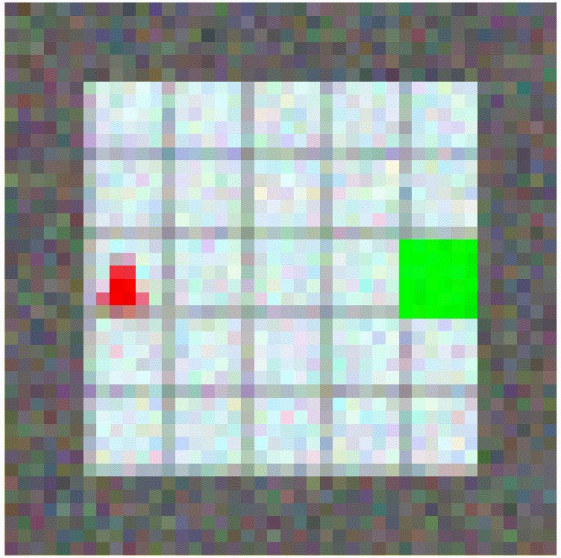
Robust Asymmetric Learning in POMDPs

Andrew Warrington*, J. Wilder Lavington*, Adam Ścibior, Mark Schmidt & Frank Wood

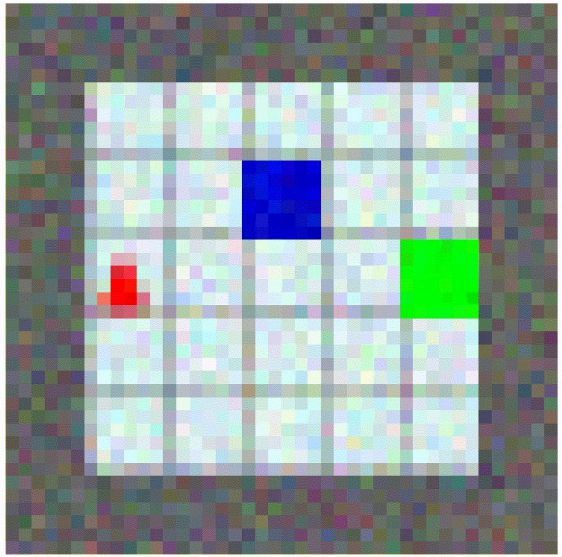
ICML 2021

Frozen Lake



Average reward: 4.00

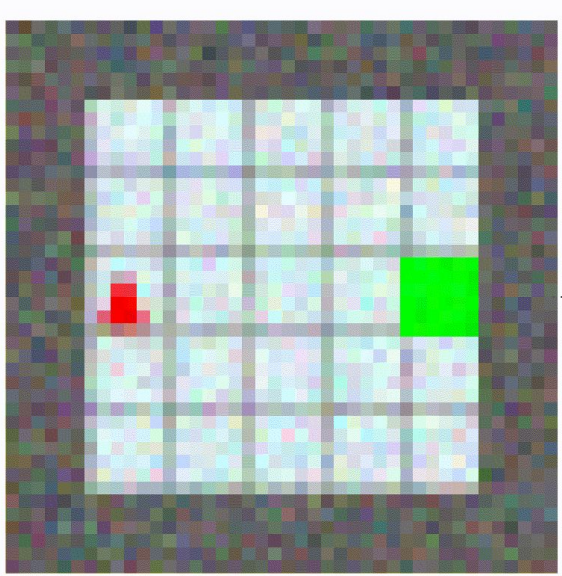
Frozen Lake



Average reward: 10.66

$$\texttt{flatten}\left(\left\{\begin{matrix}0&0&0&0&0\\0&0&0&0&0\\1&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\end{matrix}\right\},\left\{\begin{matrix}0&0&0\\0&0&0\\0&0&1\end{matrix}\right\}\right)$$

