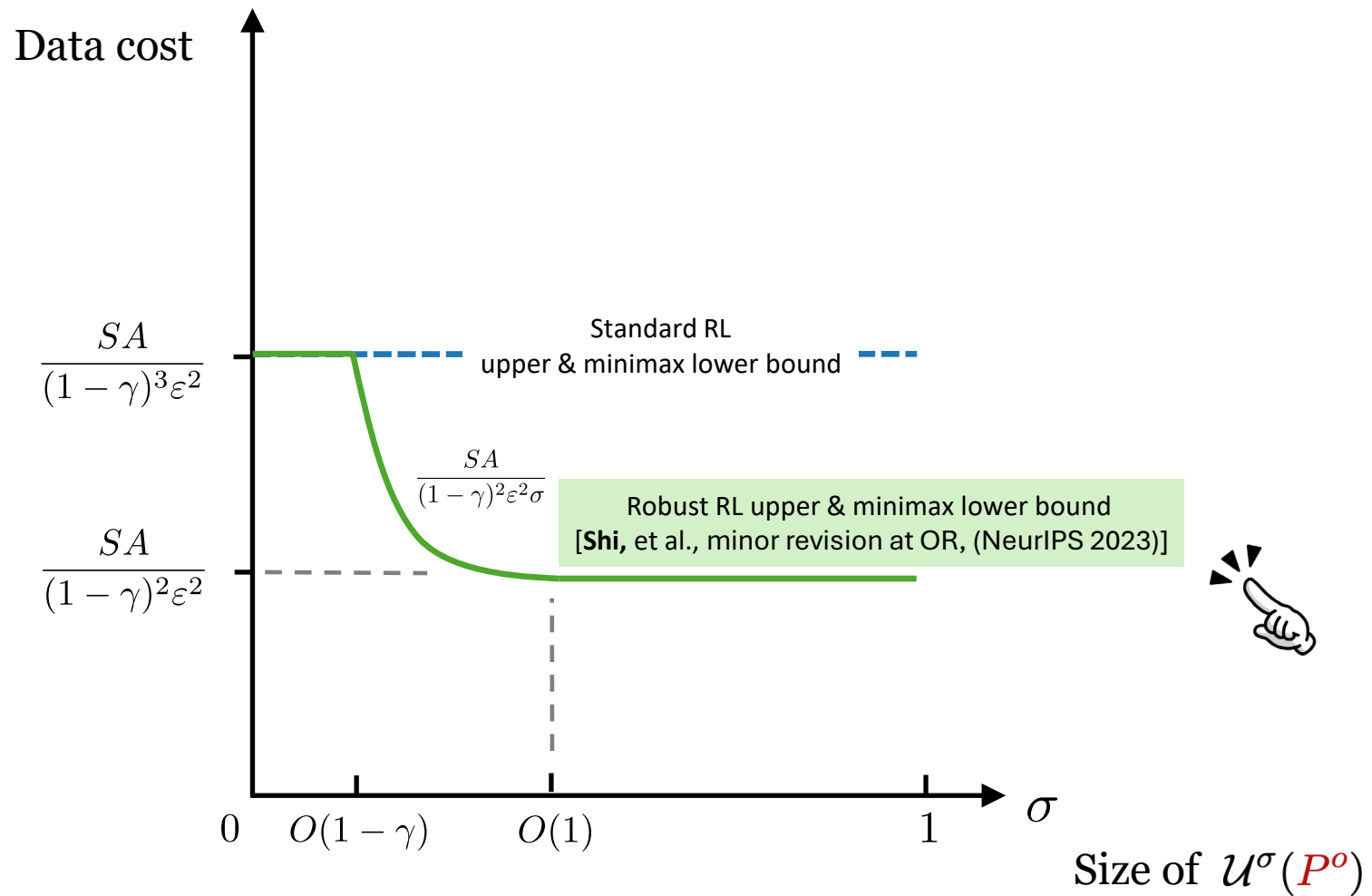
TV:

$$\rho(x, y) = \frac{1}{2} \|x - y\|_1$$



Unclear benchmarking with standard RL

Our Insights on Robust RL with Total Variation



Robust RL Requires No Additional Data Cost than Standard RL

Can we Always Achieve Robustness without Paying More Data?

Robust RL with L_p Norms ($p \geq 2$)

