Nonparametric Policy Improvement in Continuous Action Spaces via Expert Demonstrations

Enrique Mallada

Johns Hopkins University

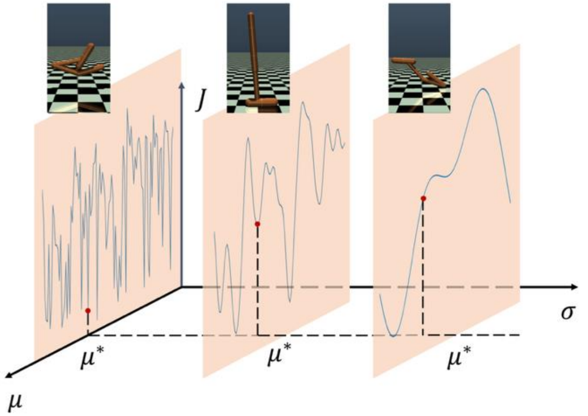
Informs

October 26th, 2025

Challenges of "modern" Policy Optimization (P.O.)

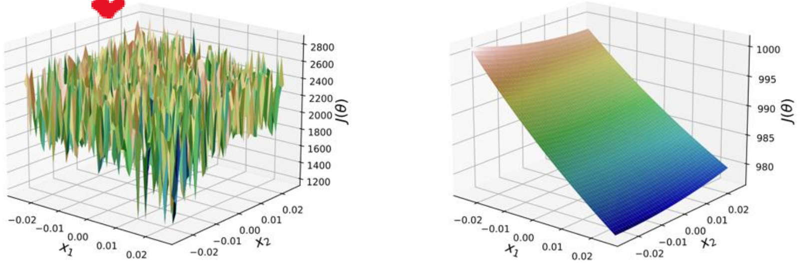
P.O. in **continuous spaces:**

- Largely based on **policy gradients.**
- Choice of **parametrization:** $a \sim \mathcal{N}\left(\mu_\theta(s), \sigma_\theta^2(s)\right)$
  - Limits expressivity.
  - Local improvement.
  - May yield non-smooth landscapes.



"**Mollification**" effects [1]

Fractal!



Non-smooth landscapes [2]

P.O. in **finite spaces** was great!

- Policy Iteration = Policy eval. + Policy improvement.

  $\pi \overset{\text{eval.}}{\rightarrow} Q^\pi(s,a)$ $\quad\quad$ $\pi'(s) \in \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q^\pi(s,a)$

- Monotonic improvement **everywhere** $V^{\pi'}(s) \geq V^\pi(s) \quad \forall s \in \mathcal{S}$.

[1] Tao Wang, Sylvia Hebert, Sicun Gao, Mollification effects of policy gradient, ICML 24
[2] — , Fractal landscapes in policy optimization, NeurIPS 23

Could we get:

1. **Benefits** of **policy iteration**
2. **Avoid drawbacks** of gradient methods?

Acknowledgements

Agustin Castellano

JOHNS HOPKINS UNIVERSITY

Sohrab Rezaei

JOHNS HOPKINS UNIVERSITY

Jared Markowitz

APL JOHNS HOPKINS APPLIED PHYSICS LABORATORY

Problem Setup

Goal: find optimal policy

$$\max_{\theta} J(\theta) := E_{s_0 \sim \rho, a_0 \sim \pi_\theta(s_0)}\Big[Q^{\pi_\theta}(s_0, a_0)\Big]$$

Problem Setup

Goal: find optimal **nonparametric** policy

$$\max_{\mathcal{D}} J(\pi_{\mathcal{D}}) := E_{s_0 \sim \rho, a_0 \sim \pi_{\mathcal{D}}(s_0)} \left[Q^{\pi_{\mathcal{D}}}(s_0, a_0) \right]$$

