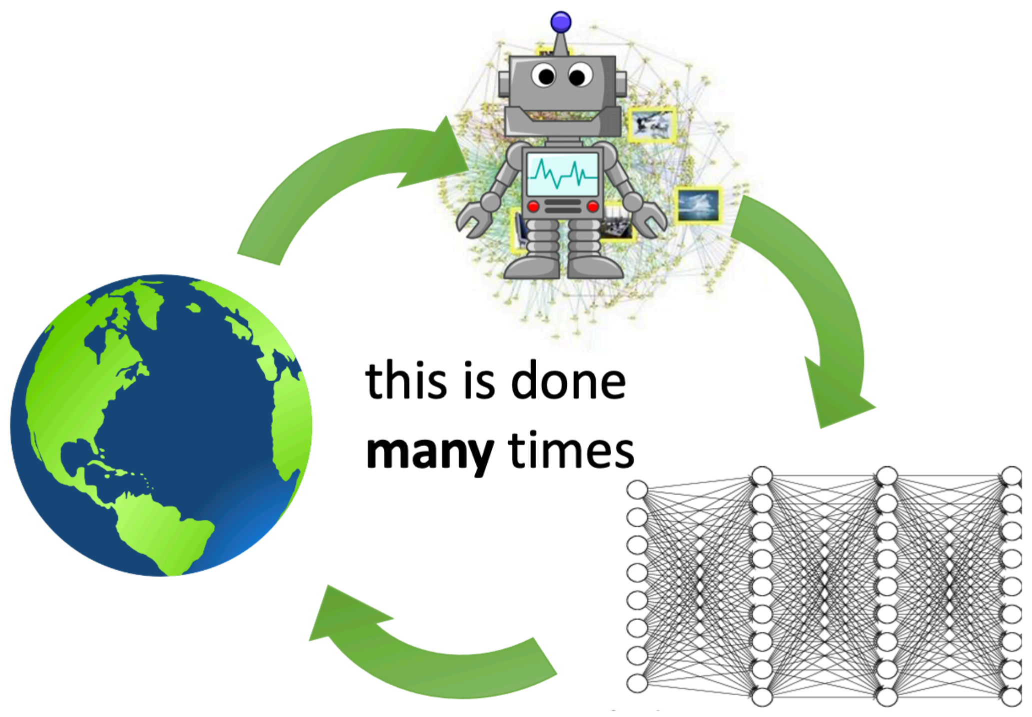
Hybrid RL breaks sample size barriers in linear MDPs

