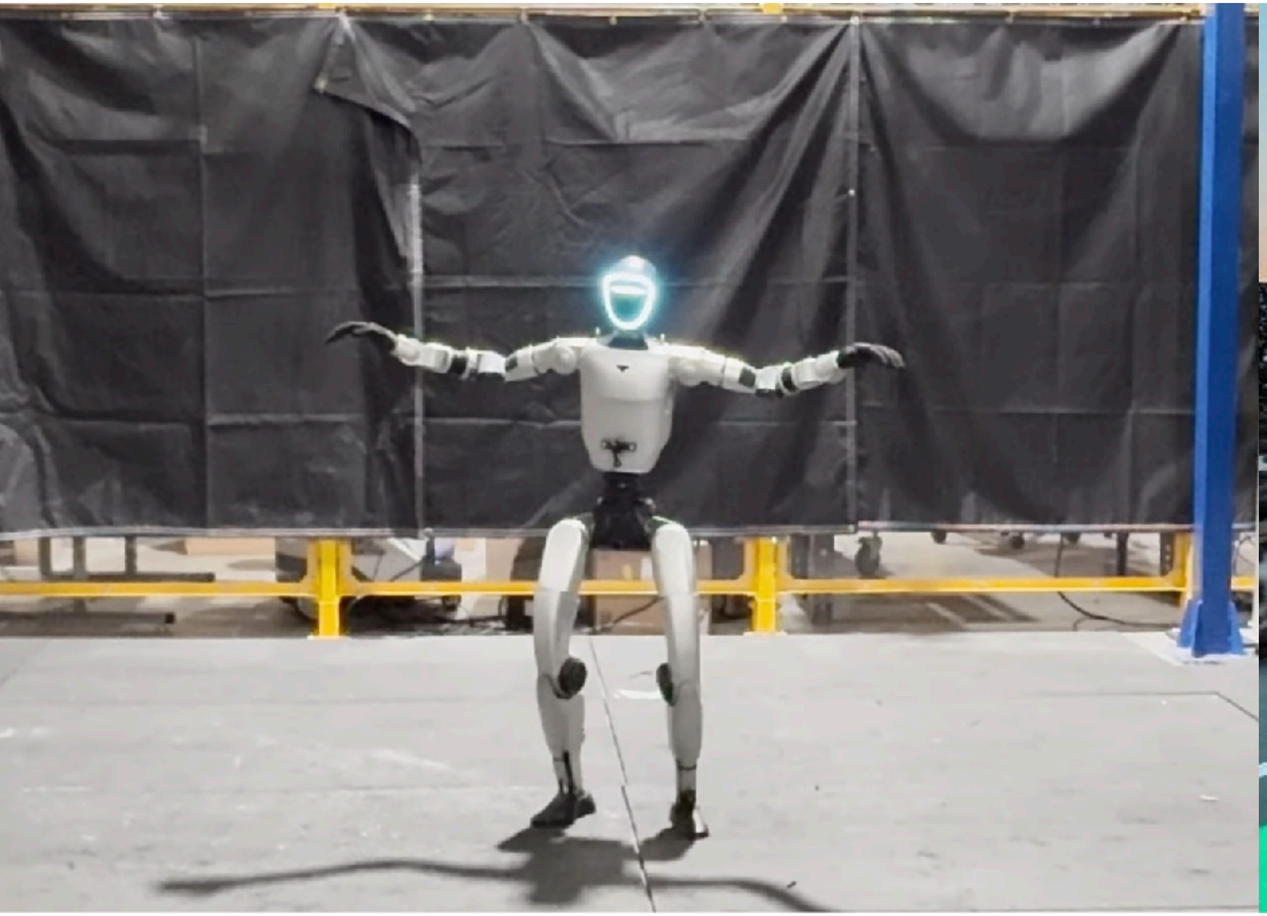
To Distill or Decide?

Understanding the Algorithmic Trade-off in Partially Observable Reinforcement Learning

Yuda Song, RL Theory Seminar, March 17th

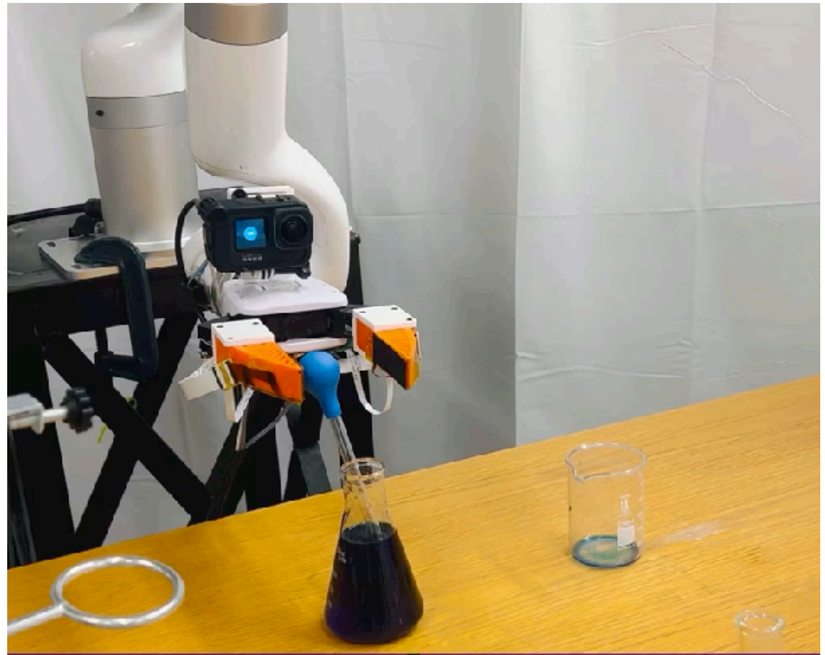
Practical Applications Involve High Dimensional Inputs

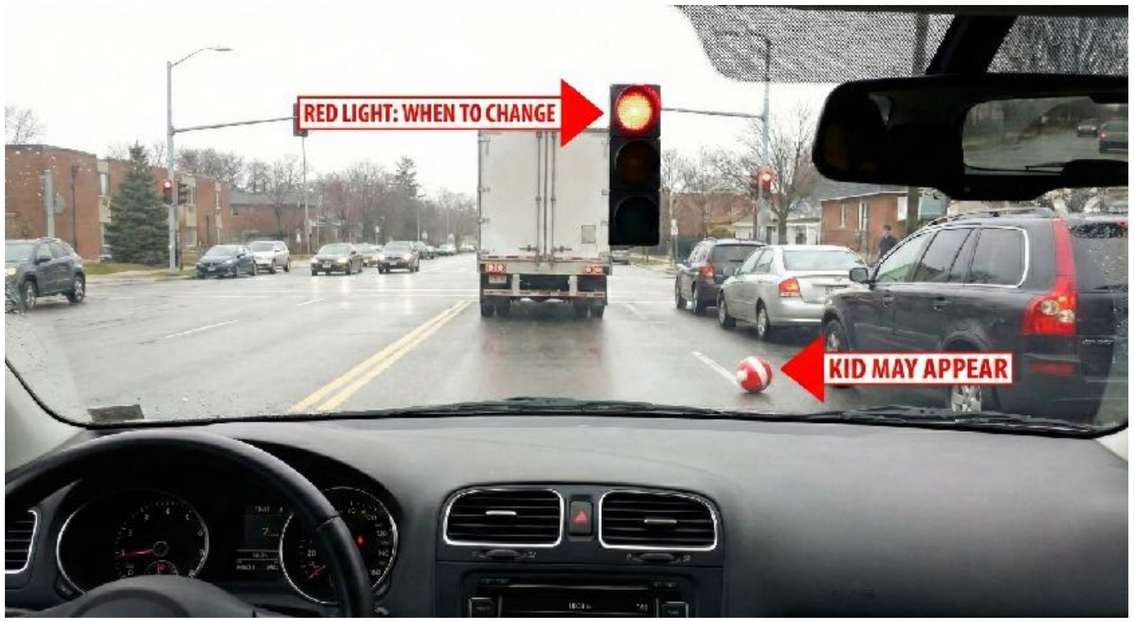


Rich Observation Environments with Latent Space

Partially Observable Models

- Such mapping doesn't always exist because of partial observation.
  - I.e., states might not be unique given (history of) observations.
  - Lack of modalities — vision (observed) vs. tactile (not observed).
  - Occlusion in autonomous driving and robotics.

