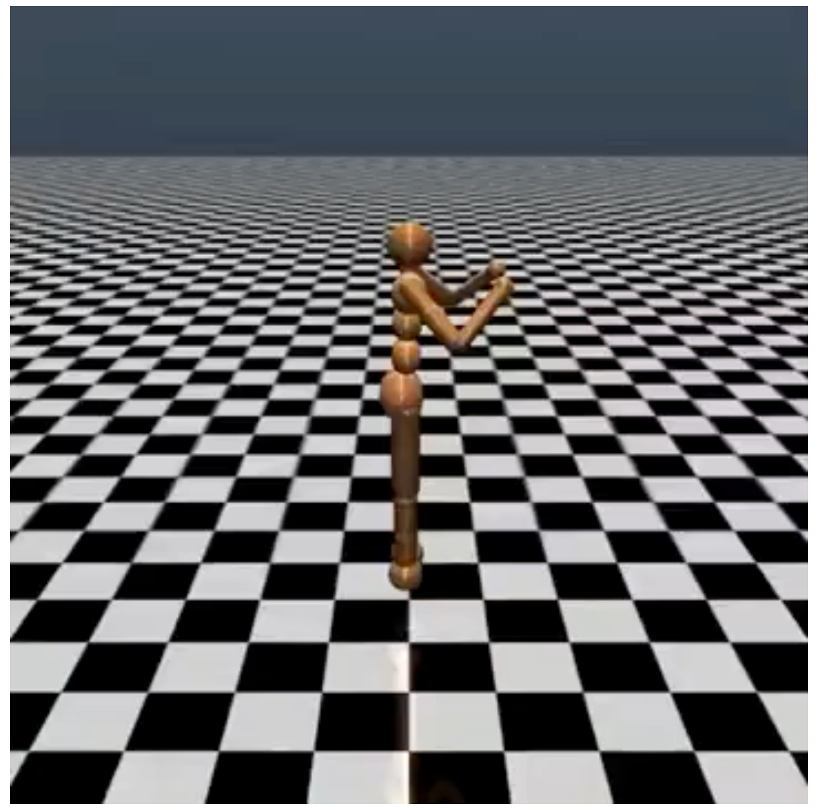
NeORL

Efficient Exploration for Nonepisodic RL

Bhavya Sukhija, Lenart Treven,
Florian Dörfler, Stelian Coros, Andreas Krause



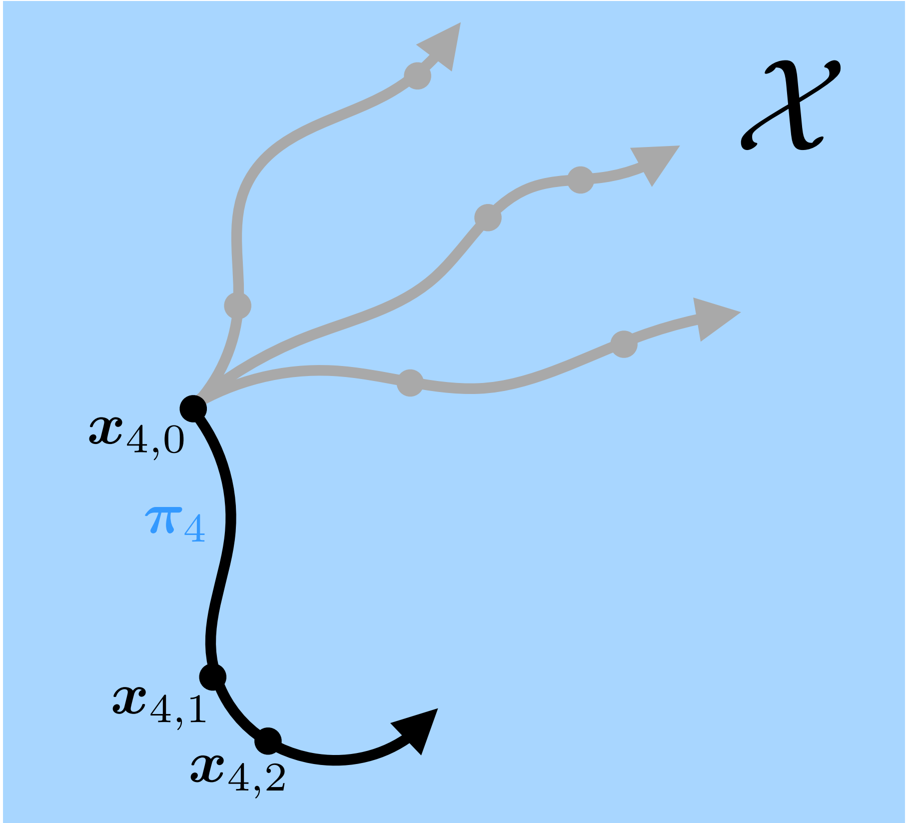

Thousands of robots learn
to walk in parallel…

Model Based RL (episodic)

- Episodes $n = 1, \ldots, N$.