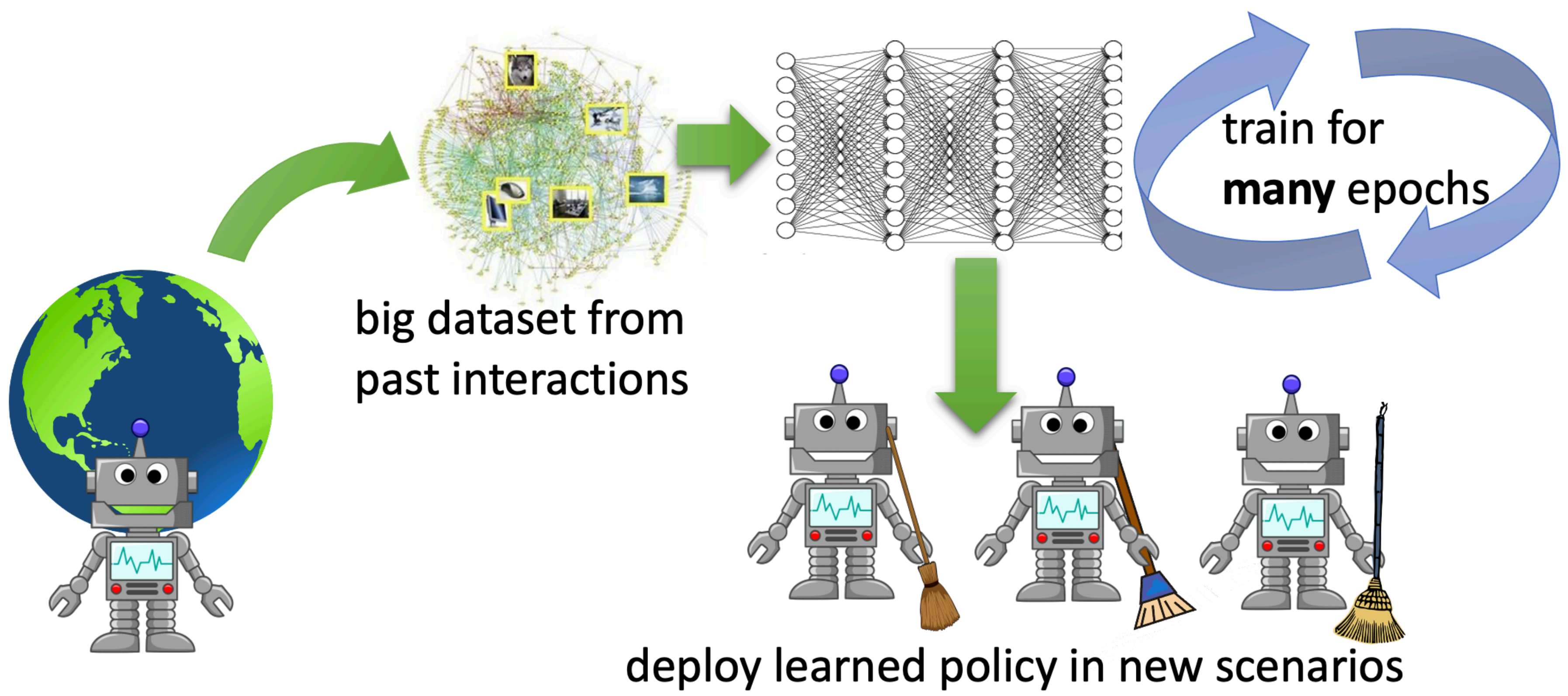
Kevin Tan, Wei Fan, Yuting Wei

Recap: Hybrid RL

reinforcement learning



offline reinforcement learning



https://bair.berkeley.edu/blog/2020/12/07/offline/

- Online RL:
  - Study of how machines learn by doing.
  - At time h, see state s_h, take action $a_h \sim \pi_h(s_h \mid \cdot\,)$ according to a policy, receive reward $r_h(s_h, a_h)$, see next state $s_{h+1} \sim \mathbb{P}(\,\cdot \mid s_h, a_h)$.
- Offline RL:
  - Study of how machines learn by watching.
  - Given a dataset $\mathscr{D} = \{(s_h^{(n)}, a_h^{(n)}, r_h^{(n)})_{n=1}^N\}_{h=1}^H$, learn a policy $\hat{\pi}_h$ whose value $V_h^{\hat{\pi}}$ is close to optimal: $V_h^* - V_h^{\hat{\pi}} \leq \epsilon$.

This work

Learning from offline and online data lets you do better than the **minimax lower bounds** in offline and online RL

Even with function approximation and without a good-quality offline dataset

Linear MDPs

Tractable function approximation

- Access to features of states and actions:
  - $\phi_h(s, a) \in \mathbb{R}^d$
- Probability transitions and reward functions are linear functions of features.
- Why is this useful?
  - Value function (how good a policy is) and Q-function (what happens if I take action a now, and follow the policy after) are linear functions of the features.
  - Can learn these via ridge regression!

Splitting the state-action space

- We consider partitions $\mathscr{X}_{\text{off}} \cup \mathscr{X}_{\text{on}} = \mathscr{X} = [H] \times \mathscr{S} \times \mathscr{A}$ of the state-action space.
- Strategy: Bound the regret/error on each partition separately.
    - Use the offline data for the offline partition, and online data for the online partition.
- Offline measure of learning complexity, $c_{\text{off}} = \max_h 1/\lambda_{d_{\text{off}}}(\mathbb{E}_{\mu_h}(P_{\text{off}}\phi_h)(P_{\text{off}}\phi_h)^T)$.
    - Inverse of the d_{off}-th largest eigenvalue of the covariance matrix of the feature maps projected down to the offline partition. By Kiefer-Wolfowitz, no worse than d.
- Online measure of learning complexity:
    - Dimension of the online partition d_{on}, no larger than d.