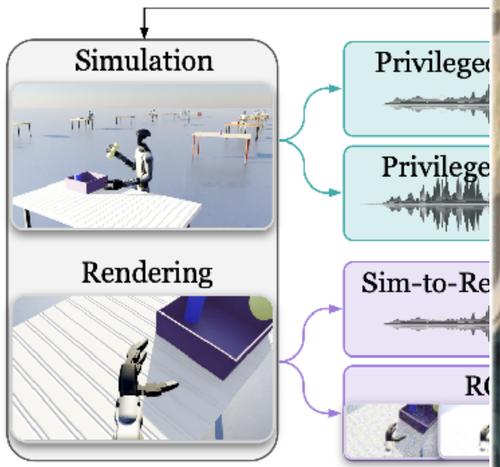




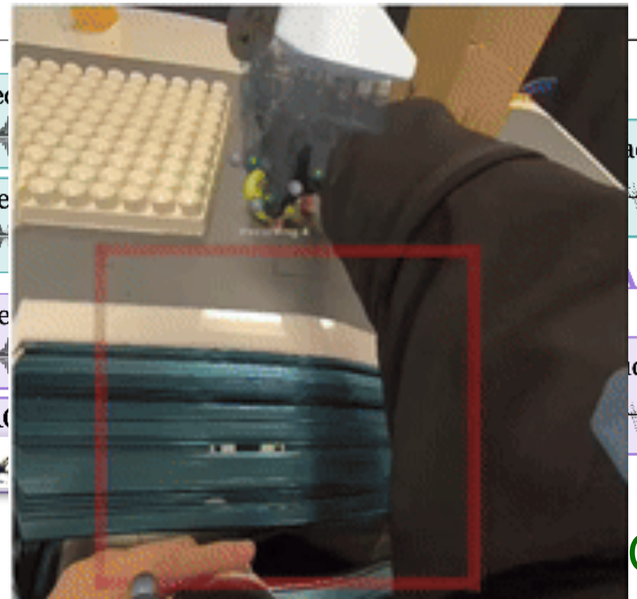
What do People Do in Practice?

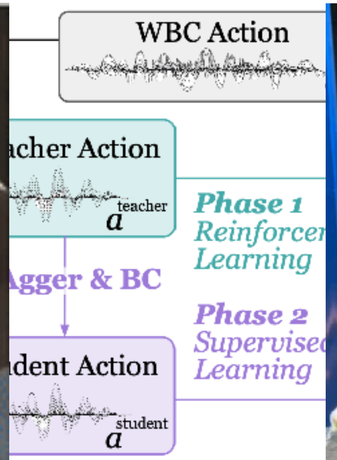
- End-to-end RL (expensive).
- Expert Distillation / Teacher Student Learning / Learning by Cheating:
  - Sim2Real, T
  - Imitation learning by mapping observation to expert action, produced by privileged policy seeing state.

What is the algorithmic tradeoff between the two in a realistic setting?

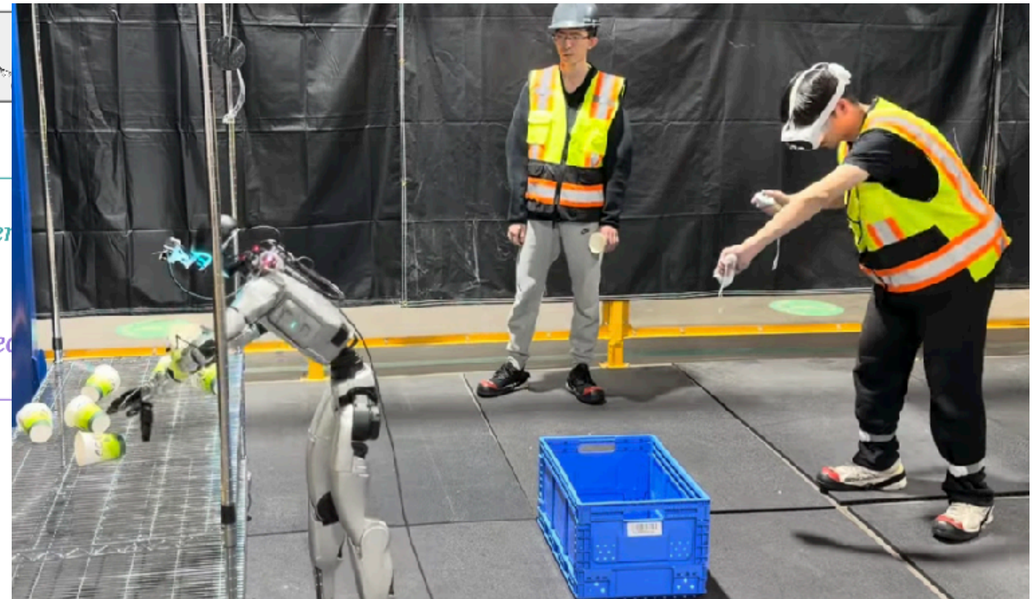


HWXB+25





CWNF+25



Part 1: Overview of the results (theoretical and empirical)

Part 2: Details of the theoretical results

Part 1: Overview of the results (theoretical and empirical)

Part 2: Details of the theoretical results

Overview: Expert Distillation Requires Stronger Error Conditions

Expert Distillation

$$\text{Poly}(\epsilon_{\text{h}}^{\text{dec}}(\pi), \epsilon_{\text{h}}^{\text{con}}(\pi; \text{L}))$$

RL

$$\text{Poly}(\epsilon_{\text{h}}^{\text{con}}(\pi; \text{L}))$$

Belief Contraction Error:
Error from using recent frames instead of full history.

Decodability Error:
Measuring how non-concentrated belief state is.

Intuition: 0 if no partial observation — misspecification to true latent expert.

Overview: Expert Distillation Requires Stronger Error Conditions

Expert Distillation

$\mathrm{Poly}(\epsilon_{\mathrm{h}}^{\mathrm{dec}}(\pi), \epsilon_{\mathrm{h}}^{\mathrm{con}}(\pi; \mathrm{L}))$

RL

$\mathrm{Poly}(\epsilon_{\mathrm{h}}^{\mathrm{con}}(\pi; \mathrm{L}))$

Under benign observability:
(Perturbed Block MDPs)

Belief Contraction Error:
Error from using recent frames instead of full history.

Decodability Error:
Measuring how non-concentrated belief state is.

Overview: Expert Distillation Requires Stronger Error Conditions

Expert Distillation

$\mathrm{Poly}(\epsilon_{\mathrm{h}}^{\mathrm{dec}}(\pi), \epsilon_{\mathrm{h}}^{\mathrm{con}}(\pi; \mathrm{L}))$

RL

$\mathrm{Poly}(\epsilon_{\mathrm{h}}^{\mathrm{con}}(\pi; \mathrm{L}))$ ✓

Under benign observability:
(Perturbed Block MDPs)

Belief Contraction Error:
controllable.

Decodability Error:
Measuring how non-concentrated belief state is.

Overview: Expert Distillation Requires Stronger Error Conditions

Expert Distillation

$\mathrm{Poly}(\epsilon_{\mathrm{h}}^{\mathrm{dec}}(\pi), \epsilon_{\mathrm{h}}^{\mathrm{con}}(\pi; \mathrm{L}))$

RL

$\mathrm{Poly}(\epsilon_{\mathrm{h}}^{\mathrm{con}}(\pi; \mathrm{L}))$ ✓

Under benign observability:
(Perturbed Block MDPs)

Belief Contraction Error:
controllable.

Decodability Error:
determined by the latent dynamics.

Overview: Expert Distillation Requires Stronger Error Conditions

Expert Distillation

$\mathrm{Poly}(\epsilon_{\mathrm{h}}^{\mathrm{dec}}(\pi), \epsilon_{\mathrm{h}}^{\mathrm{con}}(\pi; \mathrm{L}))$

RL

$\mathrm{Poly}(\epsilon_{\mathrm{h}}^{\mathrm{con}}(\pi; \mathrm{L}))$ ✓

Under benign observability:
(Perturbed Block MDPs)

Belief Contraction Error:
controllable.