Frozen Lake

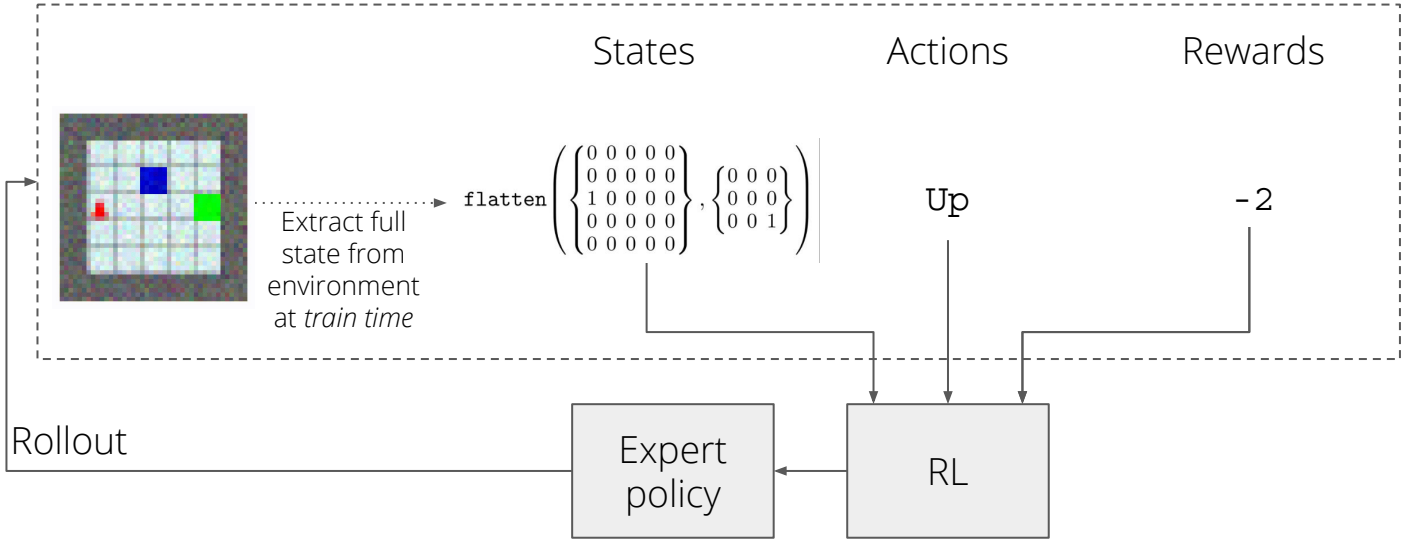


Average reward: 4.00

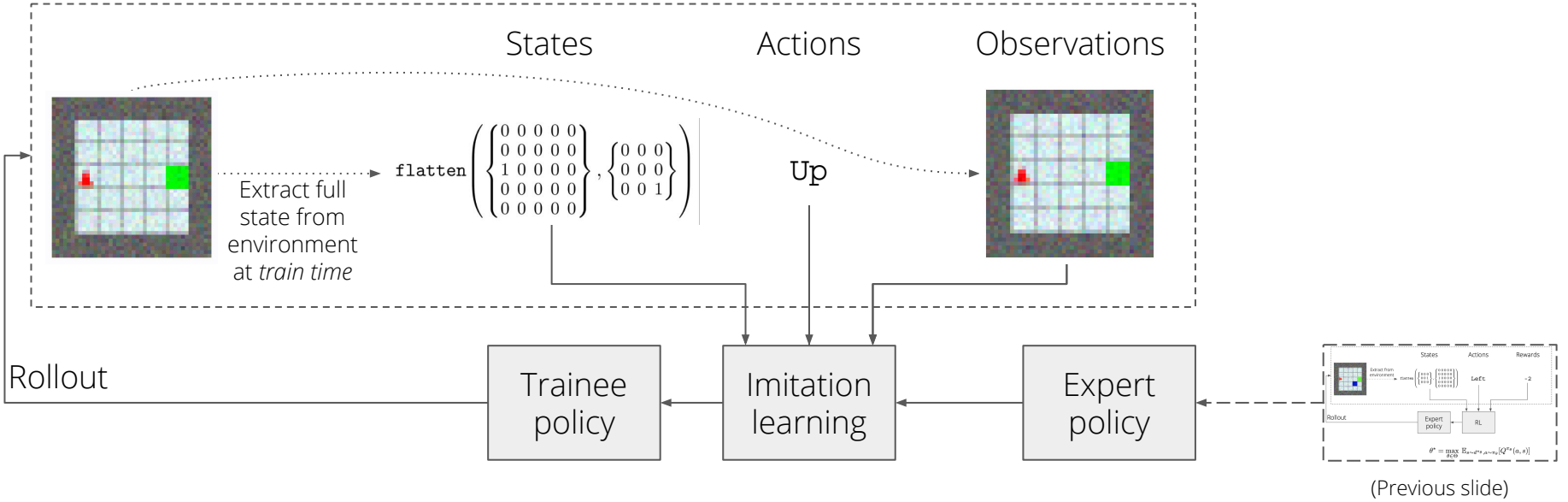
Extract full state from environment at *train time*

$$\texttt{flatten}\left(\left\{\begin{matrix}0&0&0&0&0\\0&0&0&0&0\\1&0&0&0&0\\0&0&0&0&0\\0&0&0&0&0\end{matrix}\right\},\left\{\begin{matrix}0&0&0\\0&0&0\\0&0&1\end{matrix}\right\}\right)$$

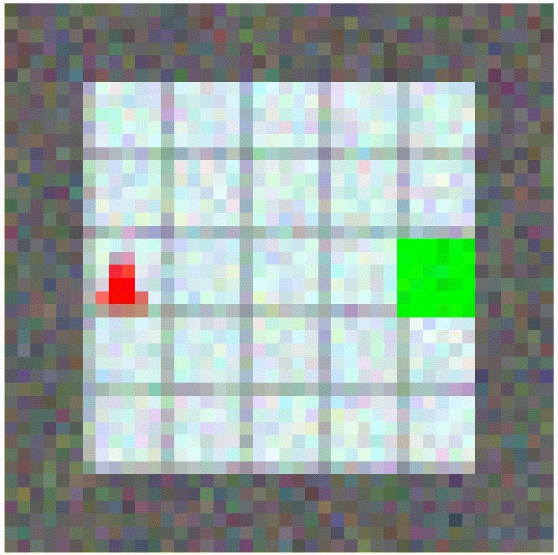
Learning *Experts* in MDPs with RL



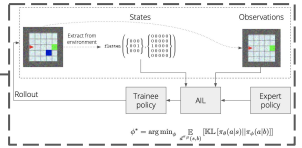
Learning *Trainees* in POMDPs with RL+AIL



Uh-oh



Average reward: -26.66



(Previous slide)

Why the failure?

Optimal POMDP policy

≠

Expectation of optimal MDP policy over missing information

Why the failure?

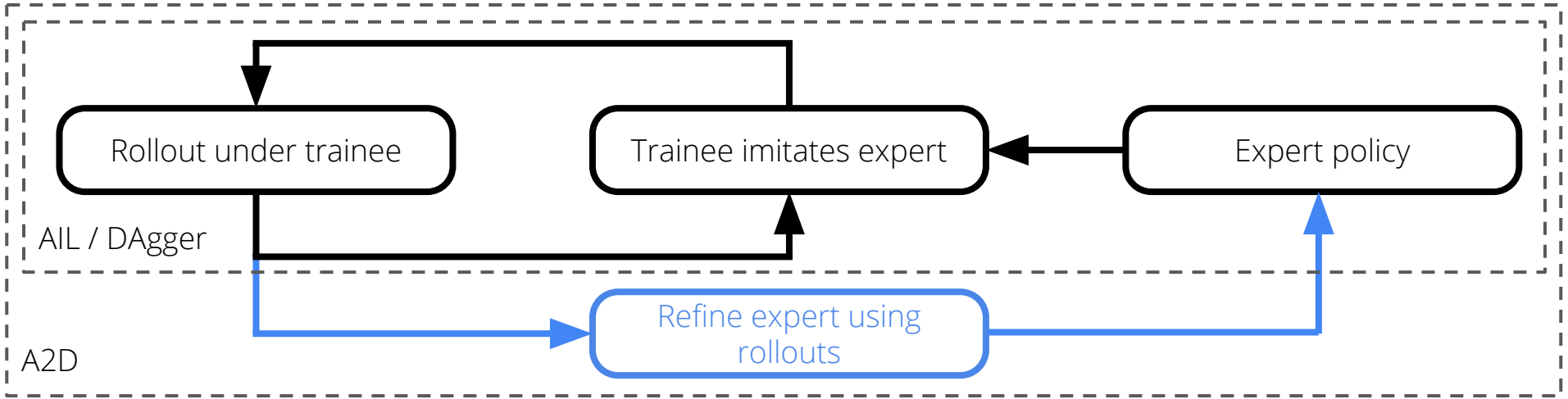
Optimal POMDP policy

≠

Expectation of optimal MDP policy over missing information

Update expert policy to provide correct supervision

Adaptive Asymmetric DAgger (A2D)