What we provide

Two algorithms

- Two algorithms that do better than lower bounds for offline-only and online-only RL
- Online-to-offline approach:
    - Use reward-agnostic exploration informed by the offline dataset to collect data with good coverage, then do minimax-optimal offline RL on combined dataset.
    - Better guarantee than minimax-optimal offline RL alone!
- Offline-to-online approach:
    - Include the offline dataset in the experience replay buffer of a minimax-optimal regret-minimizing online RL algorithm.
    - Better guarantee than minimax-optimal online RL alone!

Online-to-offline approach

The algorithm

Algorithm 1 Reward-Agnostic Pessimistic PAC Exploration-initialized Learning (RAPPEL)

1: **Input:** Offline dataset $\mathcal{D}_{\text{off}}$, samples sizes N_{on}, N_{off}, feature maps ϕ_h, tolerance parameter for reward-agnostic exploration τ.

2: **Initialize:** $\mathcal{D}_h^{(0)} \leftarrow \varnothing \ \ \forall h \in [H]$, $\lambda = 1/H^2$, $\beta_2 = \tilde{O}(\sqrt{d})$.

3: **for** horizon $h = 1, \ldots, H$ **do**

4: Run an exploration algorithm (OPTCOV, Wagenmaker and Jamieson (2023)) to collect covariates $\mathbf{\Lambda}_h$ such that

$$\max_{\phi_h \in \Phi} \phi_h^\top (\mathbf{\Lambda}_h + \lambda \mathbf{I} + \mathbf{\Lambda}_{\text{off},h})^{-1} \phi_h \leqslant \tau.$$

5: **end for** **Perform reward-agnostic exploration (informed by offline data) to collect data with good coverage**

6: **Output:** $\widehat{\pi}$ from running a pessimistic offline RL algorithm (LinPEVI-ADV+, Xiong et al. (2023)) with hyperparameters λ, β_2 on the combined dataset $\mathcal{D}_{\text{off}} \cup \{\mathcal{D}_h^{(N_{\text{on}})}\}_{h \in [H]}$.

Then use offline RL to learn a policy from the combined offline+online dataset

Online-to-offline approach

The guarantee

- For any partitions $\mathscr{X}_{\text{off}} \cup \mathscr{X}_{\text{on}} = \mathscr{X} = [H] \times \mathscr{S} \times \mathscr{A}$ of the state-action space.
- If you run the exploration algorithm for enough iterations (burn-in cost), we get w.h.p:
  - $V_1^*(s) - V_1^{\hat{\pi}}(s) \lesssim \sqrt{d} \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \|\phi(s_h, a_h)\|_{(\boldsymbol{\Sigma}^*_{\text{off},h} + \boldsymbol{\Sigma}^*_{\text{on},h})^{-1}}.$
    - Better than the minimax-optimal offline RL rate $\sqrt{d} \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \|\phi(s_h, a_h)\|_{\boldsymbol{\Sigma}^{*-1}_{\text{off},h}}$!
  - $V_1^*(s) - V_1^{\hat{\pi}}(s) \lesssim \sqrt{c_{\text{off}}\, dH^3 \min\left\{c_{\text{off}}, H\right\}/N_{\text{off}}} + \sqrt{d_{\text{on}}\, dH^3 \min\left\{d_{\text{on}}, H\right\}/N_{\text{on}}}$
    - Better than the upper bound of $\sqrt{d^2H^4/N_{\text{off}}}$ from minimax-optimal offline RL!

Offline-to-online approach

The algorithm

Algorithm 2 Hybrid Regression for Upper-Confidence Reinforcement Learning (HYRULE)

Input: Offline dataset $\mathcal{D}_{\text{off}}$, samples sizes N_{on}, N_{off}, feature maps ϕ_h. Regularization parameter $\lambda > 0$, confidence radii $\beta, \bar{\beta}, \tilde{\beta}$, $t_{\text{last}} = 0$. **Estimate parameters from offline data**

Initialize: For $h \in [H]$, estimate $\widehat{\mathbf{w}}_{1,h}, \check{\mathbf{w}}_{1,h}, Q_{1,h}, \check{Q}_{1,h}, \sigma_{1,h}, \bar{\sigma}_{1,h}$ from $\mathcal{D}_{\text{off}}$, and assign $\boldsymbol{\Sigma}_{0,h} = \boldsymbol{\Sigma}_{1,h} = \boldsymbol{\Sigma}_{\text{off}} + \lambda \mathbf{I} = \sum_{n=1}^{N_{\text{off}}} \bar{\sigma}_{n,h}^{-2} \phi_{n,h} \phi_{n,h}^{\top} + \lambda \mathbf{I}$.