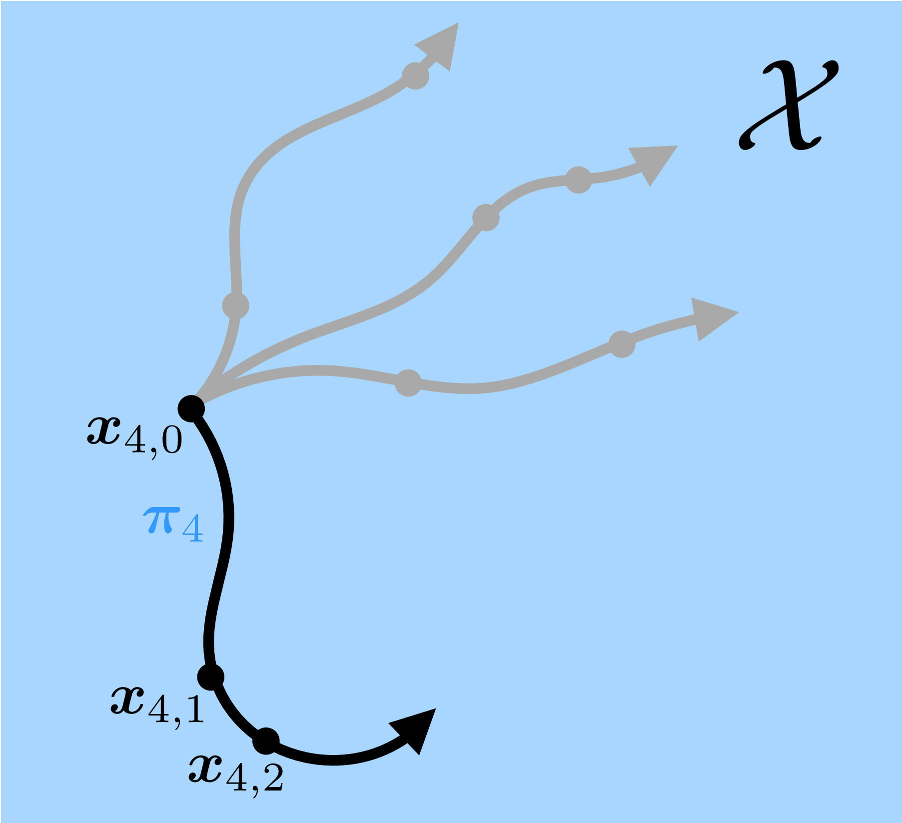
Model Based RL (episodic)

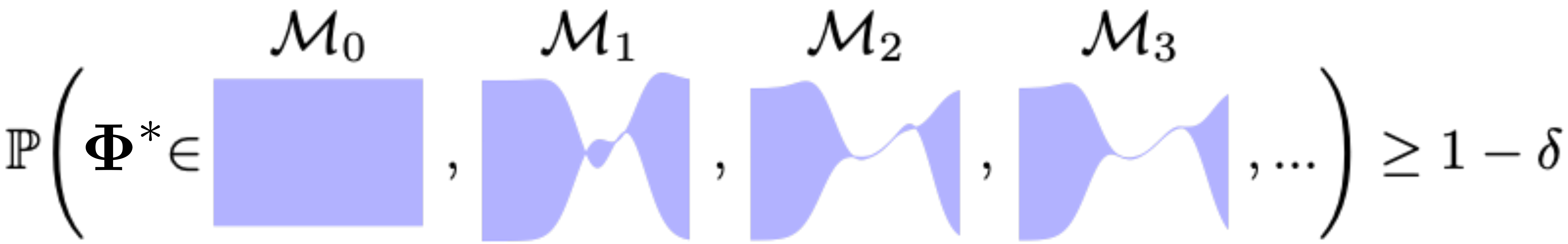
- Episodes $n = 1, \ldots, N$.
- Execute policy $\boldsymbol{\pi}_n$ and collect measurements $(\boldsymbol{x}_{n,0}, b_{n,0}), \ldots, (\boldsymbol{x}_{n,k_n}, b_{n,k_n})$



Model Based RL (episodic)

- Episodes $n = 1, \ldots, N$.
- Execute policy $\boldsymbol{\pi}_n$ and collect measurements $(\boldsymbol{x}_{n,0}, b_{n,0}), \ldots, (\boldsymbol{x}_{n,k_n}, b_{n,k_n})$
- Prepare dataset $\mathcal{D}_n = \{(\boldsymbol{z}_{n,1}, \boldsymbol{y}_{n,1}), \ldots, (\boldsymbol{z}_{n,k_n}, \boldsymbol{y}_{n,k_n})\}$, where $\boldsymbol{z}_{n,i} = (\boldsymbol{x}_{n,i-1}, \boldsymbol{\pi}_n(\boldsymbol{x}_{n,i-1}))$ and $\boldsymbol{y}_{n,i} = (\boldsymbol{x}_{n,i}, b_{n,i})$.