Outline

- Deriving A2D:
  - Establish asymptotic behaviour of AIL
  - Manipulate to derive expert update
  - Full A2D algorithm

- Experiments
  - Gridworld
  - Autonomous vehicles

- Conclusion

Symmetric IL

Symmetric IL

Problem definition:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathop{\mathbb{E}}_{d^{\pi_\eta}(s)} \left[\mathbb{KL} \left[\pi_\theta(a|s) || \pi_\phi(a|s) \right] \right]$$

Symmetric IL

Problem definition:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathop{\mathbb{E}}_{d^{\pi_\eta}(s)} [\mathbb{KL}[\pi_\theta(a|s)||\pi_\phi(a|s)]]$$

Symmetric IL

Problem definition:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathop{\mathbb{E}}_{d^{\pi_\eta}(s)} \left[\mathbb{KL} \left[\pi_\theta(a|s) || \pi_\phi(a|s) \right] \right]$$

Symmetric IL

Problem definition:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathop{\mathbb{E}}_{d^{\pi_\eta}(s)} \left[\mathbb{KL} \left[\pi_\theta(a|s) || \pi_\phi(a|s) \right] \right]$$

Symmetric IL

Problem definition:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathop{\mathbb{E}}_{d^{\pi_\eta}(s)} \left[\mathbb{KL} \left[\pi_\theta(a|s) || \pi_\phi(a|s) \right] \right]$$

Symmetric IL

Problem definition:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathop{\mathbb{E}}_{d^{\pi_\eta}(s)} \left[\mathbb{KL} \left[\pi_\theta(a|s) || \pi_\phi(a|s) \right] \right]$$

Solution: the expert policy

$$\pi_{\phi^*}(a|s) = \pi_\theta^\eta(a|s)$$

Symmetric IL

Problem definition:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathop{\mathbb{E}}_{d^{\pi_\eta}(s)} \left[\mathbb{KL} \left[\pi_\theta(a|s) || \pi_\phi(a|s) \right] \right]$$

Solution: the expert policy

$$\pi_{\phi^*}(a|s) = \pi_\theta^{\eta}(a|s)$$

Symmetric IL

Asymmetric IL

Problem definition:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathop{\mathbb{E}}_{d^{\pi_\eta}(s)} \left[\mathbb{KL} \left[\pi_\theta(a|s) || \pi_\phi(a|s) \right] \right]$$

Solution: the expert policy

$$\pi_{\phi^*}(a|s) = \pi_\theta^\eta(a|s)$$

Symmetric IL

Problem definition:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathop{\mathbb{E}}_{d^{\pi_\eta}(s)} \left[\mathbb{KL} \left[\pi_\theta(a|s) || \pi_\phi(a|s) \right] \right]$$

Solution: the expert policy

$$\pi_{\phi^*}(a|s) = \pi_\theta^\eta(a|s)$$

Asymmetric IL

Problem definition:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathop{\mathbb{E}}_{d^{\pi_\eta}(s,b)} \left[\mathbb{KL} \left[\pi_\theta(a|s) || \pi_\phi(a|b) \right] \right]$$

Symmetric IL

Problem definition:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathop{\mathbb{E}}_{d^{\pi_\eta}(s)} \left[\mathbb{KL} \left[\pi_\theta(a|s) || \pi_\phi(a|s) \right] \right]$$

Solution: the expert policy

$$\pi_{\phi^*}(a|s) = \pi_\theta^\eta(a|s)$$

Asymmetric IL

Problem definition:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathop{\mathbb{E}}_{d^{\pi_\eta}(s,b)} \left[\mathbb{KL} \left[\pi_\theta(a|s) || \pi_\phi(a|b) \right] \right]$$

Symmetric IL

Problem definition:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathop{\mathbb{E}}_{d^{\pi_\eta}(s)} \left[\mathbb{KL} \left[\pi_\theta(a|s) || \pi_\phi(a|s) \right] \right]$$

Solution: the expert policy

$$\pi_{\phi^*}(a|s) = \pi_\theta^\eta(a|s)$$

Asymmetric IL

Problem definition:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathop{\mathbb{E}}_{d^{\pi_\eta}(s,b)} \left[\mathbb{KL} \left[\pi_\theta(a|s) || \pi_\phi(a|b) \right] \right]$$

Symmetric IL

Problem definition:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathop{\mathbb{E}}_{d^{\pi_\eta}(s)} \left[\mathbb{KL} \left[\pi_\theta(a|s) || \pi_\phi(a|s) \right] \right]$$

Solution: the expert policy

$$\pi_{\phi^*}(a|s) = \pi_\theta^\eta(a|s)$$

Asymmetric IL

Problem definition:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathop{\mathbb{E}}_{d^{\pi_\eta}(s,b)} \left[\mathbb{KL} \left[\pi_\theta(a|s) || \pi_\phi(a|b) \right] \right]$$

Solution: the **implicit policy**

$$\pi_{\phi^*}(a|b) = \hat{\pi}_\theta^\eta(a|b) = \mathbb{E}_{d^{\pi_\eta}(s|b)} \left[\pi_\theta(a|s) \right]$$

A similar derivation is also presented by Weichs *et al.* 2020 [wei2020a]

Symmetric IL

Problem definition:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathop{\mathbb{E}}_{d^{\pi_\eta}(s)} \left[\mathbb{KL} \left[\pi_\theta(a|s) || \pi_\phi(a|s) \right] \right]$$

Solution: the expert policy

$$\pi_{\phi^*}(a|s) = \pi_\theta^\eta(a|s)$$

Asymmetric IL

Problem definition:

$$\phi^* = \arg\min_{\phi \in \Phi} \mathop{\mathbb{E}}_{d^{\pi_\eta}(s,b)} \left[\mathbb{KL} \left[\pi_\theta(a|s) || \pi_\phi(a|b) \right] \right]$$

Solution: the **implicit policy**

$$\pi_{\phi^*}(a|b) = \hat{\pi}_\theta^\eta(a|b) = \mathbb{E}_{d^{\pi_\eta}(s|b)} \left[\pi_\theta(a|s) \right]$$

Theorem 1

A similar derivation is also presented by Weichs *et al.* 2020 [wei2020a]

Asymmetric IL Algorithm

$$\psi_{k+1} = \arg\min_{\psi \in \Psi} \mathop{\mathbb{E}}_{d^{\pi_{\psi_k}}(s,b)} \left[\mathbb{KL} \left[\pi_{\theta*}(a|s) || \pi_{\psi}(a|b) \right] \right]$$