for episodes $t = 1, \ldots, T$ **do**

  Update optimistic and pessimistic weights $\widehat{\mathbf{w}}_{t,h}, \check{\mathbf{w}}_{t,h}$ for all h. **Do variance-aware regret minimization**

  if there exists a stage $h' \in [H]$ such that $\det(\boldsymbol{\Sigma}_{t,h'}) \geqslant 2 \det(\boldsymbol{\Sigma}_{t_{\text{last}},h'})$ **then**

    Update optimistic and pessimistic Q-functions $Q_{t,h}(s,a), \check{Q}_{t,h}(s,a)$, set $t_{\text{last}} = t$.

  end if

  for horizon $h = 1, \ldots, H$ **do**

    Play action $a_h^{(t)} \leftarrow \arg\max_a Q_{t,h}(s_h^{(t)}, a)$, receive reward $r_h^{(t)}$, next state $s_{h+1}^{(t)}$

    Estimate $\sigma_{t,h}, \bar{\sigma}_{t,h} \leftarrow \max\{\sigma_{t,h}, \sqrt{H}, 2d^3 H^2 \|\boldsymbol{\phi}(s_h^{(t)}, a_h^{(t)})\|_{\boldsymbol{\Sigma}_{t,h}^{-1}}^{1/2}\}$[1], update $\boldsymbol{\Sigma}_{t+1,h}$.

  end for

end for

Output: Greedy policy $\widehat{\pi} = \pi^{Q_{T,h}}$, $\text{Unif}(\pi^{Q_{1,h}}, \ldots, \pi^{Q_{T,h}})$ for PAC guarantee.

Offline-to-online approach

The guarantee

- For any partitions $\mathcal{X}_{\text{off}} \cup \mathcal{X}_{\text{on}} = \mathcal{X} = [H] \times \mathcal{S} \times \mathcal{A}$ of the state-action space.
- After a burn-in cost we get w.h.p:
- $\mathrm{Reg}\,(T) \lesssim \inf_{\mathcal{X}_{\text{on}}, \mathcal{X}_{\text{off}}} \left(\sqrt{c_{\text{off}}^2 H^3 T^2 / N_{\text{off}}} + \sqrt{d_{\text{on}} d H^3 T} \right).$
  - Better than the minimax-optimal online RL rate $\sqrt{d^2 H^3 T}$!
- $V_1^*(s) - V_1^{\hat{\pi}}(s) \lesssim \inf_{\mathcal{X}_{\text{on}}, \mathcal{X}_{\text{off}}} \left(\sqrt{c_{\text{off}}^2 H^3 / N_{\text{off}}} + \sqrt{d_{\text{on}} d H^3 / T} \right).$
  - Guarantee for error of learned policy via an online-to-batch conversion.

How was this done?

- Dimensional dependence sharpened from d to d_{on} and c_{off}.
  - Via projections onto online and offline partitions.
- H^3 dependence achieved by combining law of total variance and a novel truncation argument.
  - Average variance lower than worst-case variance, and by truncating we can "ignore" the worst-case variance on average.

Comparison with other work out there

	Upper Bound	Lower Bound
Offline (Error)	$\sqrt{d} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left\Vert \phi(s_h, a_h) \right\Vert_{\Sigma_{\mathrm{off},h}^{*-1}}$ $\boxed{\lesssim \sqrt{C^* d^2 H^4 / N_{\mathrm{off}}}}$ (Xiong et al., 2023)	$\sqrt{d} \cdot \sum_{h=1}^{H} \mathbb{E}_{\pi^*} \left\Vert \phi(s_h, a_h) \right\Vert_{\Sigma_{\mathrm{off},h}^{*-1}}$ $\boxed{\gtrsim \sqrt{C^* d^2 H^2 / N_{\mathrm{off}}}}$ (Xiong et al., 2023)
Online (Regret)	$\boxed{\sqrt{d^2 H^3 T}}$ (He et al., 2023)	$\boxed{\sqrt{d^2 H^3 T}}$ (Zhou et al., 2021)