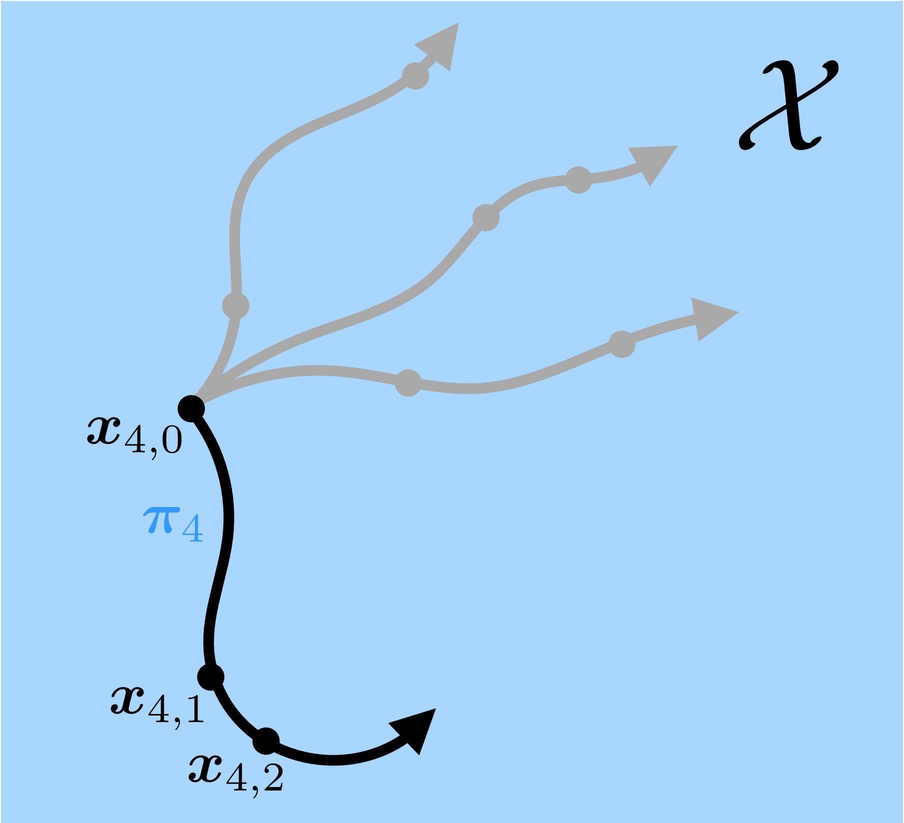


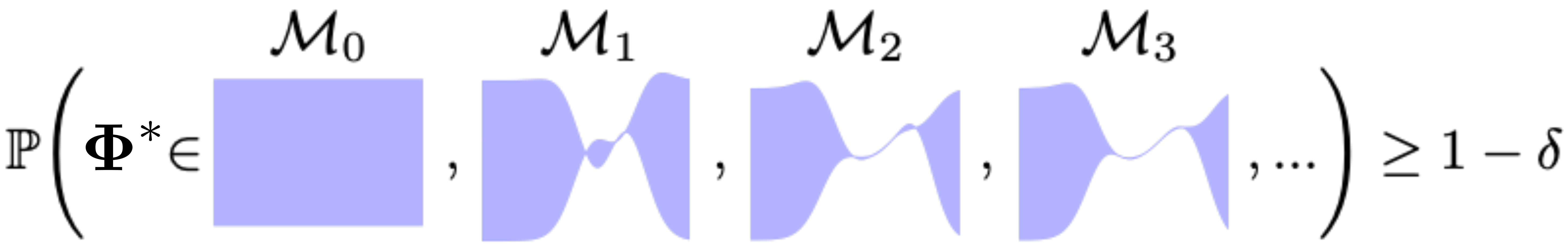


Model Based RL (non-episodic)

Iterations

- ~~Episodes~~ $n = 1, \ldots, N$.
- Execute policy $\boldsymbol{\pi}_n$ and collect measurements $(\boldsymbol{x}_{n,0}, b_{n,0}), \ldots, (\boldsymbol{x}_{n,k_n}, b_{n,k_n})$
- Prepare dataset $\mathcal{D}_n = \{(\boldsymbol{z}_{n,1}, \boldsymbol{y}_{n,1}), \ldots, (\boldsymbol{z}_{n,k_n}, \boldsymbol{y}_{n,k_n})\}$, where $\boldsymbol{z}_{n,i} = (\boldsymbol{x}_{n,i-1}, \boldsymbol{\pi}_n(\boldsymbol{x}_{n,i-1}))$ and $\boldsymbol{y}_{n,i} = (\boldsymbol{x}_{n,i}, b_{n,i})$.

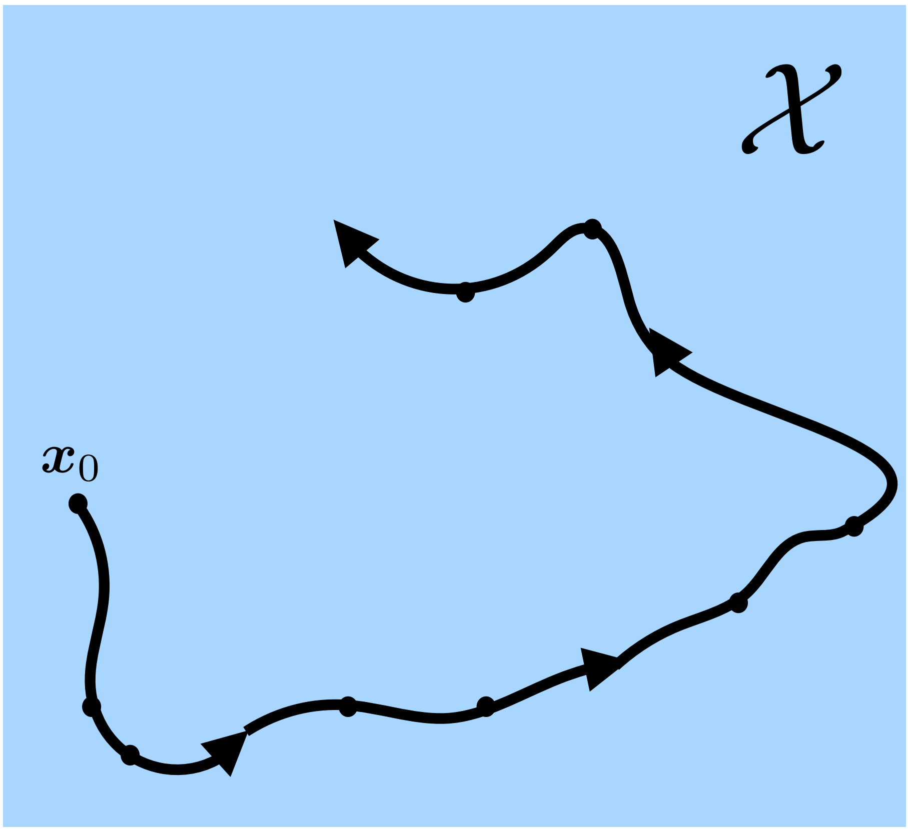


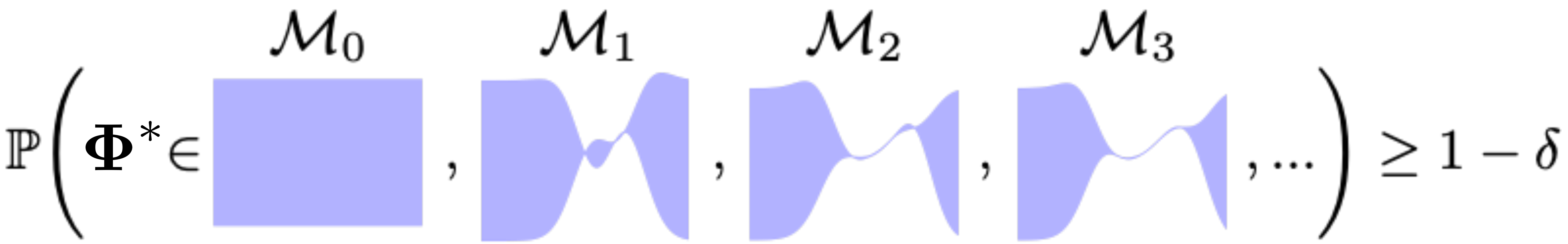


Model Based RL (non-episodic)