Decodability Error:
determined by the latent dynamics.

Deterministic ✓

controllable

Overview: Expert Distillation Requires Stronger Error Conditions

Expert Distillation

$\mathrm{Poly}(\epsilon_{\mathrm{h}}^{\mathrm{dec}}(\pi), \epsilon_{\mathrm{h}}^{\mathrm{con}}(\pi; \mathrm{L}))$

RL

$\mathrm{Poly}(\epsilon_{\mathrm{h}}^{\mathrm{con}}(\pi; \mathrm{L}))$ ✓

Under benign observability:
(Perturbed Block MDPs)

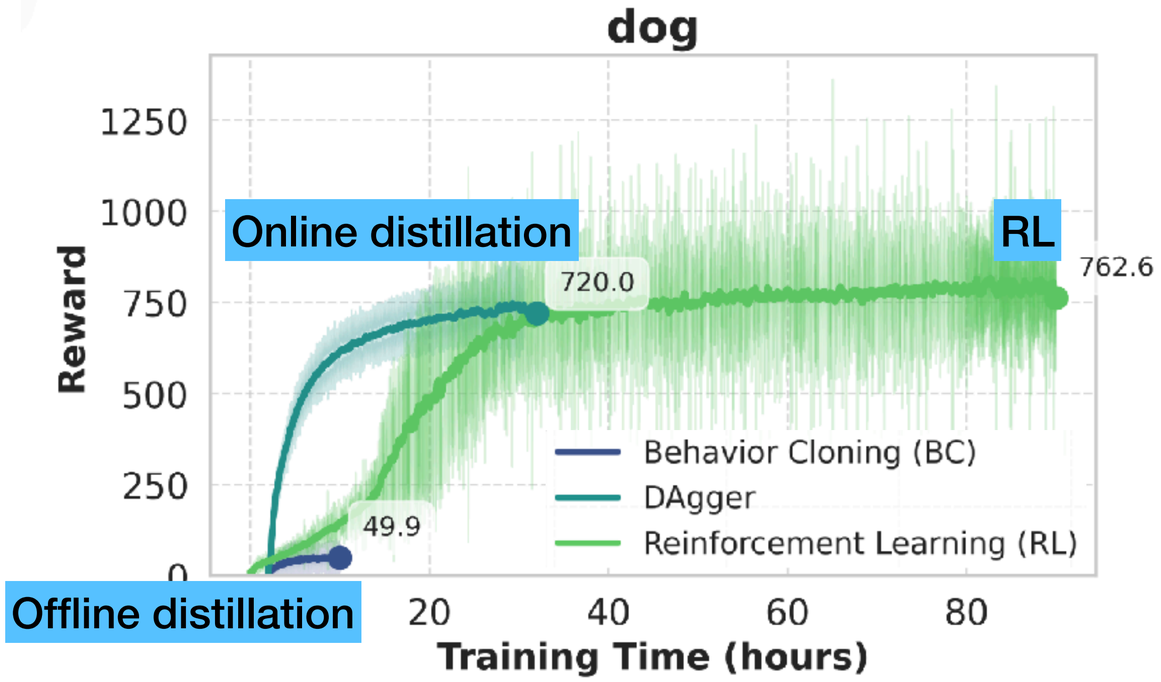
Belief Contraction Error:
controllable.

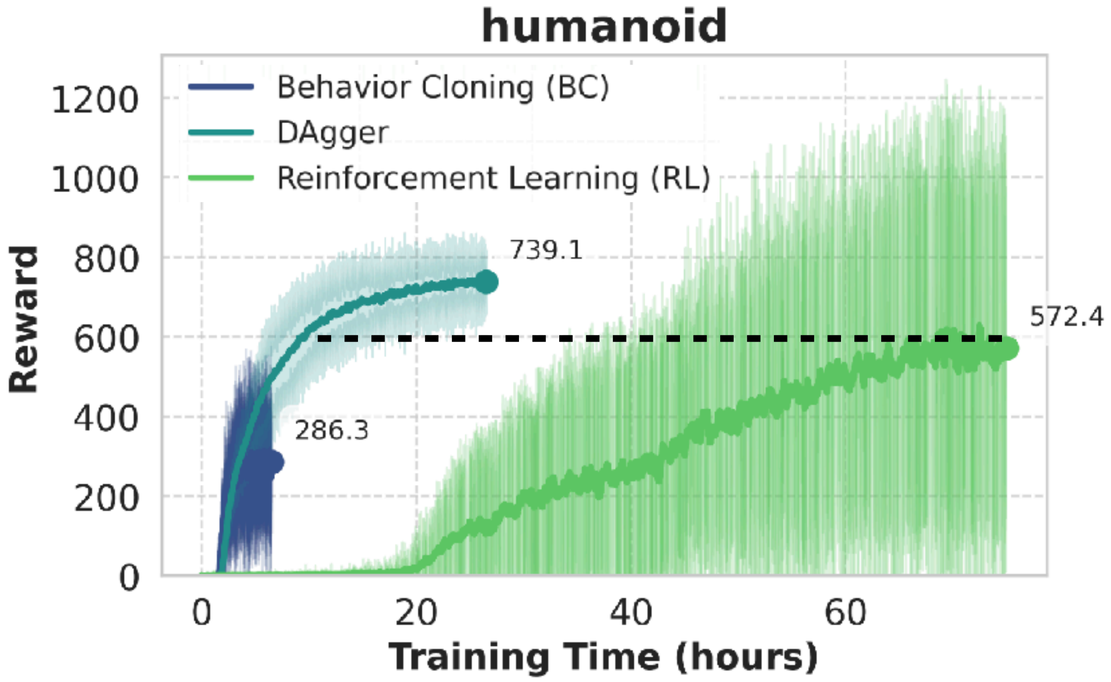
Decodability Error:
determined by the latent dynamics.

Deterministic ✓
controllable

Stochastic ✗
irreducible

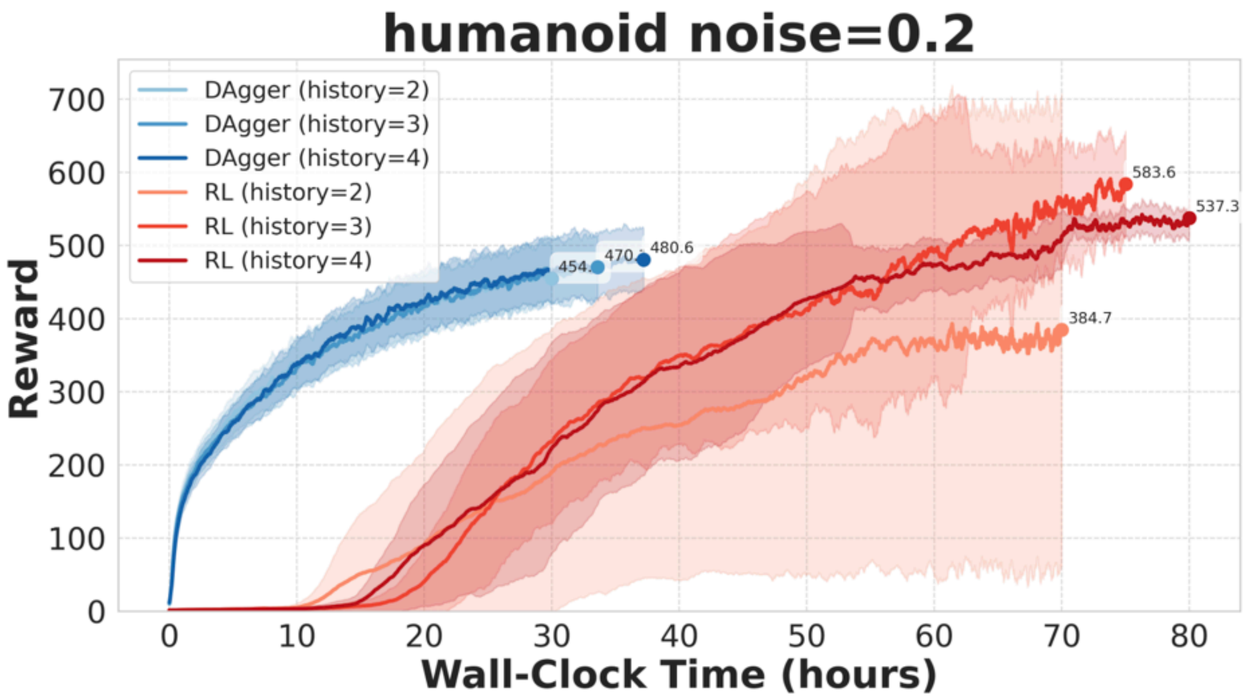
Distillation is More Efficient under Deterministic Latent