	Result
Hybrid (Online-to-offline Error)	$\sqrt{d^2 H^7 / N}$ (Wagenmaker and Pacchiano, 2023) $\boxed{\sqrt{c_{\mathrm{off}}(\mathcal{X}_{\mathrm{off}}) d H^3 \min\{c_{\mathrm{off}}(\mathcal{X}_{\mathrm{off}}), H\} / N_{\mathrm{off}}} + \sqrt{d_{\mathrm{on}} d H^3 \min\{d_{\mathrm{on}}, H\} / N_{\mathrm{on}}}}$ (Alg. 1)
Hybrid (Offline-to-online Regret)	$C^* \sqrt{d^2 H^6 N_{\mathrm{on}}}$ (Song et al., 2023; Nakamoto et al., 2023) $\sqrt{(C^* + c_{\mathrm{on}}(\mathcal{X})) d^3 H^6 N_{\mathrm{on}}}$ (Amortila et al., 2024) $\sqrt{c_{\mathrm{off}}(\mathcal{X}_{\mathrm{off}}) d H^5 N_{\mathrm{on}}^2 / N_{\mathrm{off}}} + \sqrt{d_{\mathrm{on}} d H^5 N_{\mathrm{on}}}$ (Tan and Xu, 2024) $\boxed{\sqrt{c_{\mathrm{off}}(\mathcal{X}_{\mathrm{off}})^2 d H^3 N_{\mathrm{on}}^2 / N_{\mathrm{off}}} + \sqrt{d_{\mathrm{on}} d H^3 N_{\mathrm{on}}}}$ (Alg. 2)

Table 2: Comparisons of our results to the best upper and lower bounds available, and existing results for hybrid RL, in linear MDPs. The inequalities in the offline row hold when the behavior policy satisfies C^*-single policy concentrability. Often, offline data is cheaper or easier to obtain. When this happens, $N_{\mathrm{off}} \gg N_{\mathrm{on}}$, and the second term (depending on $N_{\mathrm{on}} = T$) dominates.