Asymmetric IL Algorithm

$$\psi_{k+1} = \arg\min_{\psi \in \Psi} \mathop{\mathbb{E}}_{d^{\pi_{\psi_k}}(s,b)} \left[\mathbb{KL} \left[\pi_{\theta^*}(a|s) || \pi_\psi(a|b) \right] \right]$$

Algorithm:

Asymmetric IL Algorithm

$$\psi_{k+1} = \arg\min_{\psi \in \Psi} \mathop{\mathbb{E}}_{d^{\pi_{\psi_k}}(s,b)} \left[\mathbb{KL} \left[\pi_{\theta*}(a|s) || \pi_{\psi}(a|b) \right] \right]$$

Algorithm:

1. Sample examples from environment.

Asymmetric IL Algorithm

$$\psi_{k+1} = \arg\min_{\psi \in \Psi} \mathop{\mathbb{E}}_{d^{\pi_{\psi_k}}(s,b)} \left[\mathbb{KL} \left[\pi_{\theta*}(a|s) || \pi_{\psi}(a|b) \right] \right]$$

Algorithm:

1. Sample examples from environment.
2. Update policy to match expert using the examples gathered.

Asymmetric IL Algorithm

$$\psi_{k+1} = \arg\min_{\psi \in \Psi} \mathop{\mathbb{E}}_{d^{\pi_{\psi_k}}(s,b)} \left[\mathbb{KL} \left[\pi_{\theta*}(a|s) || \pi_\psi(a|b) \right] \right]$$

Algorithm:

1. Sample examples from environment.
2. Update policy to match expert using the examples gathered.

Converges under ***identifiability*** (Theorem 2)

Asymmetric IL Algorithm

$$\psi_{k+1} = \arg\min_{\psi \in \Psi} \underbrace{\mathop{\mathbb{E}}_{d^{\pi_{\psi_k}}(s,b)} \left[\mathbb{KL} \left[\pi_{\theta*}(a|s) || \pi_\psi(a|b) \right] \right]}_{\text{Must go to zero (Definition 3)}}$$

Algorithm:

1. Sample examples from environment.
2. Update policy to match expert using the examples gathered.

Converges under ***identifiability*** (Theorem 2)

Asymmetric IL Algorithm

$$\psi_{k+1} = \arg\min_{\psi \in \Psi} \underbrace{\mathbb{E}_{d^{\pi_{\psi_k}}(s,b)} \left[\mathbb{KL} \left[\pi_{\theta*}(a|s) || \pi_\psi(a|b) \right] \right]}_{\text{Must go to zero (Definition 3)}}$$

How do we use the AIL solution to create better policies?

Algorithm:

1. Sample examples from environment.
2. Update policy to match expert using the examples gathered.

Converges under ***identifiability*** (Theorem 2)

Deriving the A2D Update

Maximize reward of implicit policy over expert parameters:

Deriving the A2D Update

Maximize reward of implicit policy over expert parameters:

$$\rightarrow \quad \theta^* = \arg\max_{\theta \in \Theta} \left[\mathop{\mathbb{E}}_{d^{\hat{\pi}_\theta}(b)\hat{\pi}_\theta(a|b)} \left[Q^{\hat{\pi}_\theta}(a, b) \right] \right]$$

Deriving the A2D Update

Maximize reward of implicit policy over expert parameters:

$$\rightarrow \quad \theta^* = \arg\max_{\theta \in \Theta} \left[\mathop{\mathbb{E}}_{d^{\hat{\pi}_\theta}(b)\hat{\pi}_\theta(a|b)} \left[Q^{\hat{\pi}_\theta}(a,b) \right] \right]$$

Leads to “idealized” A2D algorithm:

$$\psi_{k+1} = \arg\min_{\psi \in \Psi} \mathop{\mathbb{E}}_{d^{\pi_{\psi_k}}(s,b)} \left[\mathbb{KL} \left[\pi_{\hat{\theta}^*}(a|s) || \pi_\psi(a|b) \right] \right],$$

$$\textit{where} \quad \hat{\theta}^* = \arg\max_{\theta \in \Theta} \mathop{\mathbb{E}}_{\hat{\pi}_\theta(a|b) d^{\pi_{\psi_k}}(b)} \left[Q^{\hat{\pi}_\theta}(a,b) \right].$$

Deriving the A2D Update

Maximize reward of implicit policy over expert parameters:

$$\theta^* = \arg\max_{\theta \in \Theta} \left[\mathop{\mathbb{E}}_{d^{\hat{\pi}_\theta}(b)\hat{\pi}_\theta(a|b)} \left[Q^{\hat{\pi}_\theta}(a,b) \right] \right]$$

Replace expert with refined expert

Leads to “idealized” A2D algorithm:

$$\psi_{k+1} = \arg\min_{\psi \in \Psi} \mathop{\mathbb{E}}_{d^{\pi_{\psi_k}}(s,b)} \left[\mathbb{KL} \left[\pi_{\hat{\theta}^*}(a|s) || \pi_\psi(a|b) \right] \right],$$

$$\textit{where} \quad \hat{\theta}^* = \arg\max_{\theta \in \Theta} \mathop{\mathbb{E}}_{\hat{\pi}_\theta(a|b) d^{\pi_{\psi_k}}(b)} \left[Q^{\hat{\pi}_\theta}(a,b) \right].$$

Deriving the A2D Update

Maximize reward of implicit policy over expert parameters:

$$\theta^* = \arg\max_{\theta \in \Theta} \left[\mathop{\mathbb{E}}_{d^{\hat{\pi}_\theta}(b)\hat{\pi}_\theta(a|b)} \left[Q^{\hat{\pi}_\theta}(a,b) \right] \right]$$

Leads to “idealized” A2D algorithm:

$$\psi_{k+1} = \arg\min_{\psi \in \Psi} \mathop{\mathbb{E}}_{d^{\pi_{\psi_k}}(s,b)} \left[\mathbb{KL} \left[\pi_{\hat{\theta}^*}(a|s) || \pi_\psi(a|b) \right] \right],$$

$$\textit{where} \quad \hat{\theta}^* = \arg\max_{\theta \in \Theta} \mathop{\mathbb{E}}_{\hat{\pi}_\theta(a|b) d^{\pi_{\psi_k}}(b)} \left[Q^{\hat{\pi}_\theta}(a,b) \right].$$