Iterations

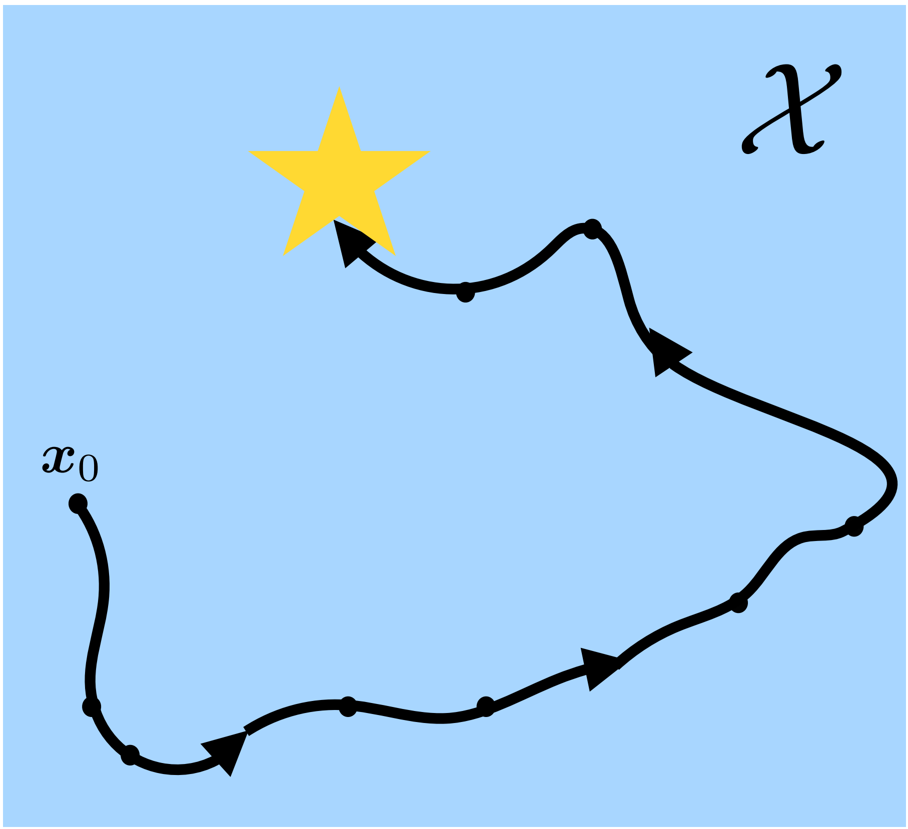
- ~~Episodes~~ $n = 1, \dots, N$.
- Execute policy $\boldsymbol{\pi}_n$ and collect measurements $(\boldsymbol{x}_{n,0}, b_{n,0}), \dots, (\boldsymbol{x}_{n,k_n}, b_{n,k_n})$
- Prepare dataset $\mathcal{D}_n = \{(\boldsymbol{z}_{n,1}, \boldsymbol{y}_{n,1}), \dots, (\boldsymbol{z}_{n,k_n}, \boldsymbol{y}_{n,k_n})\}$, where $\boldsymbol{z}_{n,i} = (\boldsymbol{x}_{n,i-1}, \boldsymbol{\pi}_n(\boldsymbol{x}_{n,i-1}))$ and $\boldsymbol{y}_{n,i} = (\boldsymbol{x}_{n,i}, b_{n,i})$.

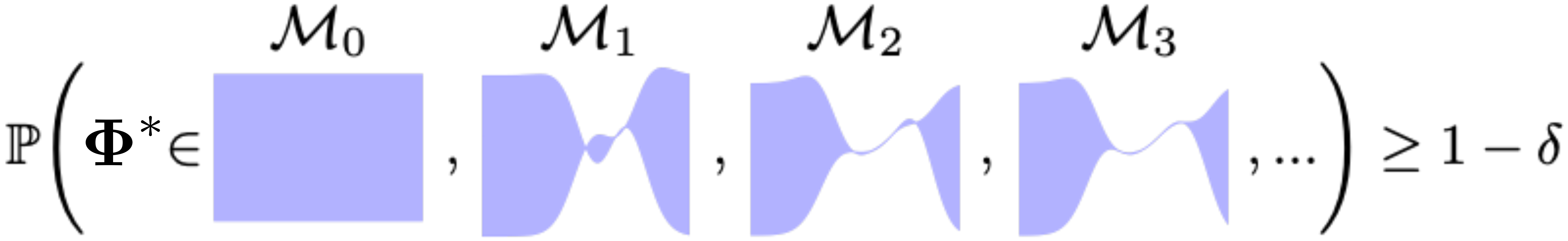




Model Based RL (non-episodic)