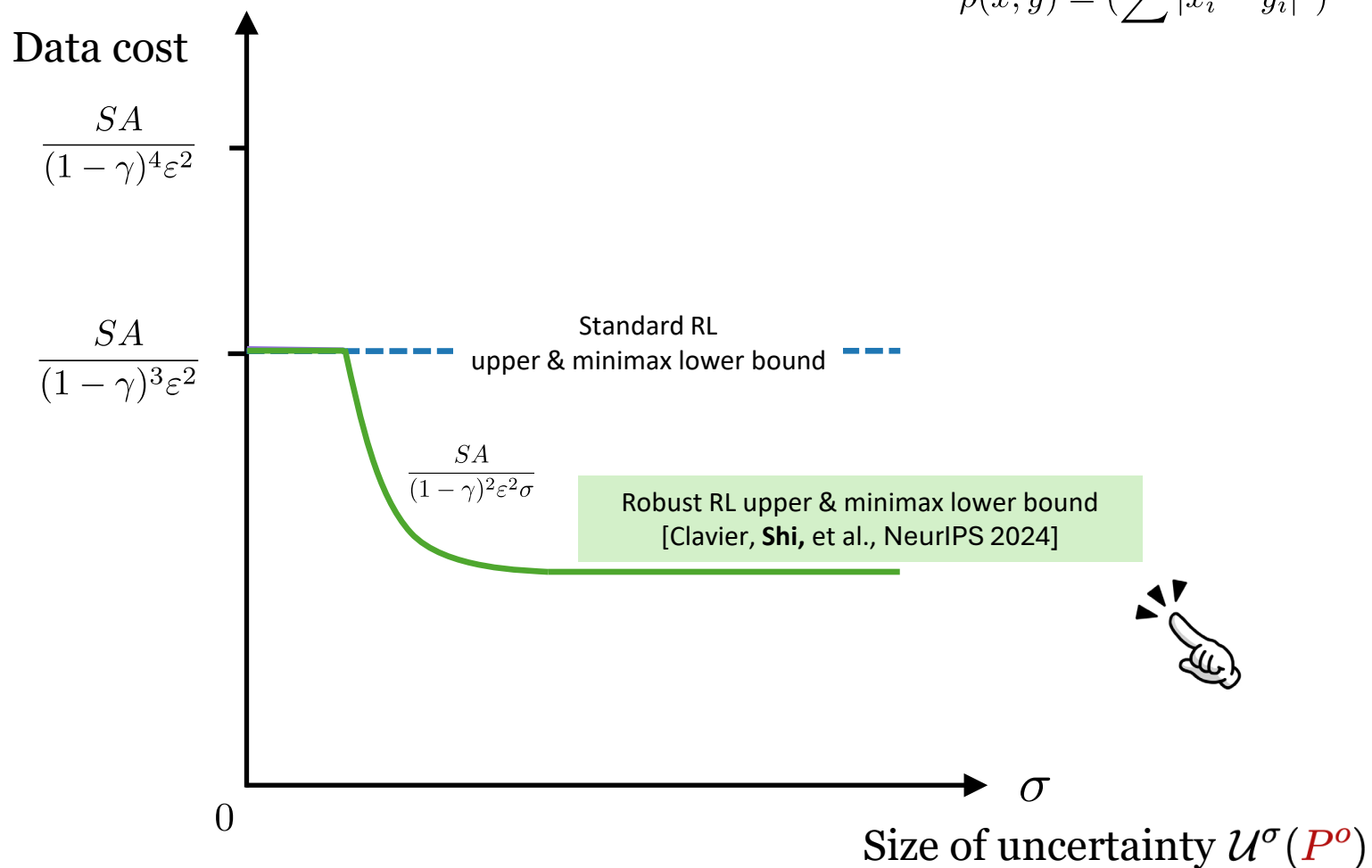
L_p Norms:

$$\rho(x, y) = (\sum |x_i - y_i|^p)^{1/p}$$

Robust RL with L_p Norms ($p \geq 2$)

L_p Norms:

$$\rho(x, y) = (\sum |x_i - y_i|^p)^{1/p}$$



Robust RL Requires No Additional Data Cost in More General Cases

Robust RL with χ^2 Divergence

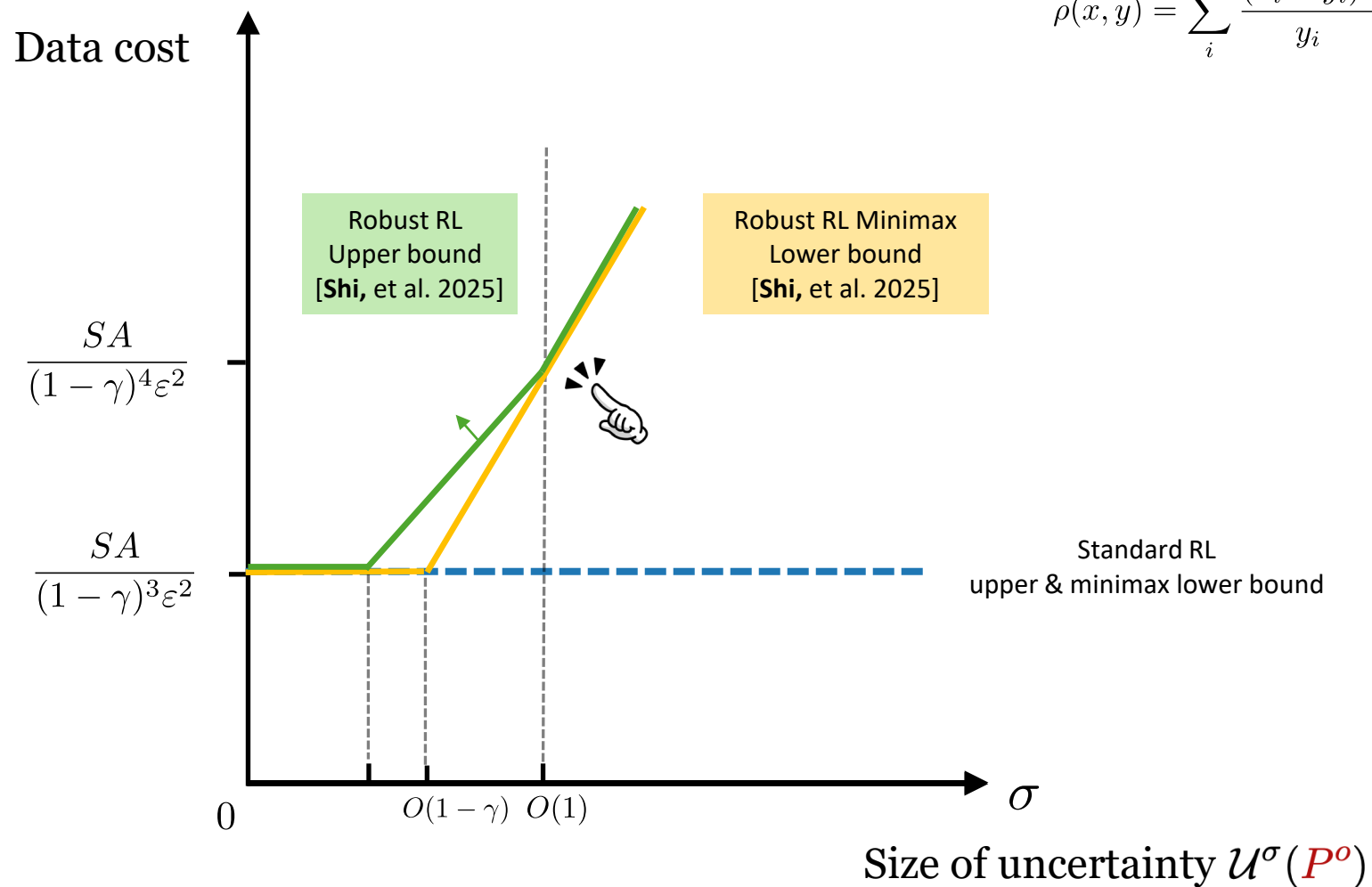
χ^2 Divergence:

$$\rho(x, y) = \sum_i \frac{(x_i - y_i)^2}{y_i}$$

Robust RL with χ^2 Divergence

χ^2 Divergence:

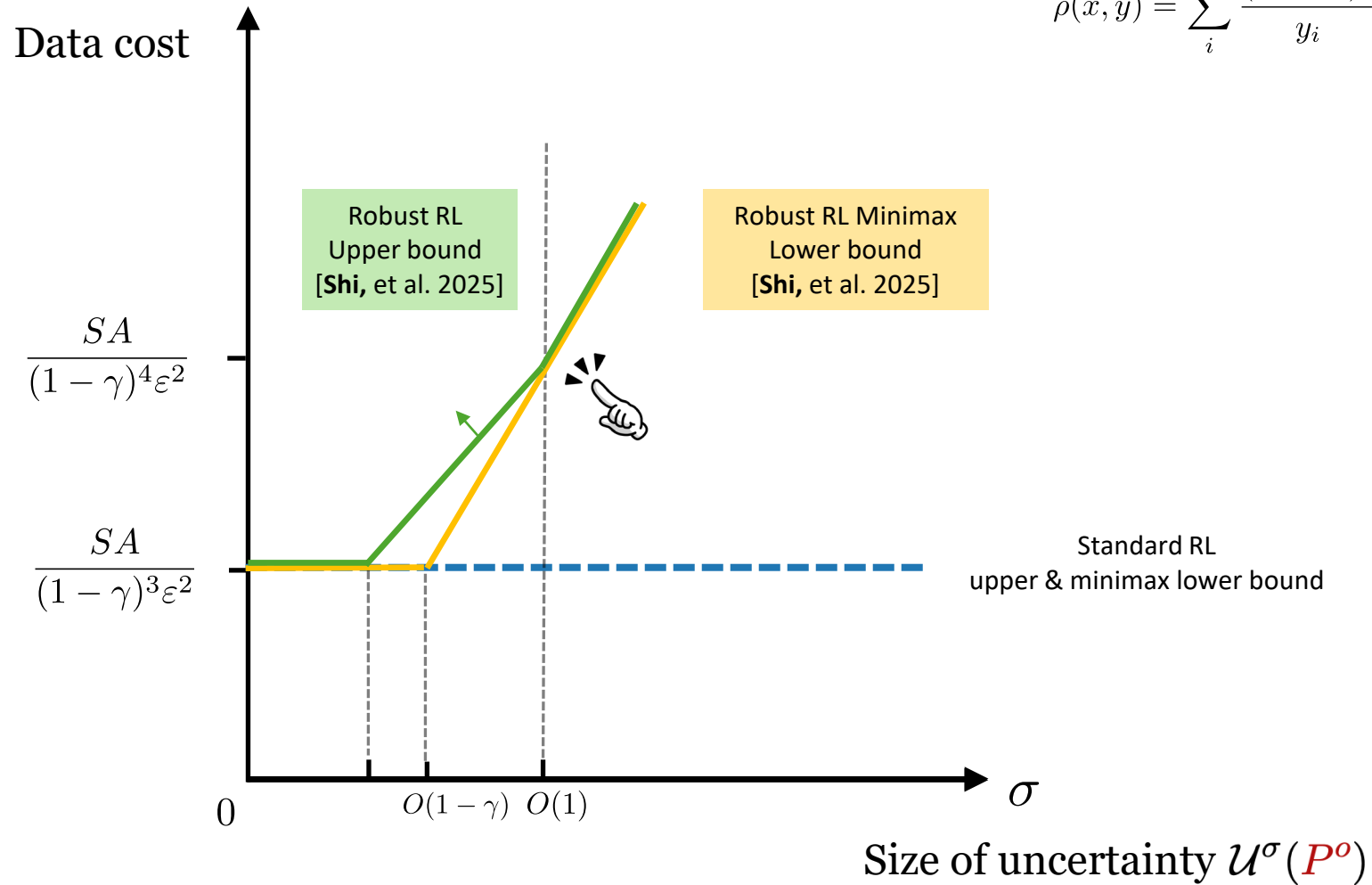
$$\rho(x, y) = \sum_i \frac{(x_i - y_i)^2}{y_i}$$