Performance of informed exploration

Reward-agnostic exploration more effective with offline data on Tetris

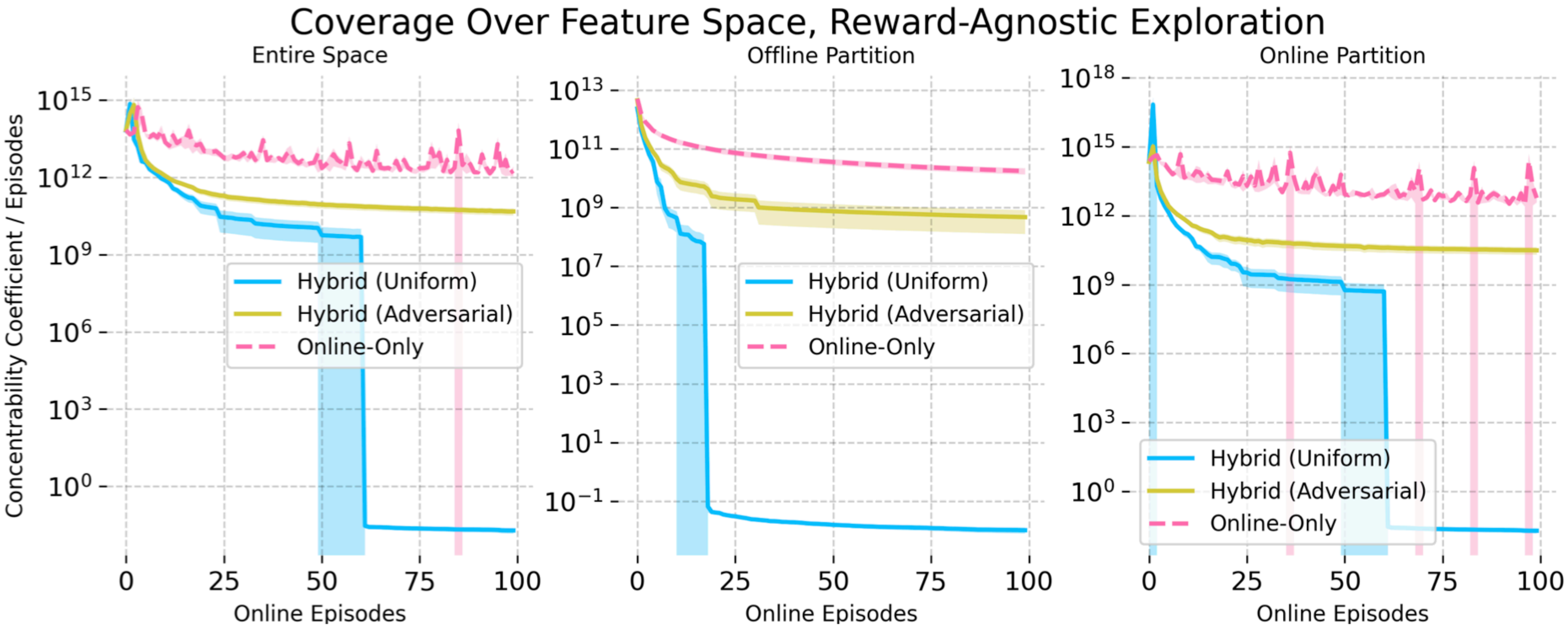


Figure 1: Coverage achieved by OPTCOV with 200 trajectories of offline data collected under a uniform and an adversarial behavior policy, and with no offline data. Results averaged over 30 trials, with the shaded area depicting 1.96-standard errors. Lower is better.

Performance of online-to-offline approach

Hybrid RL helps with learning from adversarial behavior policies on Tetris

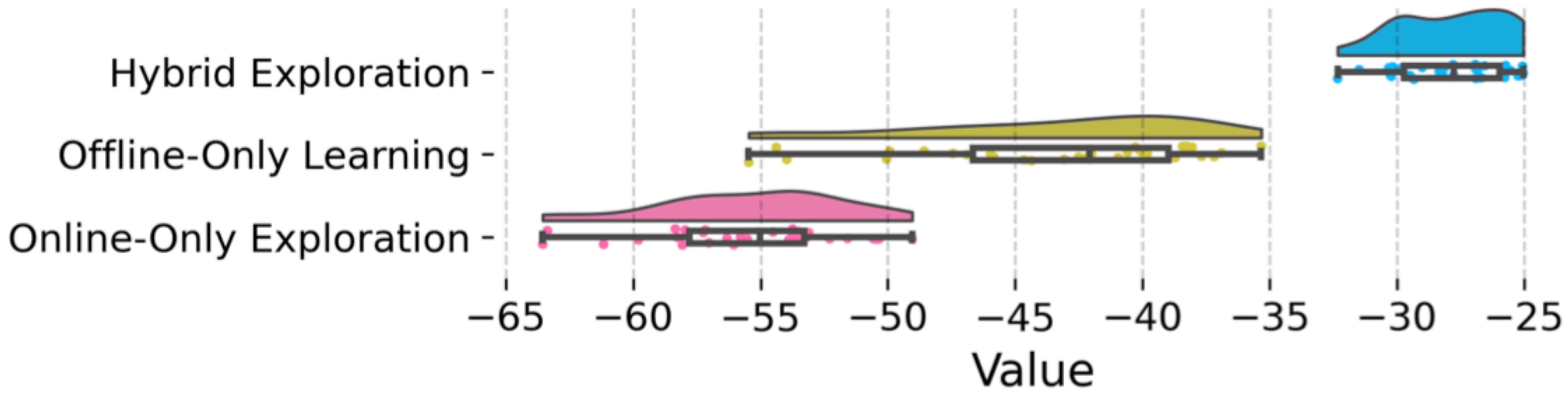


Figure 2: Value of policies learned by applying LinPEVI-ADV to the hybrid, offline, and online datasets, with an adversarial behavior policy. The reward is negative as it is the negative of the excess height. Results over 30 trials. Higher is better.