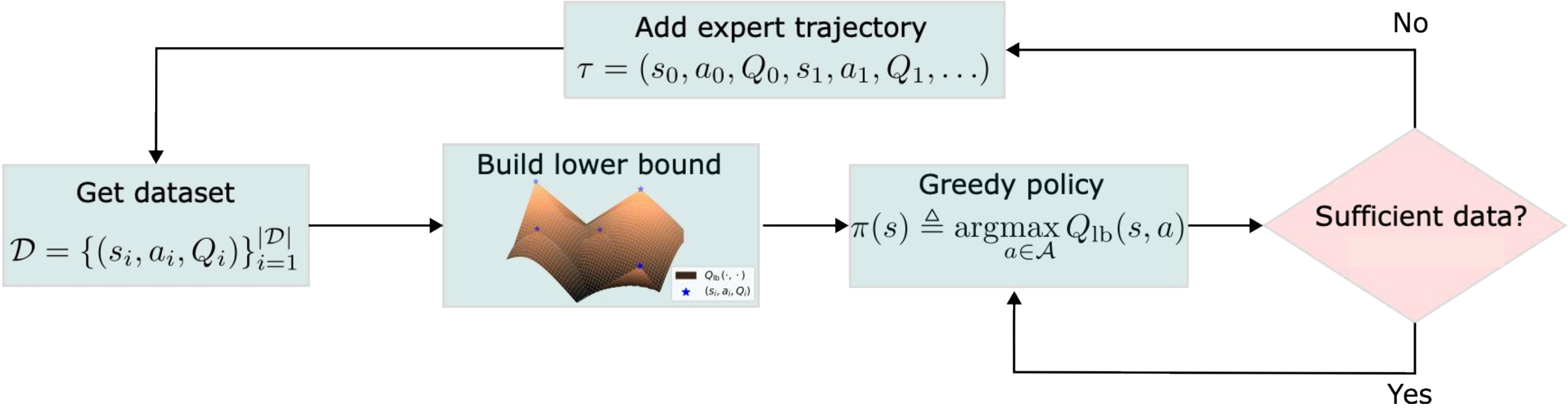
Data set: $\mathcal{D} = \{(s_i, a_i, Q_i)\}_{i=1}^{|\mathcal{D}|}$ $\quad Q_i := \sum_t \gamma^t r(s_t, a_t)$

Assumptions: distance

- C ... $, a'))$
- D
- **Expert data:** ... $= Q^{\star}(s_i, a_i)$

1. **How** can we leverage these transitions to learn a policy?
2. **What** guarantees can we get when we add more transitions?
3. **Where** should we add transitions to improve performance?

Overview of our method



1. How can we use these transitions to learn a nonparametric policy?

Building bounds & Nonparametric Policy

Expert data: we have $\mathcal{D} = \{(s_i, a_i, Q_i)\}_{i=1}^{|\mathcal{D}|}$, where: $a_i = \pi^\star(s_i); Q_i = Q^\star(s_i, a_i)$

- Use data to define **lower bounds** on optimal values:

$$V_{\text{lb}}(s) \triangleq \max_{1 \le i \le |\mathcal{D}|} \{Q_i - L \cdot d_{\mathcal{S}}(s, s_i)\} \quad Q_{\text{lb}}(s, a) \triangleq \max_{1 \le i \le |\mathcal{D}|} \{Q_i - L \cdot (d_{\mathcal{S}}(s, s_i) + d_{\mathcal{A}}(a, a_i))\}$$

- **Nonparametric Policy:**

$$\pi(s) \triangleq \operatorname*{argmax}_{a \in \mathcal{A}} Q_{\text{lb}}(s, a) = a_{i'}$$

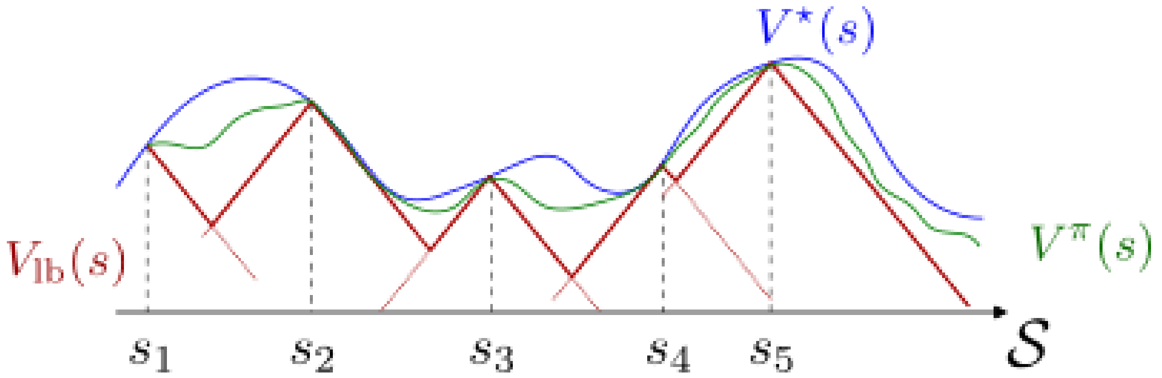
- **Remark:** Note argmax always gives actions in dataset $(s_{i'}, a_{i'}, Q_{i'})$
- **Question:** What can we say about $V^\pi(s)$?