Robust RL with χ^2 Divergence

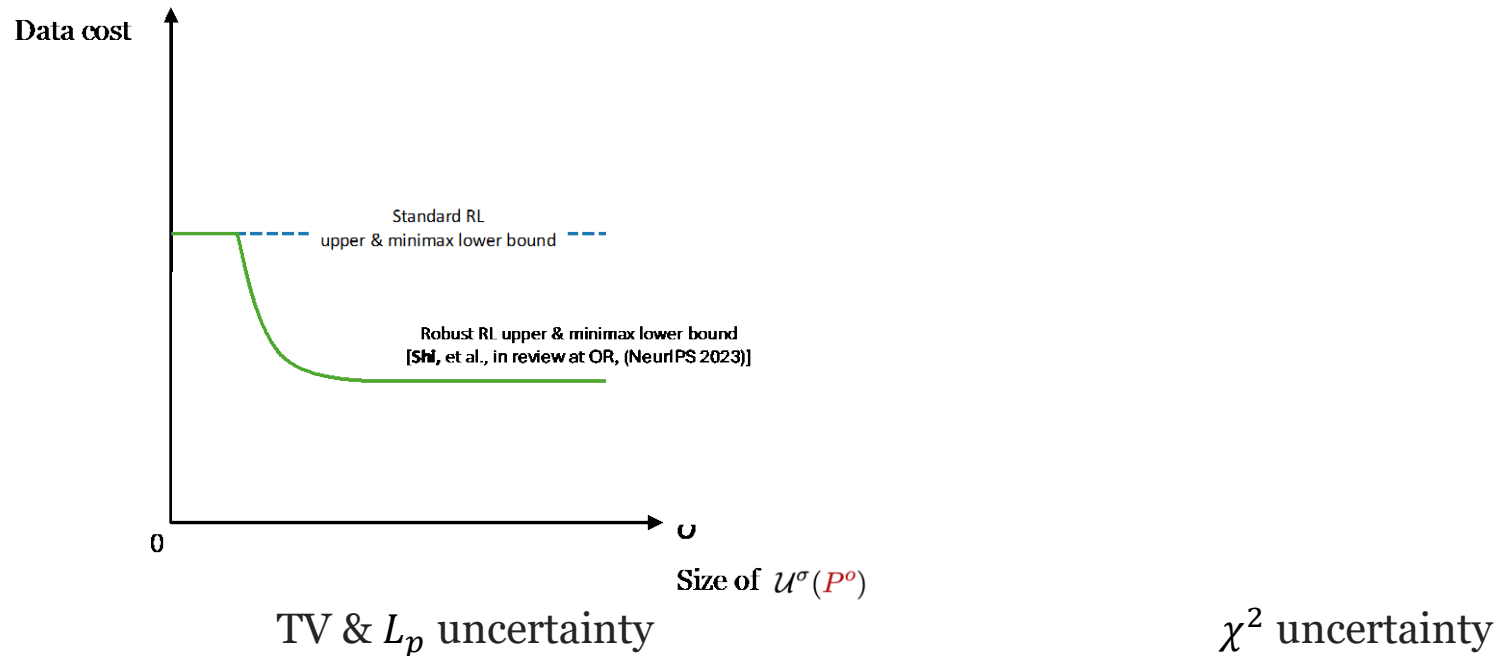
χ^2 Divergence:

$$\rho(x, y) = \sum_i \frac{(x_i - y_i)^2}{y_i}$$



Robust RL Need More Data than Standard RL

Takeaway: Enhancing Robustness in RL Without Extra Data Cost



Be mindful when designing the robust RL formulation!

Technical Challenges and Intuitions

- Goal: control the error terms based on estimate $\hat{P}^o$

$$\text{Standard RL: } \delta_{\text{RL}} = \underbrace{\left| P^o \hat{V} - \hat{P}^0 \hat{V} \right|}_{\text{linear w.r.t. } P^o - \hat{P}^0}$$

$$\text{Robust RL: } \delta_{\text{rob}} = \underbrace{\left| \inf_{\mathcal{P} \in \mathcal{U}_\rho^\sigma(P^o)} \mathcal{P} \hat{V}_{\text{rob}}^{\pi, \sigma} - \inf_{\mathcal{P} \in \mathcal{U}_\rho^\sigma(\hat{P}^0)} \mathcal{P} \hat{V}_{\text{rob}}^{\pi, \sigma} \right|}_{\text{complex form w.r.t. } P^o - \hat{P}^0 \text{ due to inner problem over } \mathcal{U}_\rho^\sigma(\cdot)}$$

Technical Challenges and Intuitions

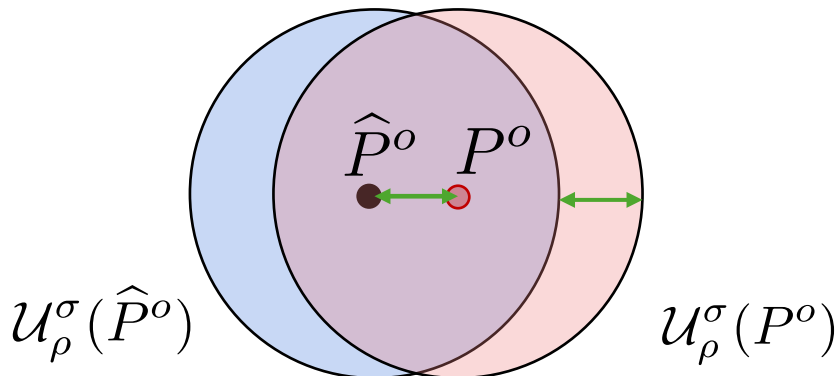
- Goal: control the error terms based on estimate $\hat{P}^o$

Standard RL: $\delta_{\text{RL}} = \underbrace{\left| P^o \hat{V} - \hat{P}^0 \hat{V} \right|}_{\text{linear w.r.t. } P^o - \hat{P}^0}$

Robust RL: $\delta_{\text{rob}} = \underbrace{\left| \inf_{\mathcal{P} \in \mathcal{U}_\rho^\sigma(P^o)} \mathcal{P} \hat{V}_{\text{rob}}^{\pi, \sigma} - \inf_{\mathcal{P} \in \mathcal{U}_\rho^\sigma(\hat{P}^0)} \mathcal{P} \hat{V}_{\text{rob}}^{\pi, \sigma} \right|}_{\text{complex form w.r.t. } P^o - \hat{P}^0 \text{ due to inner problem over } \mathcal{U}_\rho^\sigma(\cdot)}$