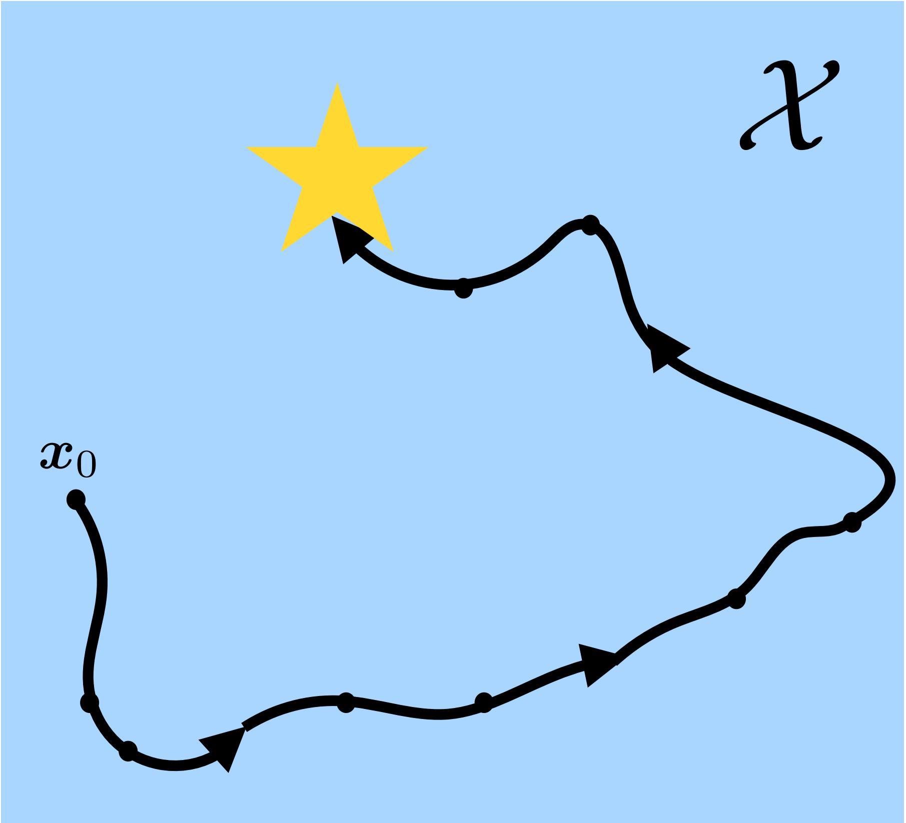
- ~~Episodes~~ Iterations $n = 1, \ldots, N$.
- Execute policy $\boldsymbol{\pi}_n$ and collect measurements $(\boldsymbol{x}_{n,0}, b_{n,0}), \ldots, (\boldsymbol{x}_{n,k_n}, b_{n,k_n})$
- Prepare dataset $\mathcal{D}_n = \{(\boldsymbol{z}_{n,1}, \boldsymbol{y}_{n,1}), \ldots, (\boldsymbol{z}_{n,k_n}, \boldsymbol{y}_{n,k_n})\}$, where $\boldsymbol{z}_{n,i} = (\boldsymbol{x}_{n,i-1}, \boldsymbol{\pi}_n(\boldsymbol{x}_{n,i-1}))$ and $\boldsymbol{y}_{n,i} = (\boldsymbol{x}_{n,i}, b_{n,i})$.

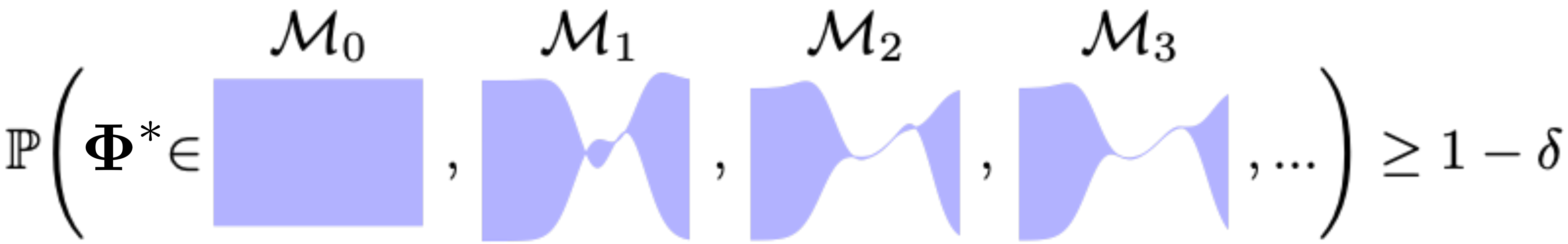




Model Based RL (non-episodic)

- ~~Episodes~~ Iterations $n = 1, \ldots, N$.
- Execute policy $\boldsymbol{\pi}_n$ and collect measurements $(\boldsymbol{x}_{n,0}, b_{n,0}), \ldots, (\boldsymbol{x}_{n,k_n}, b_{n,k_n})$
- Prepare dataset $\mathcal{D}_n = \{(\boldsymbol{z}_{n,1}, \boldsymbol{y}_{n,1}), \ldots, (\boldsymbol{z}_{n,k_n}, \boldsymbol{y}_{n,k_n})\}$, where $\boldsymbol{z}_{n,i} = (\boldsymbol{x}_{n,i-1}, \boldsymbol{\pi}_n(\boldsymbol{x}_{n,i-1}))$ and $\boldsymbol{y}_{n,i} = (\boldsymbol{x}_{n,i}, b_{n,i})$.





NeORL

NeORL

Policy Objective

$$A(\boldsymbol{\pi}^*, \boldsymbol{x}_0) = \min_{\boldsymbol{\pi}\in\Pi} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1} c(\boldsymbol{x}_t, \boldsymbol{u}_t)\right]$$

NeORL

Policy Objective

$$A(\boldsymbol{\pi}^*, \boldsymbol{x}_0) = \min_{\boldsymbol{\pi}\in\Pi} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1} c(\boldsymbol{x}_t, \boldsymbol{u}_t)\right]$$

Agent Objective

$$R_T = \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{u}_t | \boldsymbol{x}_0}[c(\boldsymbol{x}_t, \boldsymbol{u}_t) - A(\boldsymbol{\pi}^*, \boldsymbol{x}_0)]$$

NeORL

Policy Objective

$$A(\boldsymbol{\pi}^*, \boldsymbol{x}_0) = \min_{\boldsymbol{\pi} \in \Pi} \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{t=0}^{T-1} c(\boldsymbol{x}_t, \boldsymbol{u}_t) \right]$$