Nonparametric policy *improves* over lower bound

Policy Evaluation*:

- Nonparametric π satisfies $\forall s \in \mathcal{S}$:

$$V_{\text{lb}}(s) \le V^{\pi}(s) \le V^{\star}(s)$$



Policy Improvement*:

- Given data sets $\mathcal{D}, \mathcal{D}'$ with $\mathcal{D} \subset \mathcal{D}'$

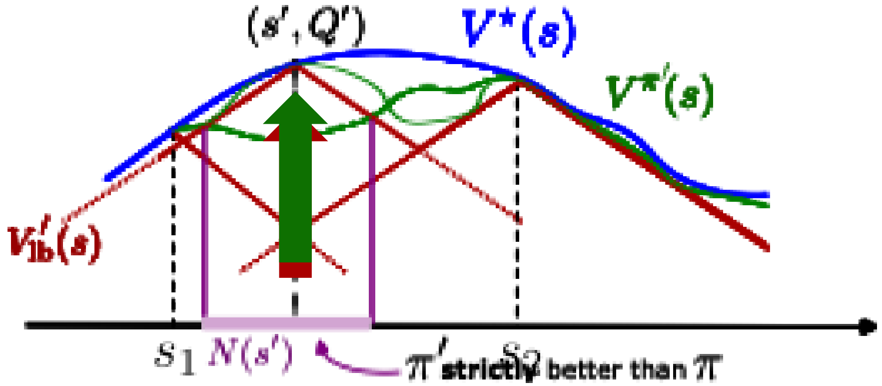
$$V_{\text{lb}}(s) \le V'_{\text{lb}}(s) \qquad \forall s \in \mathcal{S}$$

More data = better lower bounds

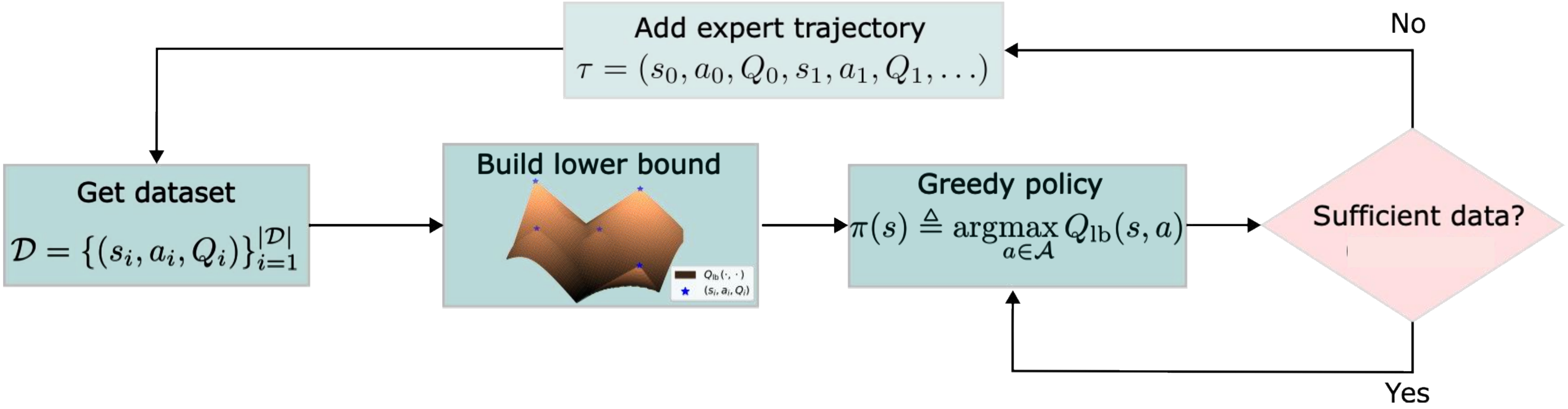
$$V^{\pi}(s') \le V^{\pi'}(s') \qquad \forall s' \in \mathcal{D}' \backslash \mathcal{D}$$