TV:

- The error is linear w.r.t. $P^o - \hat{P}^o$: $\delta_{\text{rob}} \approx \left| P^o \hat{V}_{\text{rob}}^{\pi, \sigma} - \hat{P}^0 \hat{V}_{\text{rob}}^{\pi, \sigma} \right|$



Homogeneous instance-independent geometry of the uncertainty set

$$\rho(P, P^o) = \sum_s |P(s) - P^o(s)|$$

Technical Challenges and Intuitions

- Goal: control the error terms based on estimate $\hat{P}^o$

Standard RL: $\delta_{\text{RL}} = \underbrace{\left| P^o \hat{V} - \hat{P}^0 \hat{V} \right|}_{\text{linear w.r.t. } P^o - \hat{P}^0}$

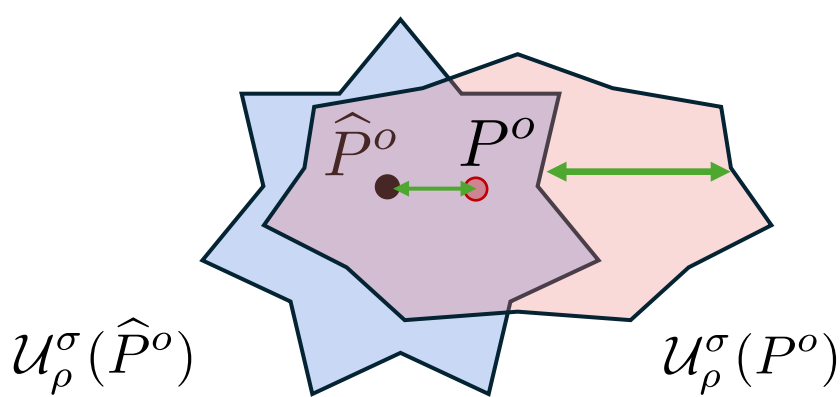
Robust RL: $\delta_{\text{rob}} = \underbrace{\left| \inf_{\mathcal{P} \in \mathcal{U}_\rho^\sigma(P^o)} \mathcal{P} \hat{V}_{\text{rob}}^{\pi, \sigma} - \inf_{\mathcal{P} \in \mathcal{U}_\rho^\sigma(\hat{P}^0)} \mathcal{P} \hat{V}_{\text{rob}}^{\pi, \sigma} \right|}_{\text{complex form w.r.t. } P^o - \hat{P}^0 \text{ due to inner problem over } \mathcal{U}_\rho^\sigma(\cdot)}$

- **TV:**

- The error is linear w.r.t. $P^o - \hat{P}^o$: $\delta_{\text{rob}} \approx \left| P^o \hat{V}_{\text{rob}}^{\pi, \sigma} - \hat{P}^0 \hat{V}_{\text{rob}}^{\pi, \sigma} \right|$
- The range of the worst-case objective $\hat{V}_{\text{rob}}^{\pi, \sigma}$ is smaller than $\hat{V}$

$$\|\hat{V}_{\text{rob}}^{\pi, \sigma}\|_\infty \leq \|\hat{V}\|_\infty$$

- χ^2 Divergence :
 - The error is **non-linear and sensitive** w.r.t. $P^o - \hat{P}^o$

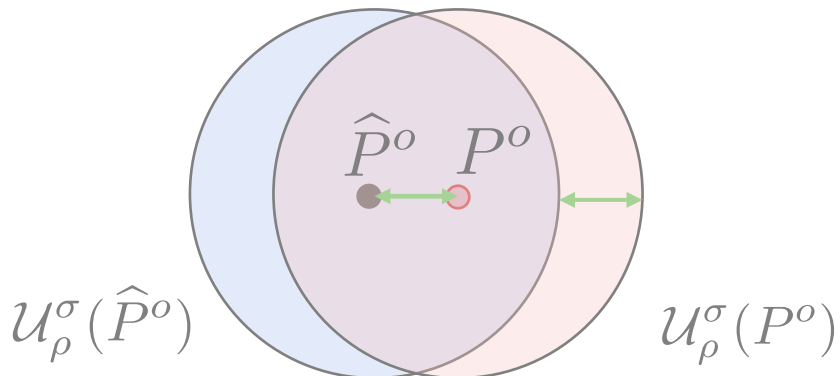


Heterogeneous instance-dependent geometry of the uncertainty set

$$\rho(P, P^o) = \sum_s \frac{(P(s) - P^o(s))^2}{P^o(s)}$$

- TV:

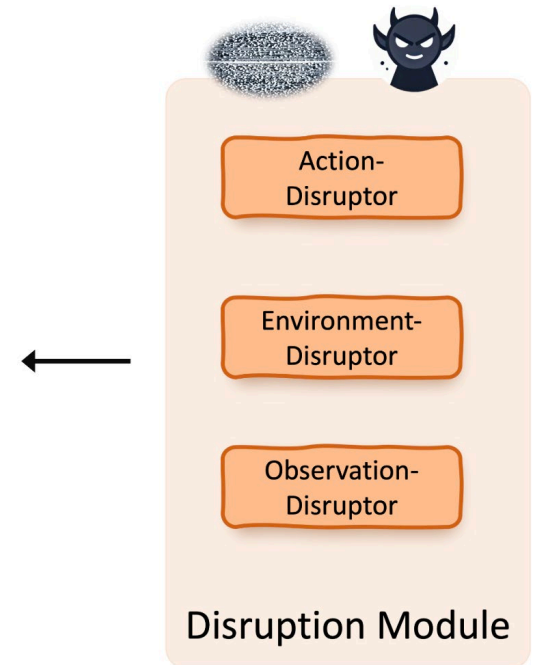
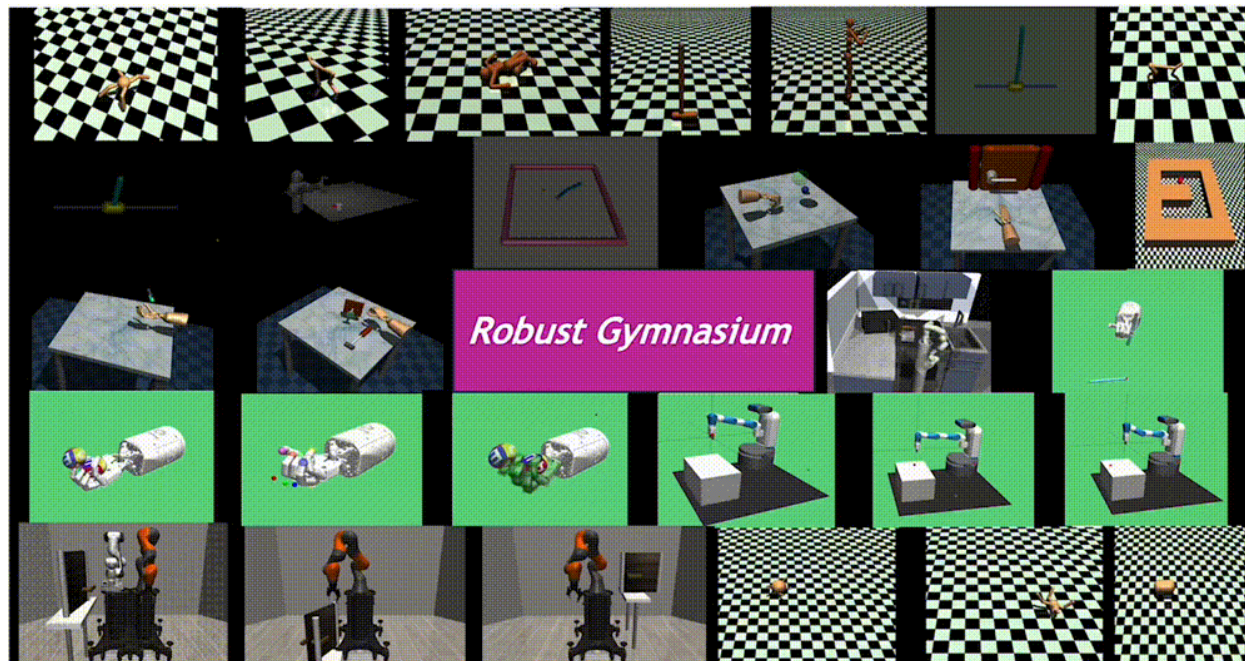
- The error is linear w.r.t. $P^o - \hat{P}^o$: $\delta_{\text{rob}} \approx \left| P^o \hat{V}_{\text{rob}}^{\pi, \sigma} - \hat{P}^o \hat{V}_{\text{rob}}^{\pi, \sigma} \right|$



Homogeneous instance-independent geometry of the uncertainty set

$$\rho(P, P^o) = \sum_s |P(s) - P^o(s)|$$

Going Beyond Theory: A Unified Robust RL Benchmark

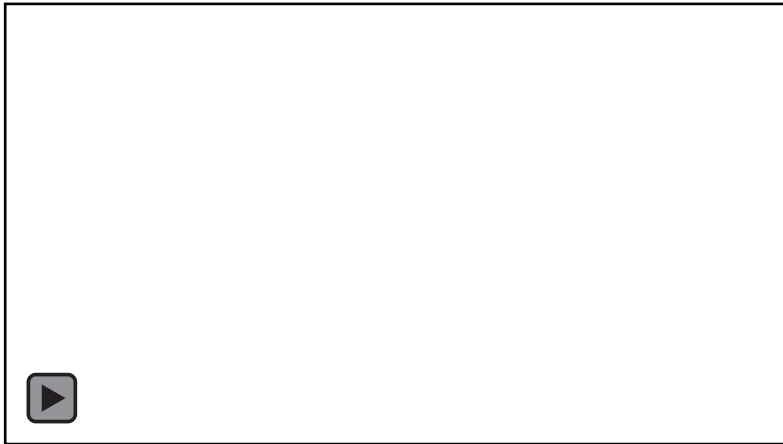


Robust Gymnasium: A Unified Modular Benchmark for Robust Reinforcement Learning

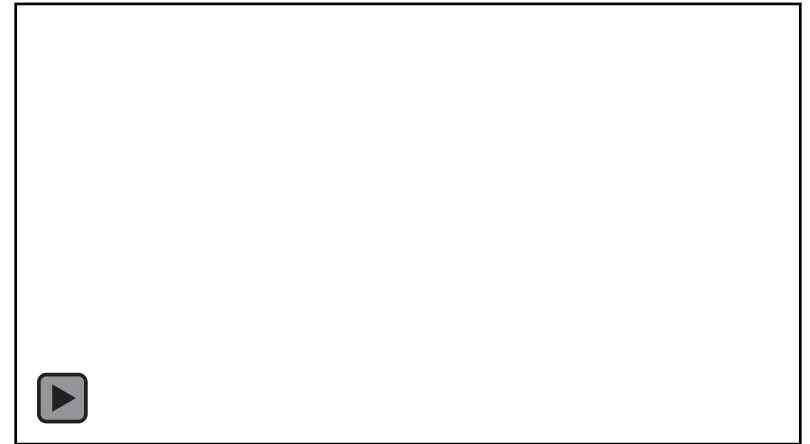
[Gu*, Shi*, et al., ICLR '25]