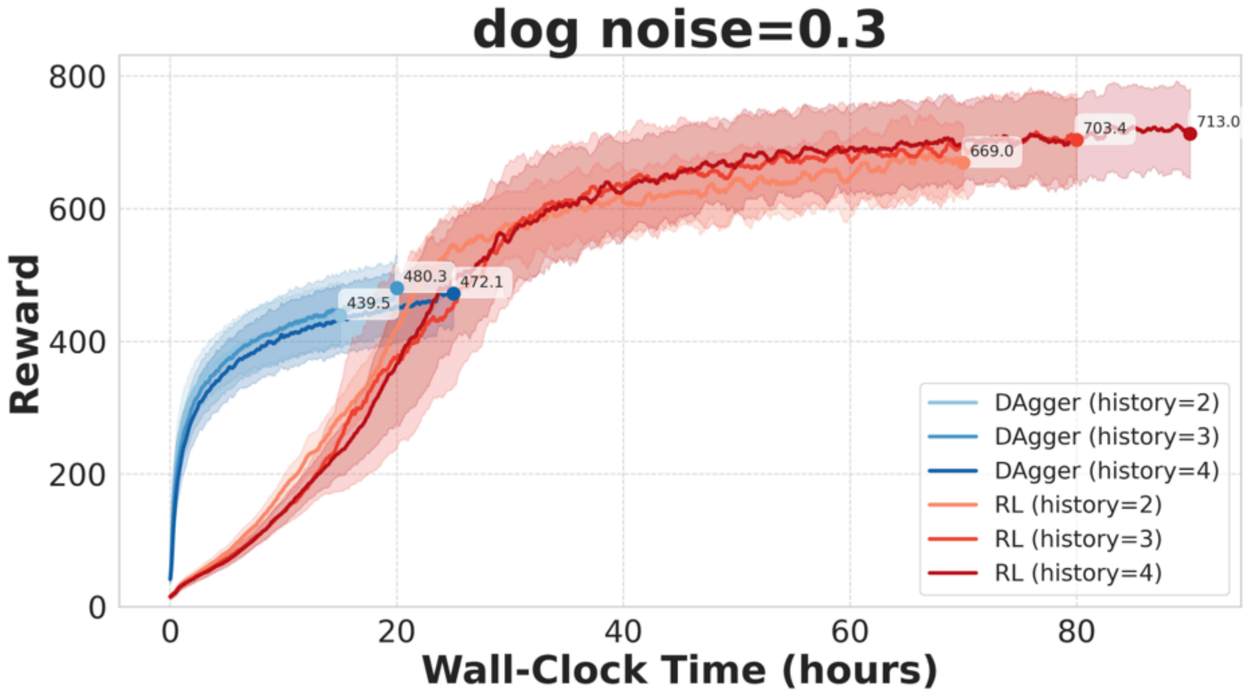




With proper algorithmic choice, distillation can be more efficient than RL in practice under high observability and deterministic latent dynamics.

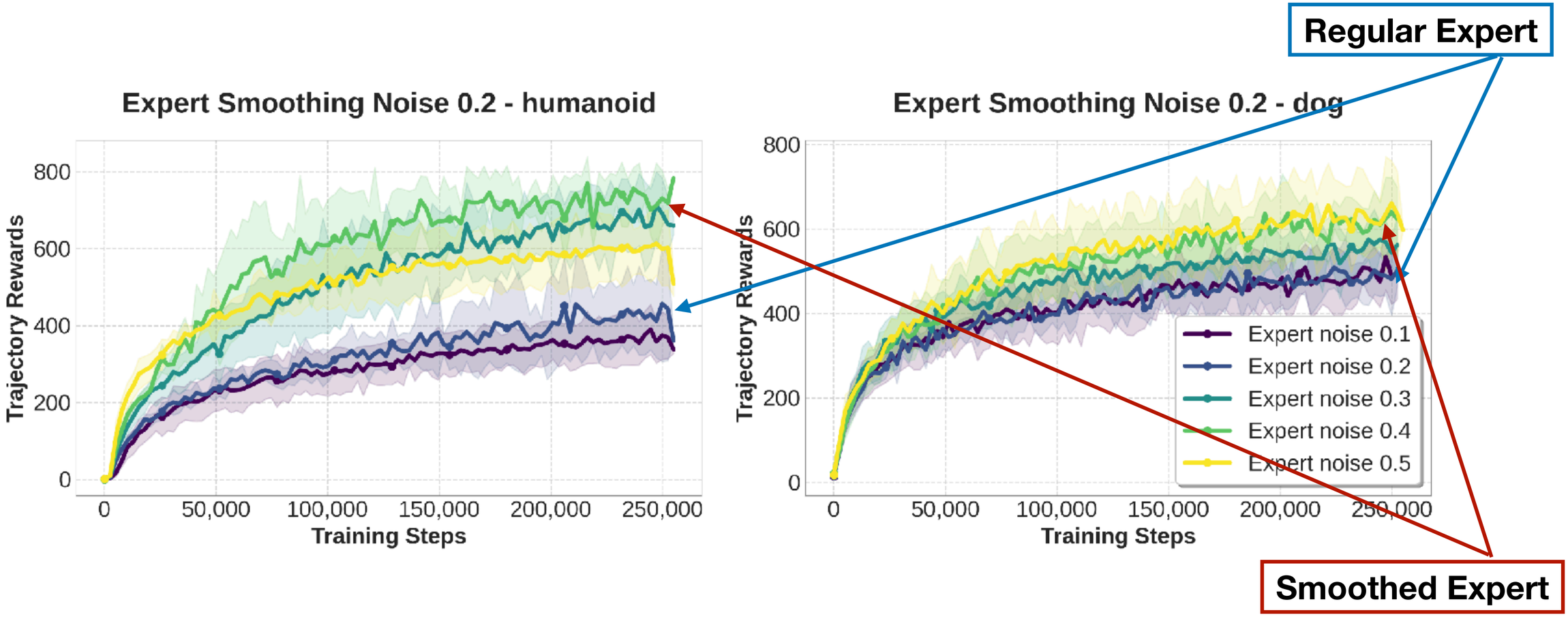
RL is Performs Better under Stochastic Latent





RL outperforms expert distillation under stochastic latent dynamics with enough frame stacks (to reduce belief contration error).

Intervention Preview: Smoothing the Expert



Smoothing the expert can reduce a variant notion of decidability error thus improve the performance of expert distillation.