Improvement on added points

- **Strict** on neighbors of new data: $\forall s \in N(s')$



(*:Data must come from trajectories)



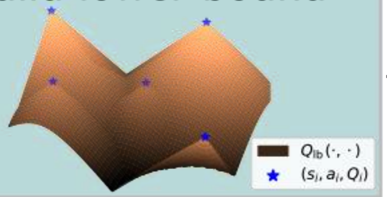
Add expert trajectory
$\tau = (s_0, a_0, Q_0, s_1, a_1, Q_1, \ldots)$
No
Get dataset
$\mathcal{D} = \{(s_i, a_i, Q_i)\}_{i=1}^{|\mathcal{D}|}$
Build lower bound
$Q_{\text{lb}}(\cdot, \cdot)$
(s_i, a_i, Q_i)
Greedy policy
$\pi(s) \triangleq \operatorname{argmax}_{a \in \mathcal{A}} Q_{\text{lb}}(s, a)$
Sufficient data?
Yes

1. **How** to learn a policy?

Get dataset

$\mathcal{D} = \{(s_i, a_i, Q_i)\}_{i=1}^{|\mathcal{D}|}$

→ Build lower bound



→ Greedy policy

$\pi(s) \triangleq \underset{a\in\mathcal{A}}{\operatorname{argmax}}\, Q_{\text{lb}}(s,a)$

→ Policy Evaluation

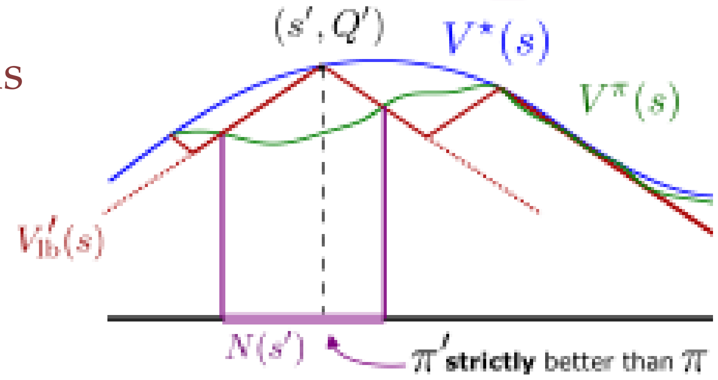
$V_{\text{lb}}(s) \leq V^{\pi}(s)$

2. **What** guarantees with more transitions?

$V_{\text{lb}}(s) \leq V'_{\text{lb}}(s) \quad \forall s \in \mathcal{S}$ — More data = better lower bounds

$V^{\pi}(s') \leq V^{\pi'}(s') \quad \forall s' \in \mathcal{D}'\backslash\mathcal{D}$ — Improvement on added points

strict improvement



3. **Where** to add transitions?