Going Beyond Theory: Robust RL Applications

Robust zero-shot safe sim-to-real transfer for robotics



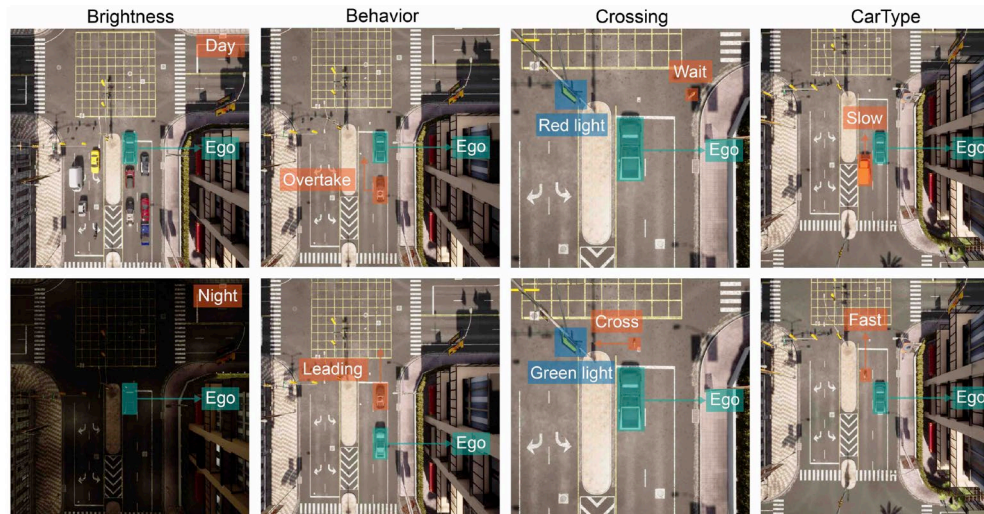
SPiDR (Ours)



Baseline

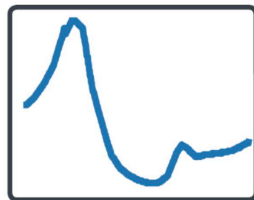
Going Beyond Theory: Robust RL Applications

Robust autonomous driving against unobserved uncertainty¹



Robust resource allocation against varying user demand in 5G²

Demand in training



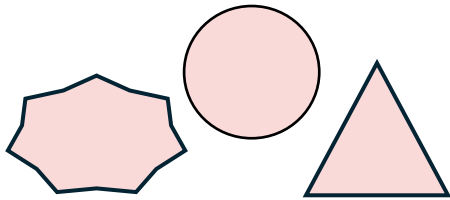
Demand in deployment



¹[Ding*, **Shi***, et al., NeurIPS '23] ²[Wang, **Shi**, et al., UAI '24]

Further Opportunities of Robust RL

Foundation and methodology



Different uncertainty sets



Data collection mechanisms



Multi-agent systems

...

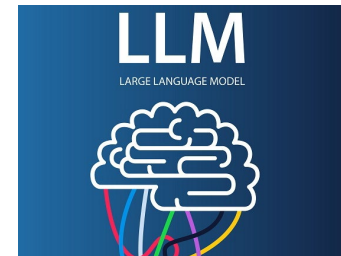
Applications for uncertain environments



Robotic safety



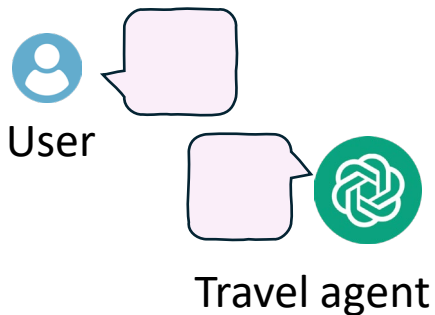
Resource allocation



Post-training

...

Agentic RL: Optimizing for More Than Just Performance

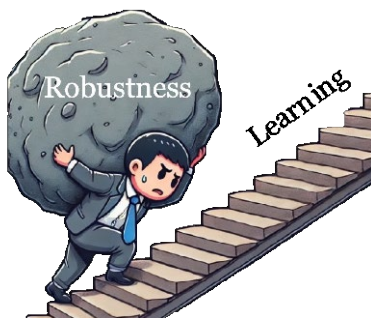


Task performance



User satisfaction

I: Agentic RL for multi-turn agent-user interaction:
behavior regularization → Pareto optimal