Part 1: Overview of the results (theoretical and empirical)

Part 2: Details of the theoretical results

Setup: Partially Observable Markov Decision Process (POMDP)

observation history:
given $h, L \in [H]$,
$\mathscr{X}^{h-L:h} = \mathscr{X}_{h-L} \times \mathscr{X}_{h-L+1} \times \ldots \times \mathscr{X}_h$

observation $\mathscr{X} = \{\mathscr{X}_h\}_{h=1}^H$

state $\mathscr{S} = \{\mathscr{S}_h\}_{h=1}^H$

action $\mathscr{A} = \{\mathscr{A}_h\}_{h=1}^H$

POMDP

$$\mathscr{P} = \{H, \mathscr{X}, \mathscr{S}, \mathscr{A}, \mathbb{P}, \mathbb{O}, R\}$$

latent dynamics:
$\mathbb{P}_h : \mathscr{S}_{h-1} \times \mathscr{A}_{h-1} \to \Delta(\mathscr{S}_h)$

emission distribution:
$\mathbb{O}_h : \mathscr{S}_h \to \Delta(\mathscr{X}_h)$

Setup: Executable Policy

observation history:
given $h, L \in [H]$,
$\mathcal{X}^{h-L:h} = \mathcal{X}_{h-L} \times \mathcal{X}_{h-L+1} \times \ldots \times \mathcal{X}_h$

observation $\mathcal{X} = \{\mathcal{X}_h\}_{h=1}^H$ state $\mathcal{S} = \{\mathcal{S}_h\}_{h=1}^H$ action $\mathcal{A} = \{\mathcal{A}_h\}_{h=1}^H$

POMDP

$$\mathcal{P} = \{H, \mathcal{X}, \mathcal{S}, \mathcal{A}, \mathbb{P}, \mathbb{O}, R\}$$

latent dynamics: $\mathbb{P}_h : \mathcal{S}_{h-1} \times \mathcal{A}_{h-1} \to \Delta(\mathcal{S}_h)$

emission distribution: $\mathbb{O}_h : \mathcal{S}_h \to \Delta(\mathcal{X}_h)$

L-step executable policy: $\Pi^L = \{\pi_h : \mathcal{X}^{h-L:h} \times \mathcal{A}^{h-L:h-1} \to \Delta(\mathcal{A}_h)\}$

Trajectory $\tau = \{s_1, x_1, a_1, r_1, \ldots, s_H, x_H, a_H, r_H\}$:

$$s_h \sim \mathbb{P}_h(s_{h-1}, a_{h-1}), x_h \sim \mathbb{O}_h(s_h), a_h \sim \pi(x_{h-L:h}, a_{h-L:h-1}), r_h = R(s_h, a_h).$$

Setup: Latent Policy

observation history:
given $h, L \in [H]$,
$\mathscr{X}^{h-L:h} = \mathscr{X}_{h-L} \times \mathscr{X}_{h-L+1} \times \ldots \times \mathscr{X}_h$

observation $\mathscr{X} = \{\mathscr{X}_h\}_{h=1}^H$ state $\mathscr{S} = \{\mathscr{S}_h\}_{h=1}^H$ action $\mathscr{A} = \{\mathscr{A}_h\}_{h=1}^H$

POMDP

$$\mathscr{P} = \{H, \mathscr{X}, \mathscr{S}, \mathscr{A}, \mathbb{P}, \mathbb{O}, R\}$$

latent dynamics:
$\mathbb{P}_h : \mathscr{S}_{h-1} \times \mathscr{A}_{h-1} \to \Delta(\mathscr{S}_h)$

emission distribution:
$\mathbb{O}_h : \mathscr{S}_h \to \Delta(\mathscr{X}_h)$

L-step executable policy: $\Pi^L = \{\pi_h : \mathscr{X}^{h-L:h} \times \mathscr{A}^{h-L:h-1} \to \Delta(\mathscr{A}_h)\}$

Trajectory $\tau = \{s_1, x_1, a_1, r_1, \ldots, s_H, x_H, a_H, r_H\}$:

$$s_h \sim \mathbb{P}_h(s_{h-1}, a_{h-1}), x_h \sim \mathbb{O}_h(s_h), a_h \sim \pi(x_{h-L:h}, a_{h-L:h-1}), r_h = R(s_h, a_h).$$

Underlying MDP $\mathscr{M} = \{H, \mathscr{S}, \mathscr{A}, \mathbb{P}, R\}$

Latent policy $\Pi^{\text{latent}} = \{\pi_h^{\text{latent}} : \mathscr{S}_h \to \Delta(\mathscr{A}_h)\}$

Setup: Belief State

observation history:
given $h, L \in [H]$,
$\mathcal{X}^{h-L:h} = \mathcal{X}_{h-L} \times \mathcal{X}_{h-L+1} \times \ldots \times \mathcal{X}_h$

observation $\mathcal{X} = \{\mathcal{X}_h\}_{h=1}^H$ state $\mathcal{S} = \{\mathcal{S}_h\}_{h=1}^H$ action $\mathcal{A} = \{\mathcal{A}_h\}_{h=1}^H$

POMDP

$$\mathcal{P} = \{H, \mathcal{X}, \mathcal{S}, \mathcal{A}, \mathbb{P}, \mathbb{O}, R\}$$

latent dynamics:
$\mathbb{P}_h : \mathcal{S}_{h-1} \times \mathcal{A}_{h-1} \to \Delta(\mathcal{S}_h)$

emission distribution:
$\mathbb{O}_h : \mathcal{S}_h \to \Delta(\mathcal{X}_h)$

Belief state: a distribution over the latent state

$$\mathbb{B}_h(b; x_h)(s_h) := \frac{\mathbb{O}_h(x_h \mid s_h)\, b(s_h)}{\sum_{z_h \in \mathcal{S}_h} \mathbb{O}_h(x_h \mid z_h)\, b(z_h)}$$

$$\mathbf{b}_h(x_{1:h}, a_{1:h-1}) = \mathbb{U}_h(\mathbf{b}_{h-1}(x_{1:h-1}, a_{1:h-2}); a_{h-1}, x_h)$$

$$\mathbb{U}_h(b; a_{h-1}, x_h) := \mathbb{B}_h\big(\mathbb{P}_h(a_{h-1}) \cdot b;\ x_h\big)$$

We can construct an executable policy given a belief b and latent policy π^{latent}:

$$(\pi_h^{\text{latent}} \circ b)(a_h) = \sum_{s_h \in \mathcal{S}_h} b(s_h)\pi_h(a_h \mid s_h)$$

Expert Distillation

Latent expert: $\pi^{\text{latent}} \in \Pi^{\text{latent}}$, training data distribution $\{s_h, x_h, a_h \sim \mu_h\}_{h=1}^{H}$.

Objective:

$$\min_{\pi \in \Pi^L} \sum_{h=1}^{H} \widehat{\mathbb{E}}_{x_{h-L:h}, s_h, a_h \sim \mu_h} \left[\ell(\pi^{\text{latent}}(\cdot \mid s_h); \pi(\cdot \mid x_{h-L:h}, a_{h-L:h-1})) \right]$$

$\mu = d^{\pi^{\text{latent}}}$: Behavior Cloning; $\mu = d^{\pi}$: DAgger/Forward

Ideal outcome:

$$\pi(x_{h-L:h}, a_{h-L:h-1}) = \pi^{\text{latent}} \circ \mathbf{b}^{\text{apx}}(x_{h-L:h}, a_{h-L:h-1})$$

optimal belief under L-step history and learning policy

(achievable with Forward)

Expert Distillation Error Decomposition

$$\pi(x_{h-L:h}, a_{h-L:h-1}) :$$

$$\pi^{\text{latent}} \circ \mathbf{b}^{\text{apx}}(x_{h-L:h}, a_{h-L:h-1})$$

Lemma (Sub-optimality of Composed Policy)

$$J(\pi^*) - J(\pi)$$

Expert Distillation Error Decomposition

$$\pi(x_{h-L:h}, a_{h-L:h-1}) :$$

$$\pi^{\text{latent}} \circ \mathbf{b}^{\text{apx}}(x_{h-L:h}, a_{h-L:h-1})$$

Lemma (Sub-optimality of Composed Policy)

$$J(\pi^*) - J(\pi) \leq J(\pi^{\text{latent}}) - J(\pi)$$

Expert Distillation Error Decomposition

$$\pi(x_{h-L:h}, a_{h-L:h-1}):$$

$$\pi^{\text{latent}} \circ \mathbf{b}^{\text{apx}}(x_{h-L:h}, a_{h-L:h-1})$$

Lemma (Sub-optimality of Composed Policy)

$$J(\pi^*) - J(\pi) \leq J(\pi^{\text{latent}}) - J(\pi) \leq \text{TV}(\mathbb{P}^{\pi^{\text{latent}}}, \mathbb{P}^{\pi})$$