Agent Objective

$$R_T = \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{u}_t | \boldsymbol{x}_0} [c(\boldsymbol{x}_t, \boldsymbol{u}_t) - A(\boldsymbol{\pi}^*, \boldsymbol{x}_0)]$$

$$\boldsymbol{\pi}_n = \operatorname*{argmin}_{\boldsymbol{\pi} \in \Pi} \min_{\boldsymbol{\Phi} \in \mathcal{M}_{n-1} \cap \mathcal{M}_0} A(\boldsymbol{\pi}, \boldsymbol{\Phi})$$

NeORL

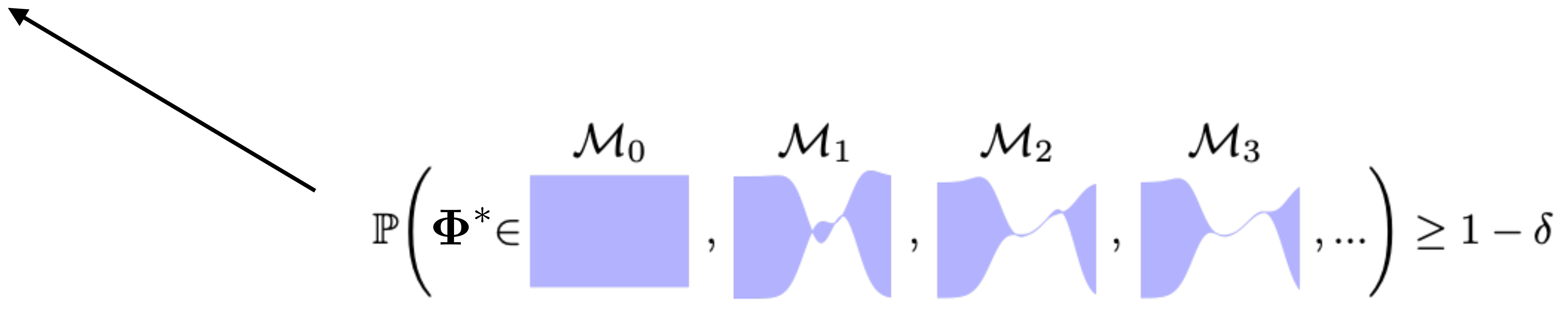
Policy Objective

$$A(\boldsymbol{\pi}^*, \boldsymbol{x}_0) = \min_{\boldsymbol{\pi}\in\Pi} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1} c(\boldsymbol{x}_t, \boldsymbol{u}_t)\right]$$

Agent Objective

$$R_T = \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{x}_t, \boldsymbol{u}_t|\boldsymbol{x}_0}[c(\boldsymbol{x}_t, \boldsymbol{u}_t) - A(\boldsymbol{\pi}^*, \boldsymbol{x}_0)]$$

$$\boldsymbol{\pi}_n = \operatorname*{argmin}_{\boldsymbol{\pi}\in\Pi} \min_{\boldsymbol{\Phi}\in\mathcal{M}_{n-1}\cap\mathcal{M}_0} A(\boldsymbol{\pi}, \boldsymbol{\Phi})$$



NeORL

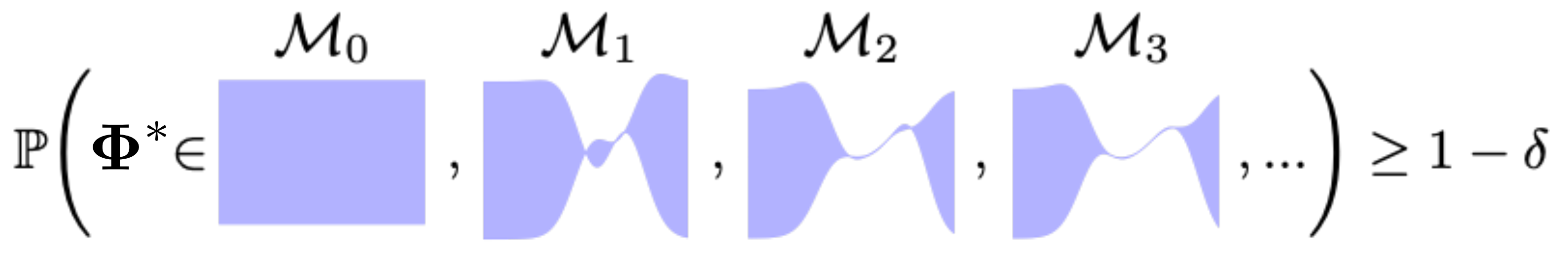
Policy Objective

$$A(\boldsymbol{\pi}^*, \boldsymbol{x}_0) = \min_{\boldsymbol{\pi}\in\Pi} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T-1} c(\boldsymbol{x}_t, \boldsymbol{u}_t)\right]$$

Agent Objective

$$R_T = \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{x}_t,\boldsymbol{u}_t|\boldsymbol{x}_0}[c(\boldsymbol{x}_t, \boldsymbol{u}_t) - A(\boldsymbol{\pi}^*, \boldsymbol{x}_0)]$$

$$\boldsymbol{\pi}_n = \operatorname*{argmin}_{\boldsymbol{\pi}\in\Pi} \min_{\boldsymbol{\Phi}\in\mathcal{M}_{n-1}\cap\mathcal{M}_0} A(\boldsymbol{\pi}, \boldsymbol{\Phi})$$



Theorem (Informal)