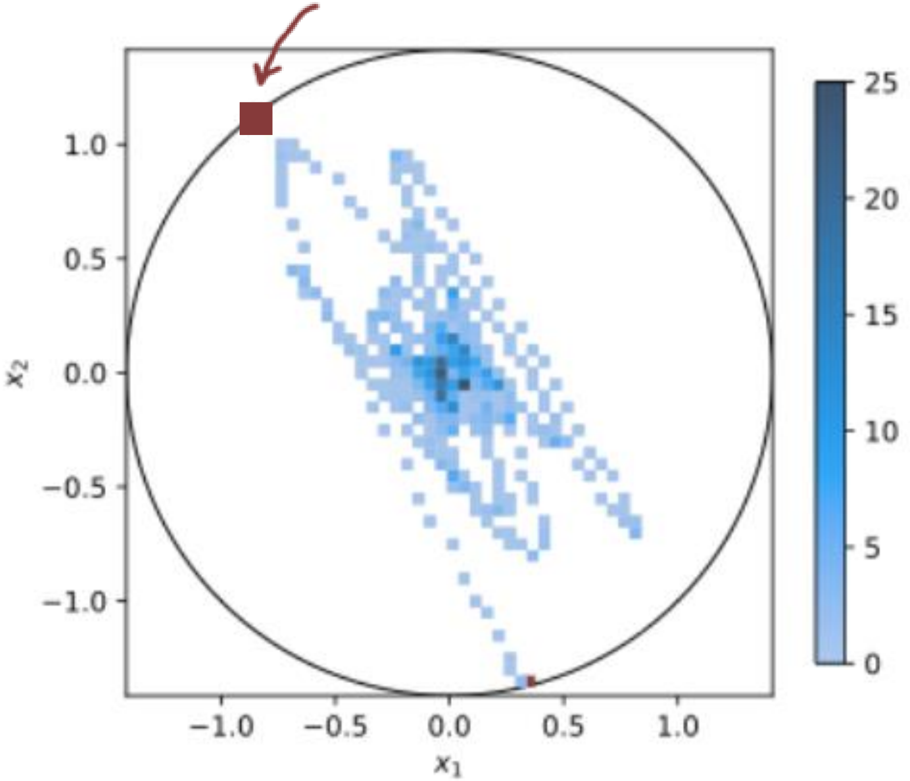
- Only *where* **sufficient improvement** is guaranteed: $\Delta(s) = V^{\star}(s) - V^{\pi}(s)$

Algorithm

Input: Lipschitz constant L

For each episode **do**:

1. Sample $s \sim \rho(\mathcal{S}_0)$
2. If $\Delta(s) > \varepsilon$:
   - Run optimal trajectory with $\pi^\star$
     $$\tau = (s_0, a_0, Q_0, s_1, a_1, Q_1, \ldots)$$
   - **Repeat:** add tuples to dataset (from $i = 0$)
     $$\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_i, a_i, Q_i)\}$$
     Until: $\Delta(s_i) \ll \varepsilon$.
3. Else:
   - **Continue**



Convergence guarantees?

$\sup_{s \in \mathcal{S}_0} |V^\star(s) - V^\pi(s)| \le \varepsilon$ w.p. $\ge 1 - \delta$ in $\mathcal{O}\left(\mathrm{N}_{\text{cover}}\left(\frac{\varepsilon}{4L}\right) \log \mathrm{N}_{\text{cover}}\left(\frac{\varepsilon}{4L}\right) \log \frac{1}{\delta}\right)$ episodes.