Expert Distillation Error Decomposition

$$\pi(x_{h-L:h}, a_{h-L:h-1}) :$$

$$\pi^{\text{latent}} \circ \mathbf{b}^{\text{apx}}(x_{h-L:h}, a_{h-L:h-1})$$

Lemma (Sub-optimality of Composed Policy)

$$J(\pi^*) - J(\pi) \leq J(\pi^{\text{latent}}) - J(\pi) \leq \text{TV}(\mathbb{P}^{\pi^{\text{latent}}}, \mathbb{P}^{\pi}) \leq \sum_{h=1}^{H} 2\epsilon_h^{\text{dec}}(\pi) + \tilde{\epsilon}_h^{\text{con}}(\pi; L))$$

Expert Distillation Error Decomposition

$\pi(x_{h-L:h}, a_{h-L:h-1})$:

$\pi^{\text{latent}} \circ \mathbf{b}^{\text{apx}}(x_{h-L:h}, a_{h-L:h-1})$

Lemma (Sub-optimality of Composed Policy)

$$J(\pi^*) - J(\pi) \le J(\pi^{\text{latent}}) - J(\pi) \le \mathrm{TV}(\mathbb{P}^{\pi^{\text{latent}}}, \mathbb{P}^{\pi}) \le \sum_{h=1}^{H} 2\epsilon_h^{\text{dec}}(\pi) + \tilde{\epsilon}_h^{\text{con}}(\pi; L))$$

Decodability error: $\epsilon_h^{\text{dec}}(\pi) := \mathbb{E}^{\pi}[1 - \|\mathbf{b}_h(x_{1:h}, a_{1:h-1})\|_\infty]$

belief with L-step history with uniform prior

Belief Contraction Error:

$\epsilon_h^{\text{con}}(\pi; L) := \mathbb{E}^{\pi}[\|\mathbf{b}_h(x_{1:h}, a_{1:h-1}) - \mathbf{b}_h^{\text{apx}}(x_{h-L+1:h}, a_{h-L:h-1}; \mathrm{Unif}(\mathcal{S}_{h-L}))\|_1]$

RL result (GMR22): $J(\pi^*) - J(\pi^{\text{RL}}) \le \epsilon^{\text{con}}(\pi; L) \cdot \mathrm{Poly}(S, X, H, \gamma^{-1})$ with quasi-poly time.

Perturbed Block MDPs

- How does these two errors behave in a practically relevant setting?
- Practical intuition: only small chance of partial observation due to, e.g., periodic occlusion.

Perturbed Block MDPs

The emission distribution in perturbed Block MDPs follows the block structure with probability $1 - \delta$ and arbitrary otherwise:

$$\mathbb{O}_h(x \mid s) = (1 - \delta)\mathbb{O}_h^{\text{block}}(x \mid s) + \delta E_h(x \mid s)\,.$$

block structure: support of $\mathbb{O}_h^{\text{block}}(\,\cdot \mid s)$ is disjoint for all $s \in \mathcal{S}$

Overview: Expert Distillation Requires Stronger Error Conditions

Expert Distillation

$\mathrm{Poly}(\epsilon_{\mathrm{h}}^{\mathrm{dec}}(\pi), \epsilon_{\mathrm{h}}^{\mathrm{con}}(\pi; \mathrm{L}))$

RL

$\mathrm{Poly}(\epsilon_{\mathrm{h}}^{\mathrm{con}}(\pi; \mathrm{L}))$ ✓

Under benign observability:
(Perturbed Block MDPs)

due to high observability

Belief Contraction Error:
decays exponentially w.r.t. frame stack.

Decodability Error:
determined by the latent dynamics.

Deterministic ✓
Decays exponentially w.r.t. current timestep

Stochastic ✗
Irreducible sub-optimality (linear in horizon)

latent dynamics does not “spread” belief

Lower Bound under Stochastic Latent Dynamics

Theorem (Lower bound of expert distillation)

There exists a δ-perturbed Block MDP with stochastic dynamics such that for all $L \in [H]$, the optimal latent policy π^{latent} satisfies:

$$\min_{\pi \in \Pi^L} \mathrm{TV}(\mathbb{P}^{\pi^{\text{latent}}}, \mathbb{P}^{\pi}) \geq \Omega(\min(1, \delta H)) \,.$$

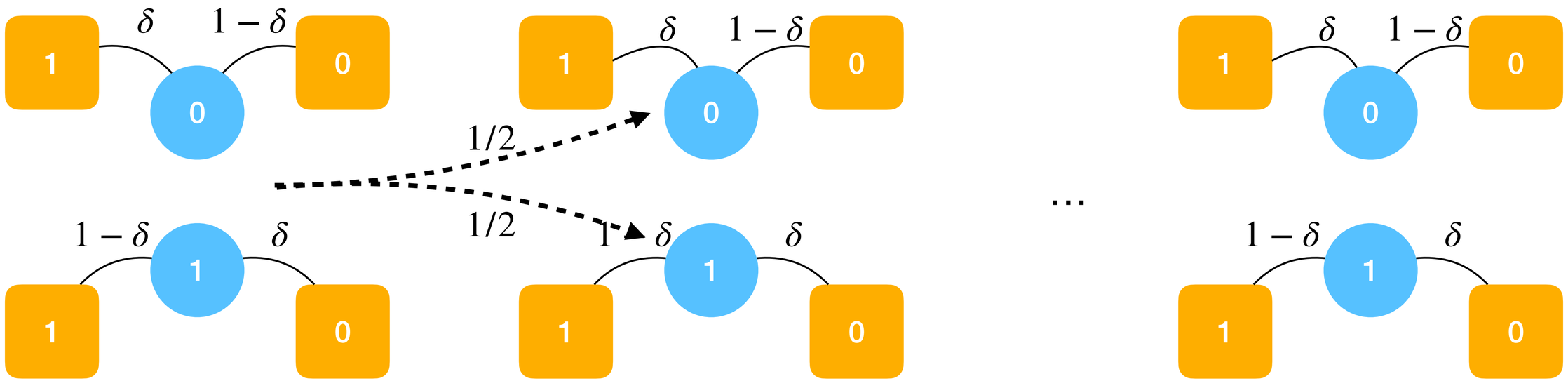
Irreducible misspecification to the latent expert!

Lower Bound under Stochastic Latent Dynamics

Theorem (Lower bound of expert distillation)

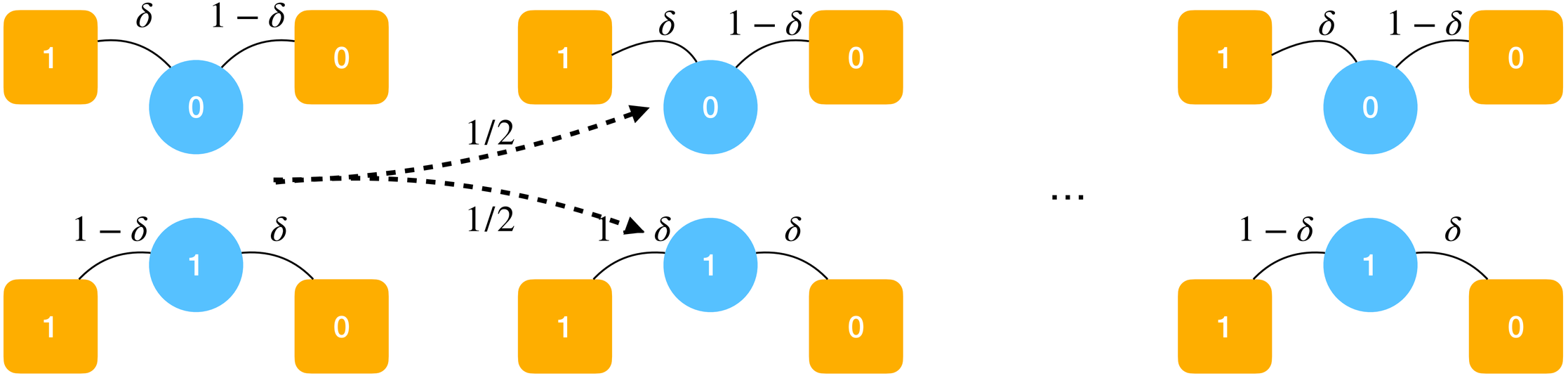
$$\min_{\pi \in \Pi^L} \mathrm{TV}(\mathbb{P}^{\pi^{\text{latent}}}, \mathbb{P}^{\pi}) \geq \Omega(\min(1, \delta H)) .$$