Replace expert with refined expert

Improvement / RL Step

Deriving the A2D Update

Maximize reward of implicit policy over expert parameters:

$$\theta^* = \arg\max_{\theta \in \Theta} \left[\mathop{\mathbb{E}}_{d^{\hat{\pi}_\theta}(b)\hat{\pi}_\theta(a|b)} \left[Q^{\hat{\pi}_\theta}(a,b) \right] \right]$$

Leads to “idealized” A2D algorithm:

$$\psi_{k+1} = \arg\min_{\psi \in \Psi} \mathop{\mathbb{E}}_{d^{\pi_{\psi_k}}(s,b)} \left[\mathbb{KL} \left[\pi_{\hat{\theta}^*}(a|s) || \pi_\psi(a|b) \right] \right],$$

$$\textit{where} \quad \hat{\theta}^* = \arg\max_{\theta \in \Theta} \mathop{\mathbb{E}}_{\hat{\pi}_\theta(a|b) d^{\pi_{\psi_k}}(b)} \left[Q^{\hat{\pi}_\theta}(a,b) \right].$$

Replace expert with refined expert

Projection / AIL Step

Improvement / RL Step

Deriving the A2D Update

Maximize reward of implicit policy over expert parameters:

$$\theta^* = \arg\max_{\theta \in \Theta} \left[\mathop{\mathbb{E}}_{d^{\hat{\pi}_\theta}(b)\hat{\pi}_\theta(a|b)} \left[Q^{\hat{\pi}_\theta}(a,b) \right] \right]$$

Leads to "idealized" A2D algorithm:

$$\psi_{k+1} = \arg\min_{\psi \in \Psi} \mathop{\mathbb{E}}_{d^{\pi_{\psi_k}}(s,b)} \left[\mathbb{KL} \left[\pi_{\hat{\theta}^*}(a|s) || \pi_\psi(a|b) \right] \right],$$

$$\textit{where} \quad \hat{\theta}^* = \arg\max_{\theta \in \Theta} \mathop{\mathbb{E}}_{\hat{\pi}_\theta(a|b) d^{\pi_{\psi_k}}(b)} \left[Q^{\hat{\pi}_\theta}(a,b) \right].$$

Replace expert with refined expert

Projection / AIL Step

Improvement / RL Step

Converges to optimal POMDP policy under weaker conditions than AIL (Theorem 3)

A2D: Practical Considerations

Exact solution to idealized A2D iteration is intractable:

A2D: Practical Considerations

Exact solution to idealized A2D iteration is intractable.

1. Update expert parameters using lower bound on improvement:

A2D: Practical Considerations

Exact solution to idealized A2D iteration is intractable.

1. Update expert parameters using lower bound on improvement:

$$\hat{\theta}^* = \arg\max_{\theta \in \Theta} \mathop{\mathbb{E}}_{\hat{\pi}_\theta(a|b) d^{\pi_\psi}(b)} \left[Q^{\hat{\pi}_\theta}(a, b) \right] \longrightarrow \theta^* = \arg\max_{\theta \in \Theta} \mathop{\mathbb{E}}_{\hat{\pi}_\theta(a|b) d^{\pi_\psi}(b)} \left[Q^{\pi_\psi}(a, b) \right]$$

A2D: Practical Considerations

Exact solution to idealized A2D iteration is intractable.

1. Update expert parameters using lower bound on improvement:

$$\hat{\theta}^* = \arg\max_{\theta \in \Theta} \mathop{\mathbb{E}}_{\hat{\pi}_\theta(a|b) d^{\pi_\psi}(b)} \left[Q^{\hat{\pi}_\theta}(a, b) \right] \longrightarrow \theta^* = \arg\max_{\theta \in \Theta} \mathop{\mathbb{E}}_{\hat{\pi}_\theta(a|b) d^{\pi_\psi}(b)} \left[Q^{\pi_\psi}(a, b) \right]$$

A2D: Practical Considerations

Exact solution to idealized A2D iteration is intractable.

1. Update expert parameters using lower bound on improvement:

$$\hat{\theta}^* = \arg\max_{\theta \in \Theta} \mathop{\mathbb{E}}_{\hat{\pi}_\theta(a|b) d^{\pi_\psi}(b)} \left[Q^{\hat{\pi}_\theta}(a, b) \right] \longrightarrow \theta^* = \arg\max_{\theta \in \Theta} \mathop{\mathbb{E}}_{\hat{\pi}_\theta(a|b) d^{\pi_\psi}(b)} \left[Q^{\pi_\psi}(a, b) \right]$$

2. Replace direct solutions to exact AIL and RL steps with iterative steps:

A2D: Practical Considerations

Exact solution to idealized A2D iteration is intractable.

1. Update expert parameters using lower bound on improvement:

$$\hat{\theta}^* = \arg\max_{\theta \in \Theta} \mathop{\mathbb{E}}_{\hat{\pi}_\theta(a|b) d^{\pi_\psi}(b)} \left[Q^{\hat{\pi}_\theta}(a,b) \right] \longrightarrow \theta^* = \arg\max_{\theta \in \Theta} \mathop{\mathbb{E}}_{\hat{\pi}_\theta(a|b) d^{\pi_\psi}(b)} \left[Q^{\pi_\psi}(a,b) \right]$$

2. Replace direct solutions to exact AIL and RL steps with iterative steps:

$$\theta_{k+1} = \theta_k + \nu_k \mathop{\mathbb{E}}_{d^{\pi_\psi}(s,b)} \left[\mathop{\mathbb{E}}_{\pi_\theta(a|s)} \left[Q^{\pi_\psi}(a,b) \nabla_\theta \log \pi_\theta(a|s) \right] \right]$$ Improvement / RL Step

$$\psi_{k+1} = \psi_k + \nu_k \mathop{\mathbb{E}}_{d^{\pi_{\psi_k}}(s,b)} \left[\mathop{\mathbb{E}}_{\pi_\theta(a|s)} \left[\nabla_\psi \log \pi_\psi(a|b) \right] \right]$$ Projection / AIL Step

A2D Algorithm

Algorithm 1 Adaptive Asymmetric DAgger (A2D)