Robust performance



Data efficiency

II: Robust RL
without sacrificing data efficiency, even enhancing it

Acknowledgments

A non-exhaustive list

Caltech

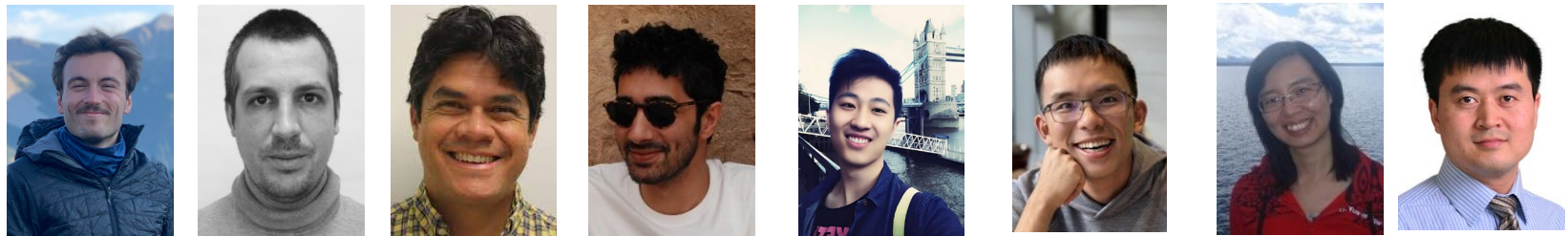
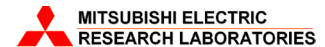
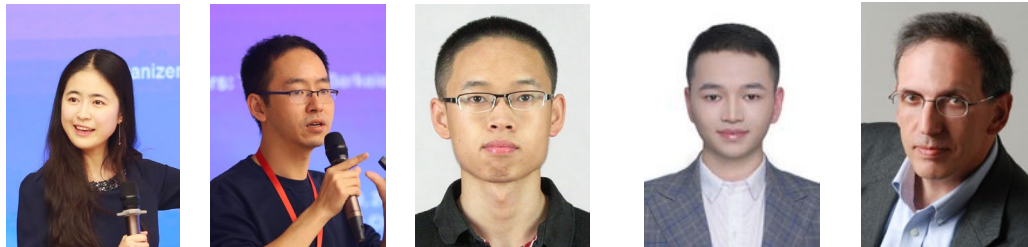


Yale

Carnegie Mellon University

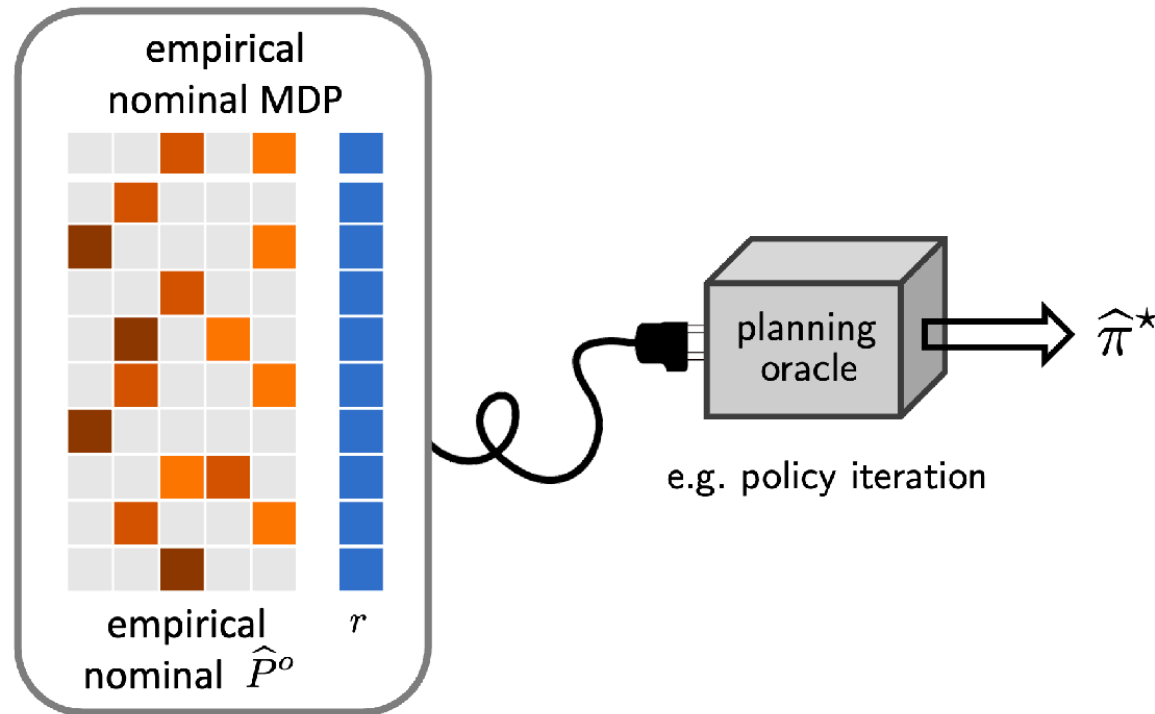


Berkeley
EECS



Fundamental Algorithm: Model-Based RL

— Azar et al., 2013, Agarwal et al., 2019



Find policy based on the **empirical** MDP
using, e.g., policy iteration $(\hat{P}^o, r)$

Planning Oracle: Distributionally Robust Value Iteration

(Iyengar. '05, Nilim and El Ghaoui. '05)

Planning by **distributionally robust value iteration (DRVI)**:

$$\hat{Q}(s, a) \leftarrow r(s, a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma(\hat{P}_{s,a}^o)} \langle P_{s,a}, \hat{V} \rangle,$$

where $\hat{V}(s) = \max_a \hat{Q}(s, a)$.

Planning Oracle: Distributionally Robust Value Iteration

(Iyengar. '05, Nilim and El Ghaoui. '05)

Planning by **distributionally robust value iteration (DRVI)**:

$$\hat{Q}(s, a) \leftarrow r(s, a) + \gamma \inf_{P_{s,a} \in \mathcal{U}^\sigma(\hat{P}_{s,a}^o)} \langle P_{s,a}, \hat{V} \rangle,$$

where $\hat{V}(s) = \max_a \hat{Q}(s, a)$.

Involves an additional inner optimization problem
($\inf_{P_{s,a} \in \mathcal{U}^\sigma(\hat{P}_{s,a}^o)}$) compared to standard RL

Study of robust RL with different divergence function ρ

$$\mathcal{U}^\sigma(P^o) = \{P : \rho(P, P^o) \leq \sigma\}$$

- [Zhou, et al., 2021]
- [Panaganti and Kalathil, 2022]
- [Yang et al., 2022]
- [Dong et al., 2022]
- [Ma et al., 2022]
- [**Shi** and Chi, 2022]
- [Xu et al., 2023]
- [Clavier et al., 2023]
- ...