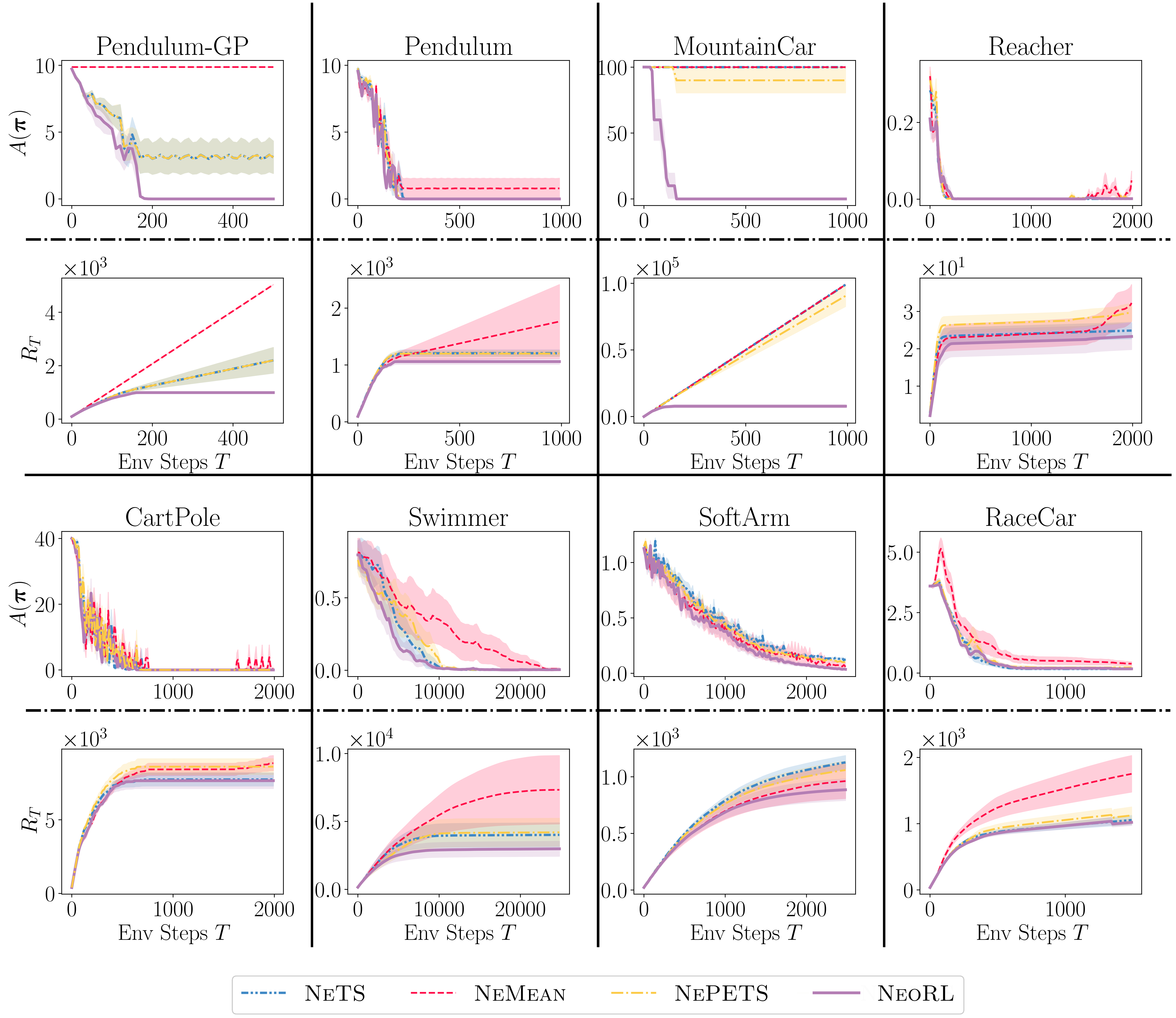
Under the assumptions, we have for NeORL *with probability at least* $1-\delta$

$$R_T = \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{x}_t,\boldsymbol{u}_t|\boldsymbol{x}_0}[c(\boldsymbol{x}_t, \boldsymbol{u}_t) - A(\boldsymbol{\pi}^*, \boldsymbol{x}_0)] \in \mathcal{O}(\beta_T\sqrt{T\Gamma_T})$$

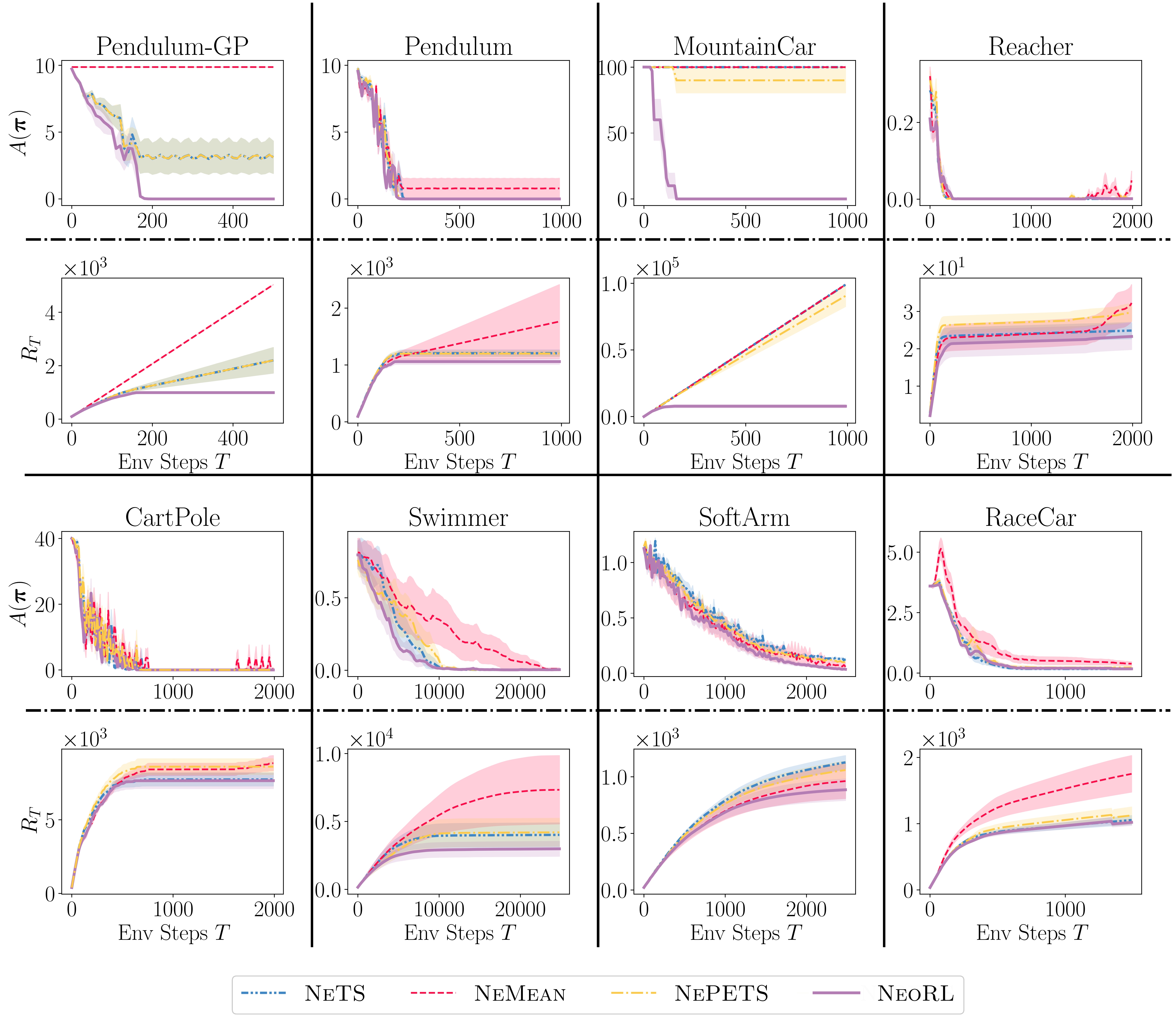
with Γ_T *being the* maximum information gain *of kernel* k, *defined as*