Input: MDP $\mathcal{M}_\Theta$, POMDP $\mathcal{M}_\Phi$, Annealing schedule $\texttt{AnnealBeta}(n, \beta)$.
Return: Variational policy parameters ψ.
$\theta, \psi, \nu_m, \nu_p, \leftarrow \texttt{InitNets}\,(\mathcal{M}_\Theta, \mathcal{M}_\Phi)$
$\beta \leftarrow 1,\ D \leftarrow \emptyset$
for $n = 0, \ldots, N$ **do**
  $\beta \leftarrow \texttt{AnnealBeta}\,(n, \beta)$
  $\pi_\beta \leftarrow \beta\pi_\theta + (1-\beta)\hat{\pi}_\psi$
  $\mathcal{T} = \{\tau_i\}_{i=1}^{\mathcal{I}} \sim q_{\pi_\beta}(\tau)$
  $D \leftarrow \texttt{UpdateBuffer}\,(D, \mathcal{T})$
  $V^{\pi_\beta} \leftarrow \beta V_{\nu_m}^{\pi_\theta} + (1-\beta) V_{\nu_p}^{\hat{\pi}_\psi}$
  $\theta, \nu_m, \nu_p \leftarrow \texttt{RLStep}\,(\mathcal{T}, V^{\pi_\beta}, \pi_\beta)$
  $\psi \leftarrow \texttt{AILStep}\,(D, \pi_\theta, \hat{\pi}_\psi)$
end for

A2D Algorithm

Algorithm 1 Adaptive Asymmetric DAgger (A2D)

Input: MDP $\mathcal{M}_\Theta$, POMDP $\mathcal{M}_\Phi$, Annealing schedule $\texttt{AnnealBeta}(n, \beta)$.
Return: Variational policy parameters ψ.
$\theta, \psi, \nu_m, \nu_p, \leftarrow \texttt{InitNets}\,(\mathcal{M}_\Theta, \mathcal{M}_\Phi)$
$\beta \leftarrow 1,\ D \leftarrow \emptyset$
for $n = 0, \ldots, N$ **do**
  $\beta \leftarrow \texttt{AnnealBeta}\,(n, \beta)$
  $\pi_\beta \leftarrow \beta\pi_\theta + (1-\beta)\hat{\pi}_\psi$
  $\mathcal{T} = \{\tau_i\}_{i=1}^{\mathcal{I}} \sim q_{\pi_\beta}(\tau)$
  $D \leftarrow \texttt{UpdateBuffer}\,(D, \mathcal{T})$
  $V^{\pi_\beta} \leftarrow \beta V_{\nu_m}^{\pi_\theta} + (1-\beta) V_{\nu_p}^{\hat{\pi}_\psi}$
  $\theta, \nu_m, \nu_p \leftarrow \texttt{RLStep}\,(\mathcal{T}, V^{\pi_\beta}, \pi_\beta)$
  $\psi \leftarrow \texttt{AILStep}\,(D, \pi_\theta, \hat{\pi}_\psi)$
end for

Blue indicates differences with original DAgger Algorithm.

A2D Algorithm

Algorithm 1 Adaptive Asymmetric DAgger (A2D)

Input: MDP $\mathcal{M}_\Theta$, POMDP $\mathcal{M}_\Phi$, Annealing schedule $\texttt{AnnealBeta}(n, \beta)$.
Return: Variational policy parameters ψ.
$\theta, \psi, \nu_m, \nu_p, \leftarrow \texttt{InitNets}\,(\mathcal{M}_\Theta, \mathcal{M}_\Phi)$
$\beta \leftarrow 1,\ D \leftarrow \emptyset$
for $n = 0, \ldots, N$ **do**
  $\beta \leftarrow \texttt{AnnealBeta}\,(n, \beta)$
  $\pi_\beta \leftarrow \beta\pi_\theta + (1-\beta)\hat{\pi}_\psi$
  $\mathcal{T} = \{\tau_i\}_{i=1}^{\mathcal{I}} \sim q_{\pi_\beta}(\tau)$
  $D \leftarrow \texttt{UpdateBuffer}\,(D, \mathcal{T})$
  $V^{\pi_\beta} \leftarrow \beta V_{\nu_m}^{\pi_\theta} + (1-\beta) V_{\nu_p}^{\hat{\pi}_\psi}$
  $\theta, \nu_m, \nu_p \leftarrow \texttt{RLStep}\,(\mathcal{T}, V^{\pi_\beta}, \pi_\beta)$
  $\psi \leftarrow \texttt{AILStep}\,(D, \pi_\theta, \hat{\pi}_\psi)$
end for

Mixing networks can accelerate convergence.

Blue indicates differences with original DAgger Algorithm.

A2D Algorithm

Additional modifications:

Algorithm 1 Adaptive Asymmetric DAgger (A2D)

Input: MDP $\mathcal{M}_\Theta$, POMDP $\mathcal{M}_\Phi$, Annealing schedule $\texttt{AnnealBeta}(n, \beta)$.
Return: Variational policy parameters ψ.
$\theta, \psi, \nu_m, \nu_p, \leftarrow \texttt{InitNets}\,(\mathcal{M}_\Theta, \mathcal{M}_\Phi)$
$\beta \leftarrow 1,\ D \leftarrow \emptyset$
for $n = 0, \ldots, N$ **do**
  $\beta \leftarrow \texttt{AnnealBeta}\,(n, \beta)$
  $\pi_\beta \leftarrow \beta\pi_\theta + (1-\beta)\hat{\pi}_\psi$
  $\mathcal{T} = \{\tau_i\}_{i=1}^{\mathcal{I}} \sim q_{\pi_\beta}(\tau)$
  $D \leftarrow \texttt{UpdateBuffer}\,(D, \mathcal{T})$
  $V^{\pi_\beta} \leftarrow \beta V_{\nu_m}^{\pi_\theta} + (1-\beta) V_{\nu_p}^{\hat{\pi}_\psi}$
  $\theta, \nu_m, \nu_p \leftarrow \texttt{RLStep}\,(\mathcal{T}, V^{\pi_\beta}, \pi_\beta)$
  $\psi \leftarrow \texttt{AILStep}\,(D, \pi_\theta, \hat{\pi}_\psi)$
end for

Mixing networks can accelerate convergence.

Blue indicates differences with original DAgger Algorithm.