$\pi^{\text{latent}}(s_h) = a_h, a_h = s_h$ $\quad \Pr^{\pi}(a_h = s_h \mid \tau_h) \leq 1 - \delta$ $\quad \Pr^{\pi}(\forall h \in [H], a_h = s_h) \leq (1-\delta)^H$



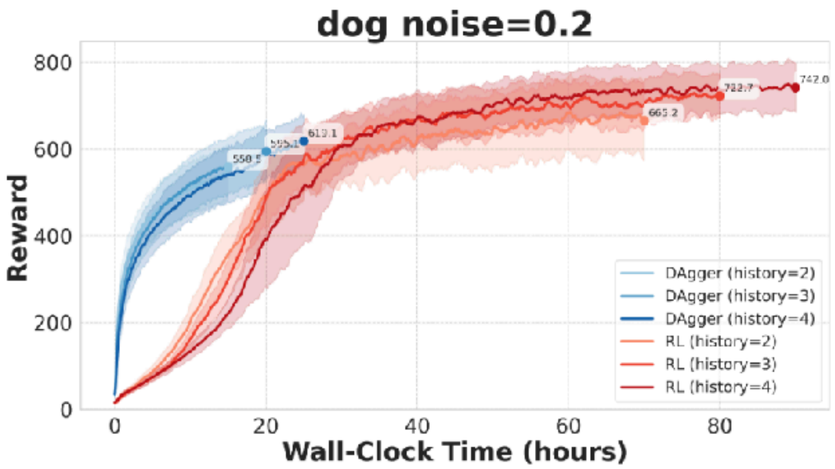
This is easy for RL (to learn the optimal executable policy): one step bandit problem for each tilmestep

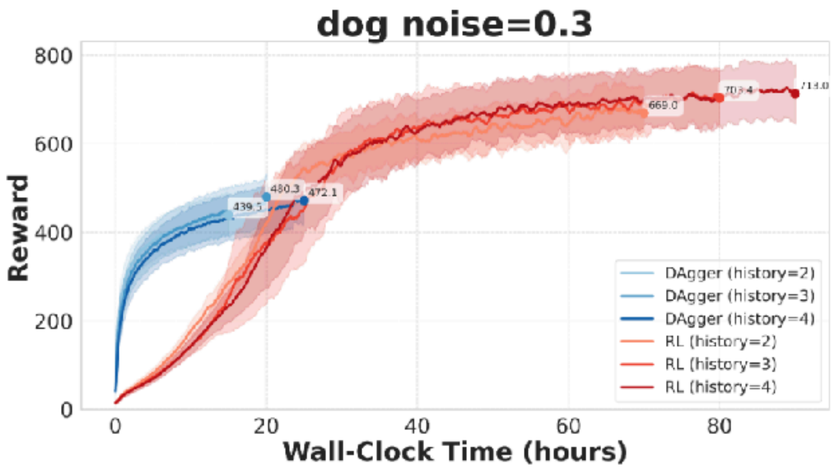
Lower Bound under Stochastic Latent Dynamics

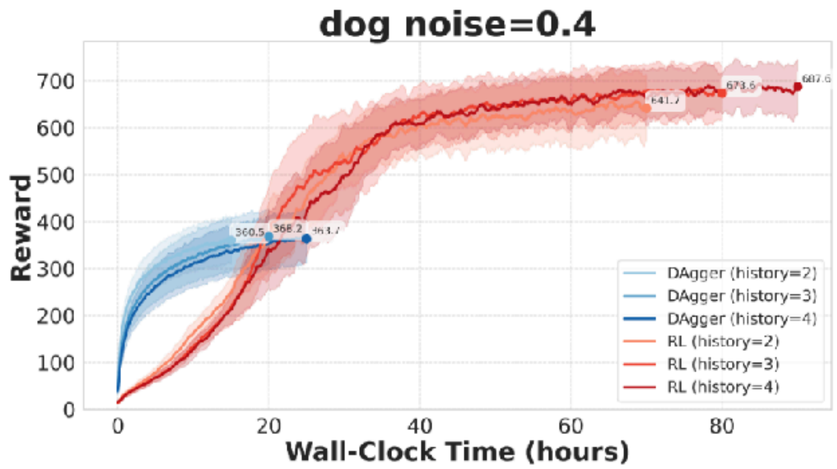


- Optimal executable policy (RL policy): $a_h = x_h$
- Expert distillation policy: $a_h = x_h$ with probability $1 - \delta$, $a_h = x_h$ with probability δ
- Suboptimality: $O(\delta)$
  - Can be made $O(\delta H)$ by adding an absorbing state

Increasing Gap with Higher Stochasticity







Decodability Error Can Be Loose

$$\pi(x_{h-L:h}, a_{h-L:h-1}):$$

$$\pi^{\text{latent}} \circ \mathbf{b}^{\text{apx}}(x_{h-L:h}, a_{h-L:h-1})$$

Lemma (Sub-optimality of Composed Policy)

$$J(\pi^*) - J(\pi) \leq J(\pi^{\text{latent}}) - J(\pi) \leq \text{TV}(\mathbb{P}^{\pi^{\text{latent}}}, \mathbb{P}^{\pi}) \leq \sum_{h=1}^{H} 2\epsilon_h^{\text{dec}}(\pi) + \tilde{\epsilon}_h^{\text{con}}(\pi; L))$$

Decodability error: $\epsilon_h^{\text{dec}}(\pi) := \mathbb{E}^{\pi}[1 - \|\mathbf{b}_h(x_{1:h}, a_{1:h-1})\|_{\infty}]$

Action Prediction Error: $\epsilon_h^{\text{act};\pi^{\text{latent}}}(\pi) := \mathbb{E}^{\pi}[1 - \|\pi^{\text{latent}} \circ \mathbf{b}_h(x_{1:h}, a_{1:h-1})\|_{\infty}]$

Decodability Error Can Be Loose

$$\pi(x_{h-L:h}, a_{h-L:h-1}):$$

$$\pi^{\text{latent}} \circ \mathbf{b}^{\text{apx}}(x_{h-L:h}, a_{h-L:h-1})$$

Lemma (Sub-optimality of Composed Policy)

$$J(\pi^*) - J(\pi) \leq J(\pi^{\text{latent}}) - J(\pi) \leq \text{TV}(\mathbb{P}^{\pi^{\text{latent}}}, \mathbb{P}^{\pi}) \leq \sum_{h=1}^{H} 2\epsilon_h^{\text{act};\pi^{\text{latent}}}(\pi) + \tilde{\epsilon}_h^{\text{con}}(\pi; L))$$

Decodability error: $\epsilon_h^{\text{dec}}(\pi) := \mathbb{E}^{\pi}[1 - \|\mathbf{b}_h(x_{1:h}, a_{1:h-1})\|_{\infty}]$

Action Prediction Error: $\epsilon_h^{\text{act};\pi^{\text{latent}}}(\pi) := \mathbb{E}^{\pi}[1 - \|\pi^{\text{latent}} \circ \mathbf{b}_h(x_{1:h}, a_{1:h-1})\|_{\infty}]$

Decodability Error Can Be Loose

$$\pi(x_{h-L:h}, a_{h-L:h-1}):$$

$$\pi^{\text{latent}} \circ \mathbf{b}^{\text{apx}}(x_{h-L:h}, a_{h-L:h-1})$$

Lemma (Sub-optimality of Composed Policy)

$$J(\pi^*) - J(\pi) \le J(\pi^{\text{latent}}) - J(\pi) \le \text{TV}(\mathbb{P}^{\pi^{\text{latent}}}, \mathbb{P}^{\pi}) \le \sum_{h=1}^{H} 2\epsilon_h^{\text{act};\pi^{\text{latent}}}(\pi) + \tilde{\epsilon}_h^{\text{con}}(\pi; L))$$

Decodability error: $\epsilon_h^{\text{dec}}(\pi) := \mathbb{E}^{\pi}[1 - \|\mathbf{b}_h(x_{1:h}, a_{1:h-1})\|_\infty]$