Performance of offline-to-online approach

Variance-aware regret-minimizing hybrid RL outperforms minimax-optimal online-only learning

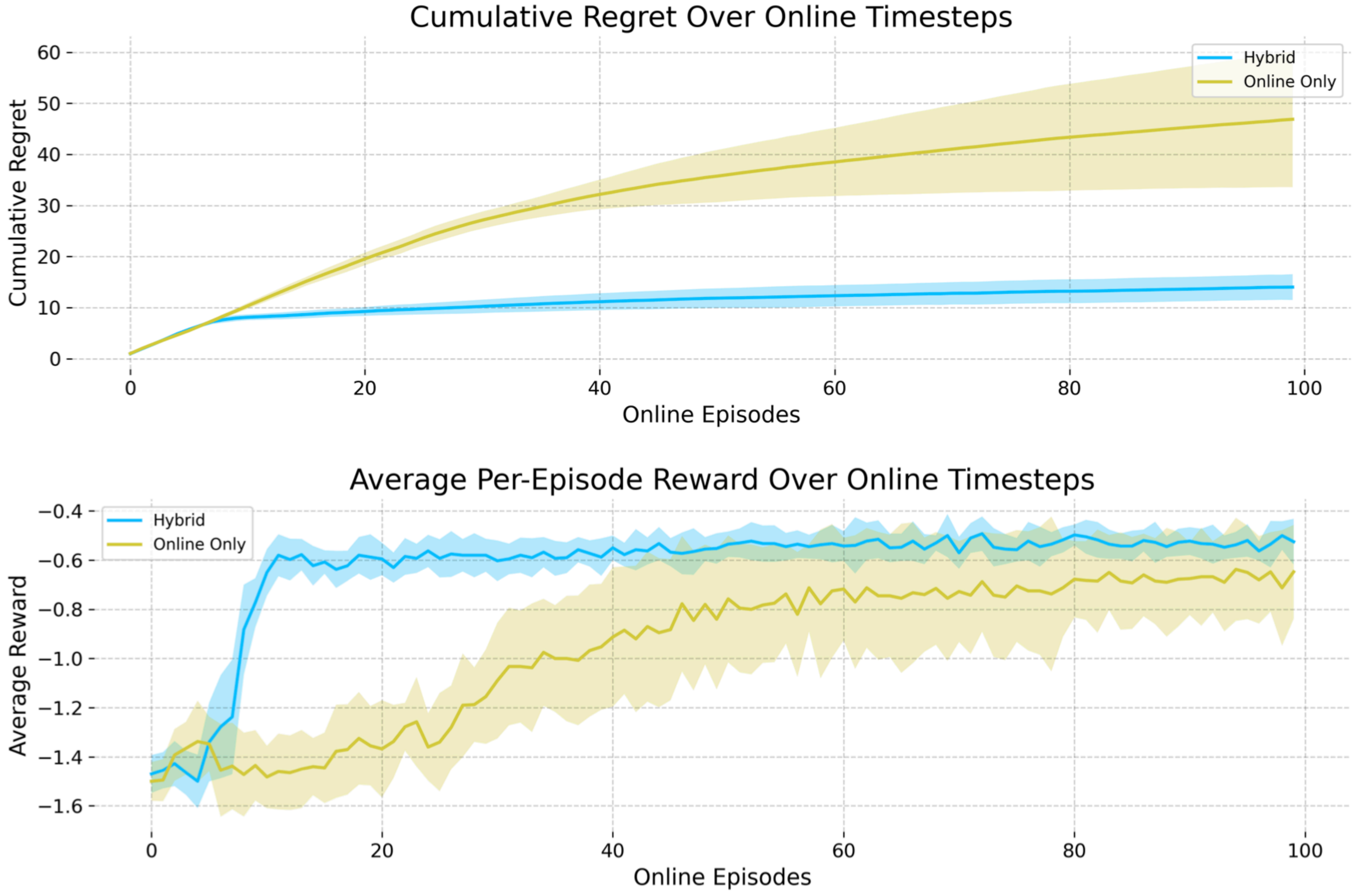


Figure 3: Comparison of LSVI-UCB++ and Algorithm 2. Results averaged over 10 trials, with 1-standard deviation error bars over 10 trials.

Bottom line and further questions

Sharpest guarantees for hybrid RL in linear MDPs thus far

- We improve over online-only or offline-only RL, but not both at the same time.
- H^3 dependence in offline RL is new, but with caveats on d dependence.
- High burn-in costs for both algorithms.
- Which is better rate-wise? Offline-to-online or online-to-offline? No clear answer here.
- Further work on other function approximation while remaining statistically efficient needed, linear only a first step.