A2D Algorithm

Additional modifications:

1. Control variates / baselines

Algorithm 1 Adaptive Asymmetric DAgger (A2D)

Input: MDP $\mathcal{M}_\Theta$, POMDP $\mathcal{M}_\Phi$, Annealing schedule $\texttt{AnnealBeta}(n, \beta)$.
Return: Variational policy parameters ψ.
$\theta, \psi, \nu_m, \nu_p, \leftarrow \texttt{InitNets}\,(\mathcal{M}_\Theta, \mathcal{M}_\Phi)$
$\beta \leftarrow 1,\ D \leftarrow \emptyset$
for $n = 0, \ldots, N$ **do**
  $\beta \leftarrow \texttt{AnnealBeta}\,(n, \beta)$
  $\pi_\beta \leftarrow \beta\pi_\theta + (1-\beta)\hat{\pi}_\psi$
  $\mathcal{T} = \{\tau_i\}_{i=1}^{\mathcal{I}} \sim q_{\pi_\beta}(\tau)$
  $D \leftarrow \texttt{UpdateBuffer}\,(D, \mathcal{T})$
  $V^{\pi_\beta} \leftarrow \beta V_{\nu_m}^{\pi_\theta} + (1-\beta) V_{\nu_p}^{\hat{\pi}_\psi}$
  $\theta, \nu_m, \nu_p \leftarrow \texttt{RLStep}\,(\mathcal{T}, V^{\pi_\beta}, \pi_\beta)$
  $\psi \leftarrow \texttt{AILStep}\,(D, \pi_\theta, \hat{\pi}_\psi)$
end for

Mixing networks can accelerate convergence.

Blue indicates differences with original DAgger Algorithm.

A2D Algorithm

Additional modifications:

1. Control variates / baselines
2. Generalized advantage estimation (GAE) [sch2016a]

Algorithm 1 Adaptive Asymmetric DAgger (A2D)

Input: MDP $\mathcal{M}_\Theta$, POMDP $\mathcal{M}_\Phi$, Annealing schedule $\texttt{AnnealBeta}(n, \beta)$.
Return: Variational policy parameters ψ.
$\theta, \psi, \nu_m, \nu_p, \leftarrow \texttt{InitNets}\,(\mathcal{M}_\Theta, \mathcal{M}_\Phi)$
$\beta \leftarrow 1,\ D \leftarrow \emptyset$
for $n = 0, \ldots, N$ **do**
  $\beta \leftarrow \texttt{AnnealBeta}\,(n, \beta)$
  $\pi_\beta \leftarrow \beta\pi_\theta + (1-\beta)\hat{\pi}_\psi$
  $\mathcal{T} = \{\tau_i\}_{i=1}^{\mathcal{I}} \sim q_{\pi_\beta}(\tau)$
  $D \leftarrow \texttt{UpdateBuffer}\,(D, \mathcal{T})$
  $V^{\pi_\beta} \leftarrow \beta V_m^{\pi_\theta} + (1-\beta) V_{\nu_m}^{\hat{\pi}_\psi}$
  $\theta, \nu_m, \nu_p \leftarrow \texttt{RLStep}\,(\mathcal{T}, V^{\pi_\beta}, \pi_\beta)$
  $\psi \leftarrow \texttt{AILStep}\,(D, \pi_\theta, \hat{\pi}_\psi)$
end for

Mixing networks can accelerate convergence.

Blue indicates differences with original DAgger Algorithm.

A2D Algorithm

Additional modifications:

1. Control variates / baselines
2. Generalized advantage estimation (GAE) [sch2016a]
3. Monte Carlo estimation (asymmetric case)

Algorithm 1 Adaptive Asymmetric DAgger (A2D)

Input: MDP $\mathcal{M}_\Theta$, POMDP $\mathcal{M}_\Phi$, Annealing schedule $\texttt{AnnealBeta}(n, \beta)$.
Return: Variational policy parameters ψ.
$\theta, \psi, \nu_m, \nu_p, \leftarrow \texttt{InitNets}\,(\mathcal{M}_\Theta, \mathcal{M}_\Phi)$
$\beta \leftarrow 1,\ D \leftarrow \emptyset$
for $n = 0, \ldots, N$ **do**
  $\beta \leftarrow \texttt{AnnealBeta}\,(n, \beta)$
  $\pi_\beta \leftarrow \beta\pi_\theta + (1-\beta)\hat{\pi}_\psi$
  $\mathcal{T} = \{\tau_i\}_{i=1}^{\mathcal{I}} \sim q_{\pi_\beta}(\tau)$
  $D \leftarrow \texttt{UpdateBuffer}\,(D, \mathcal{T})$
  $V^{\pi_\beta} \leftarrow \beta V_{\nu_m}^{\pi_\theta} + (1-\beta) V_{\nu_p}^{\hat{\pi}_\psi}$
  $\theta, \nu_m, \nu_p \leftarrow \texttt{RLStep}\,(\mathcal{T}, V^{\pi_\beta}, \pi_\beta)$
  $\psi \leftarrow \texttt{AILStep}\,(D, \pi_\theta, \hat{\pi}_\psi)$
end for

Mixing networks can accelerate convergence.

Blue indicates differences with original DAgger Algorithm.

A2D Algorithm

Additional modifications:

1. Control variates / baselines
2. Generalized advantage estimation (GAE) [sch2016a]
3. Monte Carlo estimation (asymmetric case)
4. Critic approximation

Algorithm 1 Adaptive Asymmetric DAgger (A2D)

Input: MDP $\mathcal{M}_\Theta$, POMDP $\mathcal{M}_\Phi$, Annealing schedule $\texttt{AnnealBeta}(n, \beta)$.
Return: Variational policy parameters ψ.
$\theta, \psi, \nu_m, \nu_p, \leftarrow \texttt{InitNets}\,(\mathcal{M}_\Theta, \mathcal{M}_\Phi)$
$\beta \leftarrow 1,\ D \leftarrow \emptyset$
for $n = 0, \ldots, N$ **do**
  $\beta \leftarrow \texttt{AnnealBeta}\,(n, \beta)$
  $\pi_\beta \leftarrow \beta\pi_\theta + (1-\beta)\hat{\pi}_\psi$
  $\mathcal{T} = \{\tau_i\}_{i=1}^{\mathcal{I}} \sim q_{\pi_\beta}(\tau)$
  $D \leftarrow \texttt{UpdateBuffer}\,(D, \mathcal{T})$
  $V^{\pi_\beta} \leftarrow \beta V_{\nu_m}^{\pi_\theta} + (1-\beta) V_{\nu_p}^{\hat{\pi}_\psi}$
  $\theta, \nu_m, \nu_p \leftarrow \texttt{RLStep}\,(\mathcal{T}, V^{\pi_\beta}, \pi_\beta)$
  $\psi \leftarrow \texttt{AILStep}\,(D, \pi_\theta, \hat{\pi}_\psi)$
end for

Mixing networks can accelerate convergence.