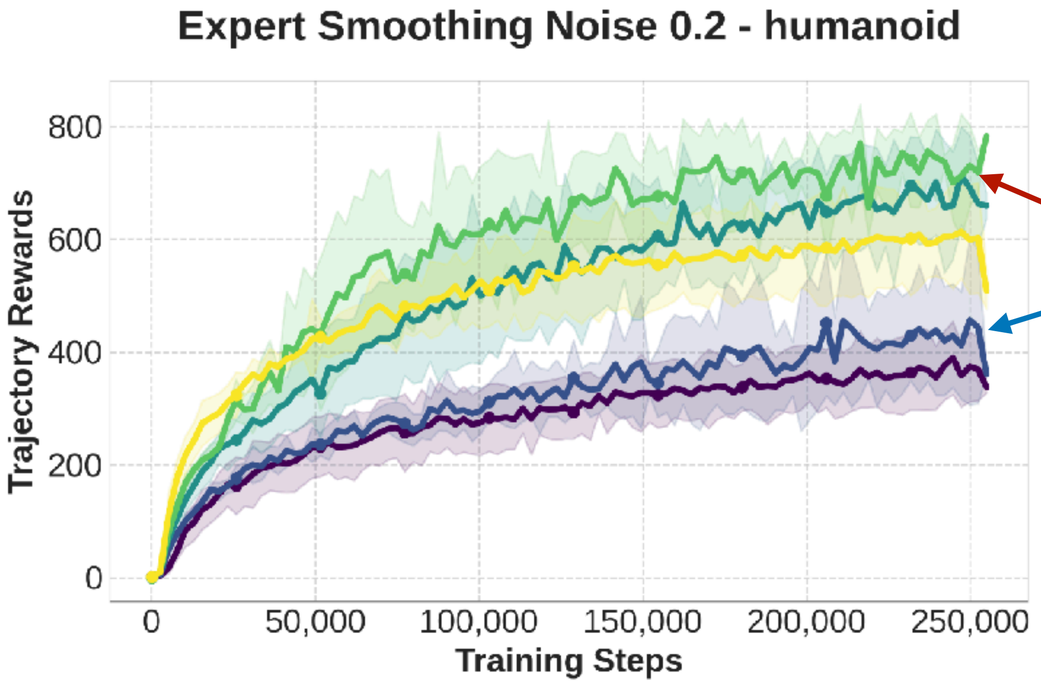
Action Prediction Error: $\epsilon_h^{\text{act};\pi^{\text{latent}}}(\pi) := \mathbb{E}^{\pi}[1 - \|\pi^{\text{latent}} \circ \mathbf{b}_h(x_{1:h}, a_{1:h-1})\|_\infty]$

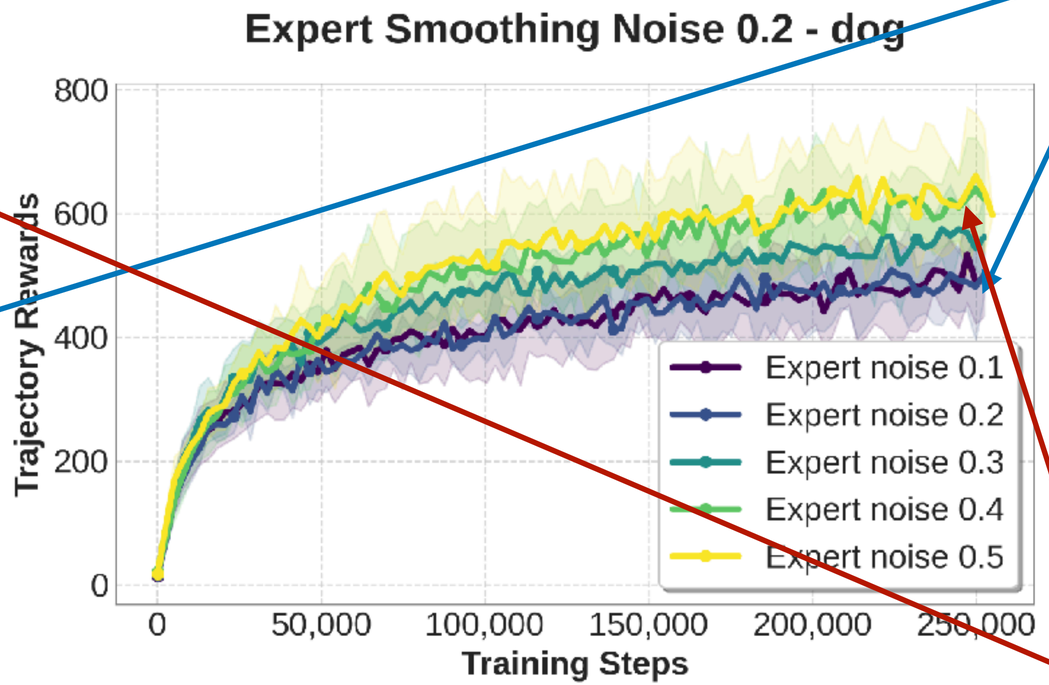
If π^{latent} is deterministic, in general we have $\epsilon_h^{\text{act};\pi^{\text{latent}}}(\pi) \le \epsilon_h^{\text{dec}}(\pi)$

Calling some notion of "smoother" expert

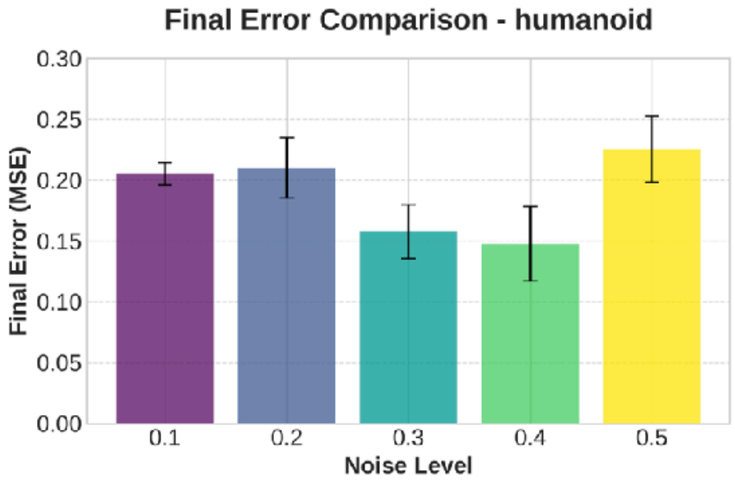
Smoothing the Expert

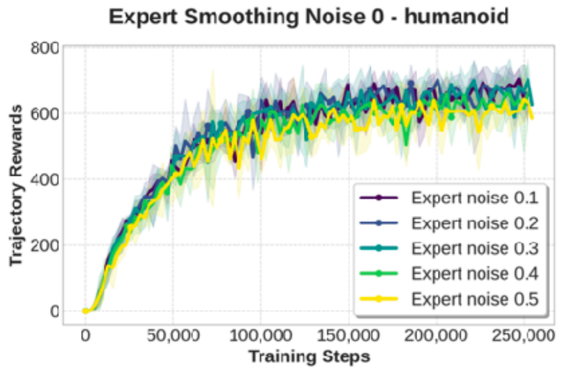
Regular Expert

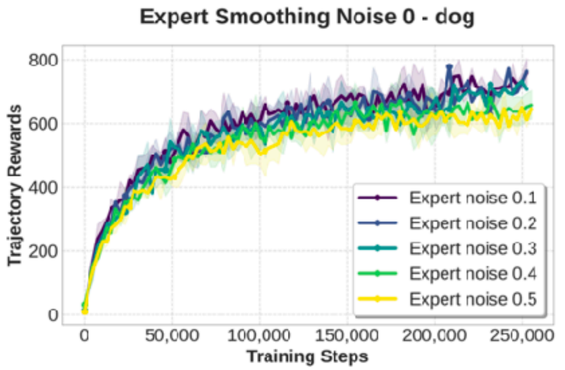




Smoothed Expert







Summary

Tools

Decodability Error

Belief Contraction Error

Perturbed Block MDPs

Tradeoffs

RL:
Computation inefficiency

Expert distillation:
Pitfall under stochastic latent dynamics

Intervention

Tighter bounds with action prediction error

Imitation smoother expert in practice