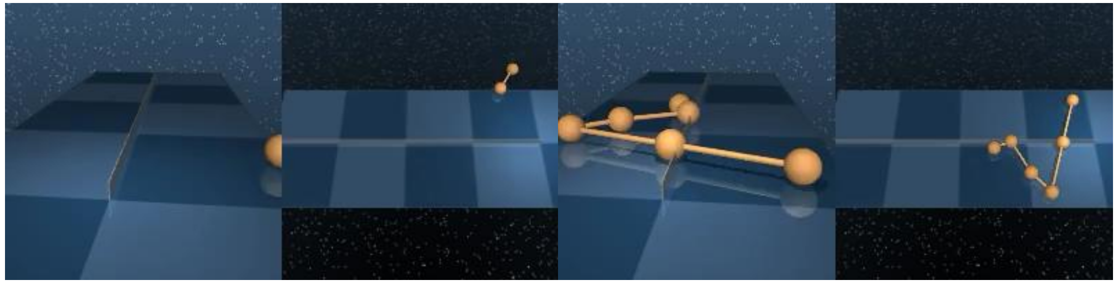
Experiments

Number of balls

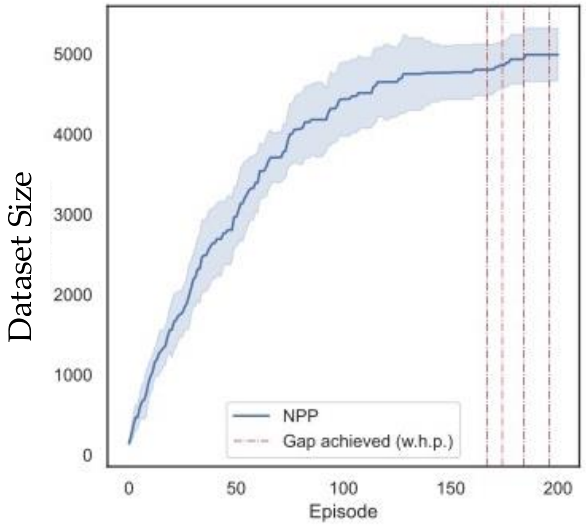
1st m actuated

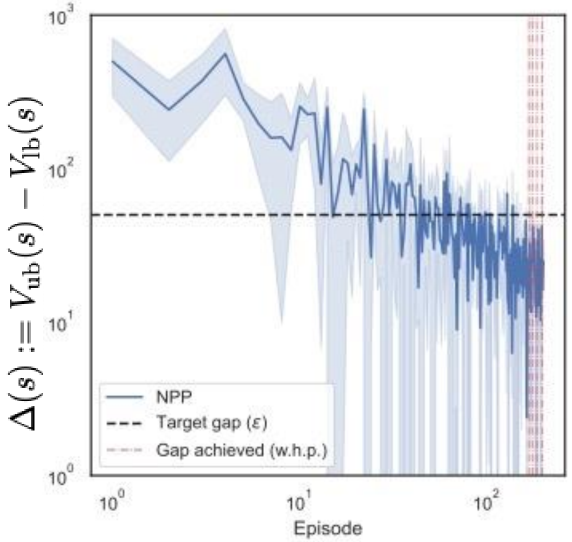
- We use the `lqr_n_m` environments from DeepMind's Control Suite

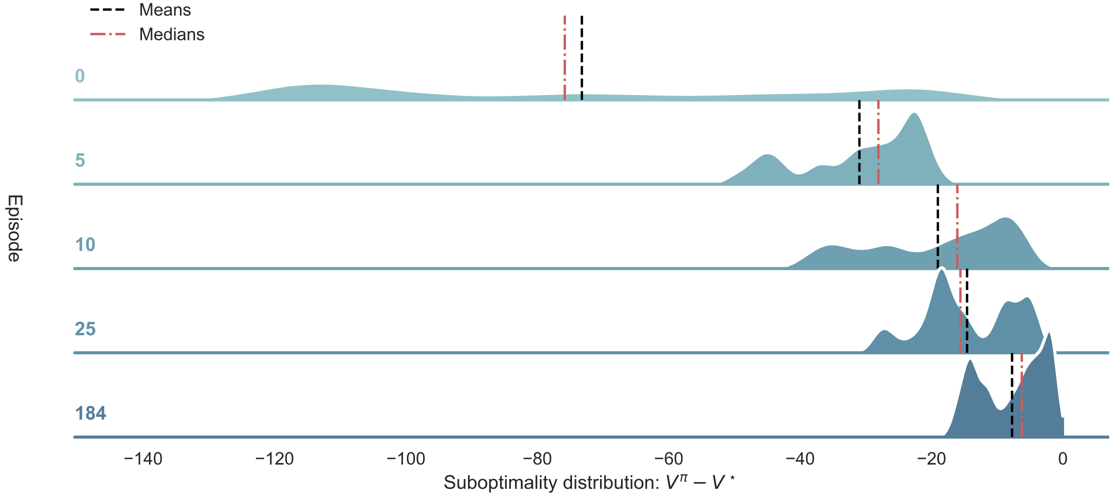
`lqr_2_1` **`lqr_6_2`**



- Results on **`lqr_2_1`**:

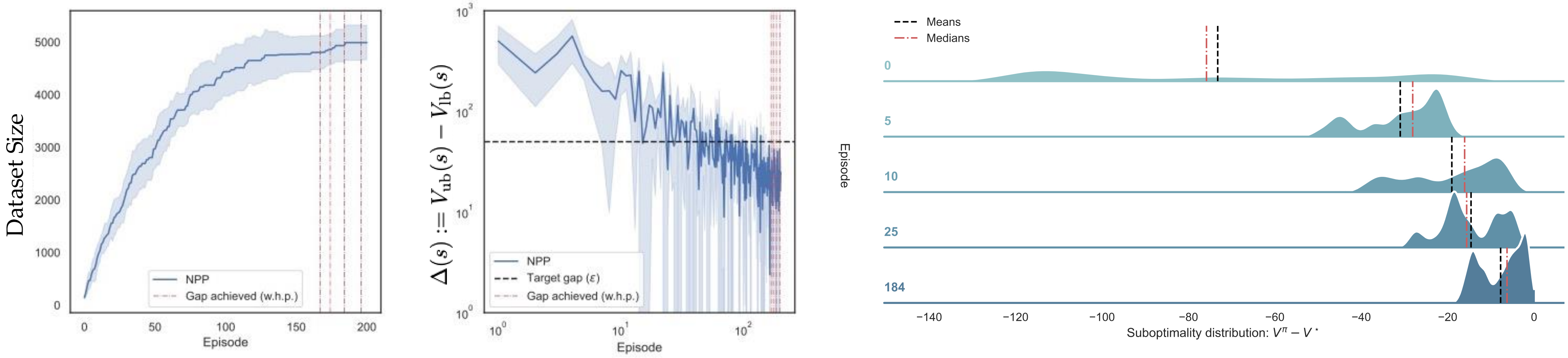






Experiments

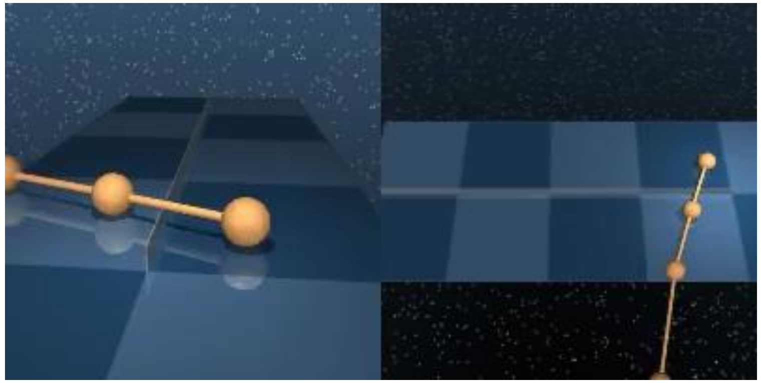
- We use the `lqr_n_m` environments from DeepMind's Control Suite
- Results on **`lqr_2_1`**:



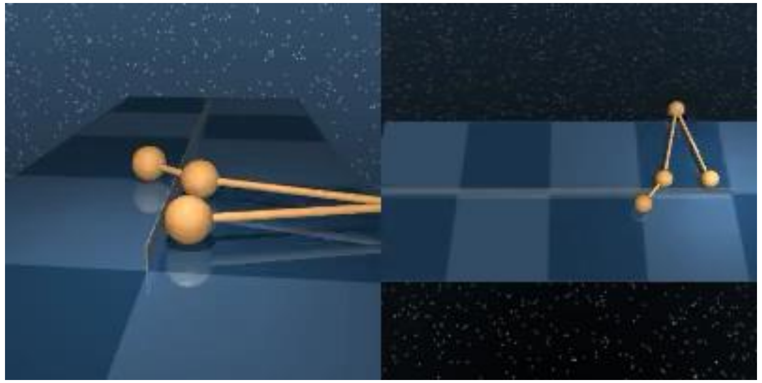
- **Remarks:**
  - **Incremental learning:** No catastrophic forgetting.
  - **Improvement** across **entire state space** (not in expectation).
  - **Only valuable data is added** (harder to find at times passes)