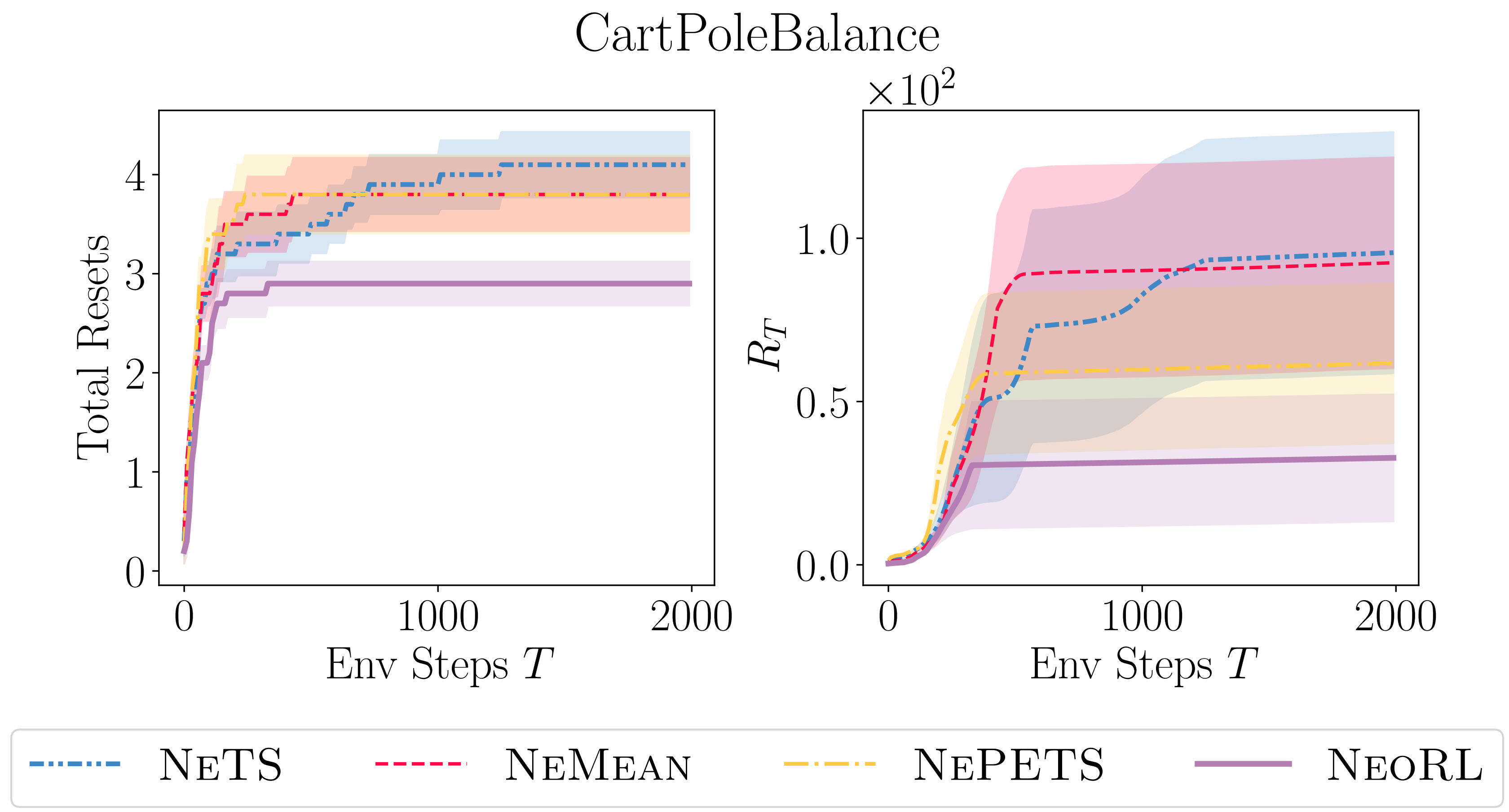
$$\Gamma_T(k) = \max_{\mathcal{A}\subset\mathcal{X}\times\mathcal{U};|\mathcal{A}|\leq T} \frac{1}{2}\log\left|\boldsymbol{I} + \sigma^{-2}\boldsymbol{K}_{\mathcal{A}}\right|.$$

NeORL — Results



NeORL — Results





Thanks for you attention!!

AUTOMATIC CONTROL LABORATORY IFA

LASA | Learning & Adaptive Systems

CRL Computational Robotics Lab