Incremental Learning

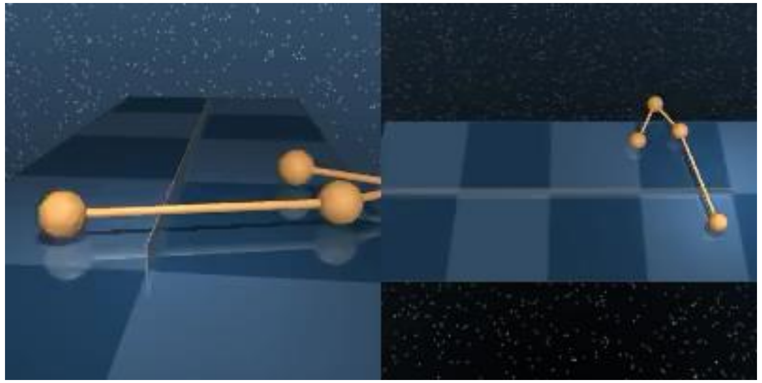
after 10 episode…



after 100 episode…

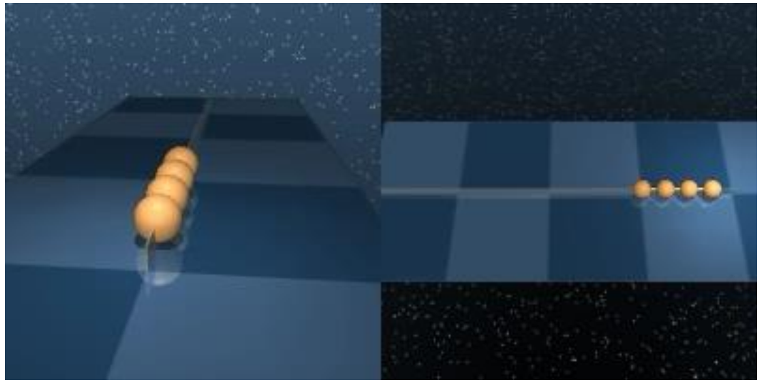


after 1000 episodes…



after 30K+

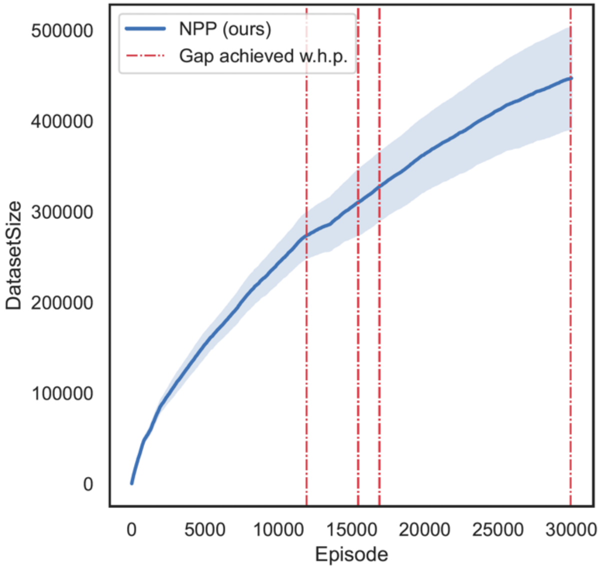
optimal control

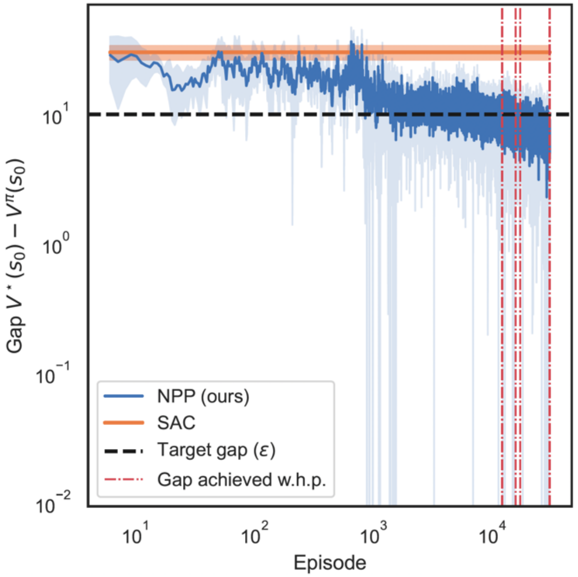


Incremental Learning

after 30K+

optimal control





Conclusions

- Proposed **nonparametric policy** based on **expert demonstrations**
- Policy is **greedy** w.r.t. **lower bound** on $\boldsymbol{Q}^{\star}$, satisfies:

  i) **policy evaluation inequality** (everywhere)

  ii) (strict) **policy improvement** (on new data)
- **Data** collection only **where it's needed.**
- Experiments show **incremental learning, no catastrophic forgetting**

Future work

- Sub-expert demonstrations.
- Bootstrapping with lower bounds.
- Stochastic MDPs.

Thank you!

Questions?

Reinforcement Learning Conference

Check out our paper

Check out our Github repo