Blue indicates differences with original DAgger Algorithm.

A2D Algorithm

Additional modifications:

1. Control variates / baselines
2. Generalized advantage estimation (GAE) [sch2016a]
3. Monte Carlo estimation (asymmetric case)
4. Critic approximation

Algorithm 1 Adaptive Asymmetric DAgger (A2D)

Input: MDP $\mathcal{M}_\Theta$, POMDP $\mathcal{M}_\Phi$, Annealing schedule $\texttt{AnnealBeta}(n, \beta)$.
Return: Variational policy parameters ψ.
$\theta, \psi, \nu_m, \nu_p, \leftarrow \texttt{InitNets}\,(\mathcal{M}_\Theta, \mathcal{M}_\Phi)$
$\beta \leftarrow 1,\ D \leftarrow \emptyset$
for $n = 0, \ldots, N$ **do**
  $\beta \leftarrow \texttt{AnnealBeta}\,(n, \beta)$
  $\pi_\beta \leftarrow \beta\pi_\theta + (1-\beta)\hat{\pi}_\psi$
  $\mathcal{T} = \{\tau_i\}_{i=1}^{\mathcal{I}} \sim q_{\pi_\beta}(\tau)$
  $D \leftarrow \texttt{UpdateBuffer}\,(D, \mathcal{T})$
  $V^{\pi_\beta} \leftarrow \beta V_{\nu_m}^{\pi_\theta} + (1-\beta) V_{\nu_p}^{\hat{\pi}_\psi}$
  $\theta, \nu_m, \nu_p \leftarrow \texttt{RLStep}\,(\mathcal{T}, V^{\pi_\beta}, \pi_\beta)$
  $\psi \leftarrow \texttt{AILStep}\,(D, \pi_\theta, \hat{\pi}_\psi)$
end for

Mixing networks can accelerate convergence.

Blue indicates differences with original DAgger Algorithm.

A2D Algorithm

Additional modifications:

1. Control variates / baselines
2. Generalized advantage estimation (GAE) [sch2016a]
3. Monte Carlo estimation (asymmetric case)
4. Critic approximation

Algorithm 1 Adaptive Asymmetric DAgger (A2D)

Input: MDP $\mathcal{M}_\Theta$, POMDP $\mathcal{M}_\Phi$, Annealing schedule $\texttt{AnnealBeta}(n, \beta)$.
Return: Variational policy parameters ψ.
$\theta, \psi, \nu_m, \nu_p, \leftarrow \texttt{InitNets}\,(\mathcal{M}_\Theta, \mathcal{M}_\Phi)$
$\beta \leftarrow 1,\ D \leftarrow \emptyset$
for $n = 0, \ldots, N$ **do**
  $\beta \leftarrow \texttt{AnnealBeta}\,(n, \beta)$
  $\pi_\beta \leftarrow \beta\pi_\theta + (1-\beta)\hat{\pi}_\psi$
  $\mathcal{T} = \{\tau_i\}_{i=1}^{\mathcal{I}} \sim q_{\pi_\beta}(\tau)$
  $D \leftarrow \texttt{UpdateBuffer}\,(D, \mathcal{T})$
  $V^{\pi_\beta} \leftarrow \beta V_{\nu_m}^{\pi_\theta} + (1-\beta) V_{\nu_p}^{\hat{\pi}_\psi}$
  $\theta, \nu_m, \nu_p \leftarrow \texttt{RLStep}\,(\mathcal{T}, V^{\pi_\beta}, \pi_\beta)$
  $\psi \leftarrow \texttt{AILStep}\,(D, \pi_\theta, \hat{\pi}_\psi)$
end for

Mixing networks can accelerate convergence.

Blue indicates differences with original DAgger Algorithm.

A2D Algorithm

Additional modifications:

1. Control variates / baselines
2. Generalized advantage estimation (GAE) [sch2016a]
3. Monte Carlo estimation (asymmetric case)
4. Critic approximation

Step 1: Rollout under trainee

Algorithm 1 Adaptive Asymmetric DAgger (A2D)

Input: MDP $\mathcal{M}_\Theta$, POMDP $\mathcal{M}_\Phi$, Annealing schedule $\texttt{AnnealBeta}(n, \beta)$.
Return: Variational policy parameters ψ.
$\theta, \psi, \nu_m, \nu_p, \leftarrow \texttt{InitNets}\,(\mathcal{M}_\Theta, \mathcal{M}_\Phi)$
$\beta \leftarrow 1,\ D \leftarrow \emptyset$
for $n = 0, \ldots, N$ **do**
  $\beta \leftarrow \texttt{AnnealBeta}\,(n, \beta)$
  $\pi_\beta \leftarrow \beta\pi_\theta + (1-\beta)\hat{\pi}_\psi$
  $\mathcal{T} = \{\tau_i\}_{i=1}^{\mathcal{I}} \sim q_{\pi_\beta}(\tau)$
  $D \leftarrow \texttt{UpdateBuffer}\,(D, \mathcal{T})$
  $V^{\pi_\beta} \leftarrow \beta V_{\nu_m}^{\pi_\theta} + (1-\beta) V_{\nu_p}^{\hat{\pi}_\psi}$
  $\theta, \nu_m, \nu_p \leftarrow \texttt{RLStep}\,(\mathcal{T}, V^{\pi_\beta}, \pi_\beta)$
  $\psi \leftarrow \texttt{AILStep}\,(D, \pi_\theta, \hat{\pi}_\psi)$
end for

Mixing networks can accelerate convergence.

Blue indicates differences with original DAgger Algorithm.

A2D Algorithm

Additional modifications:

1. Control variates / baselines
2. Generalized advantage estimation (GAE) [sch2016a]
3. Monte Carlo estimation (asymmetric case)
4. Critic approximation

Step 1: Rollout under trainee

Step 2: A2D gradient step

Algorithm 1 Adaptive Asymmetric DAgger (A2D)

Input: MDP $\mathcal{M}_\Theta$, POMDP $\mathcal{M}_\Phi$, Annealing schedule $\texttt{AnnealBeta}(n, \beta)$.
Return: Variational policy parameters ψ.
$\theta, \psi, \nu_m, \nu_p, \leftarrow \texttt{InitNets}\,(\mathcal{M}_\Theta, \mathcal{M}_\Phi)$
$\beta \leftarrow 1,\ D \leftarrow \emptyset$
for $n = 0, \ldots, N$ **do**
  $\beta \leftarrow \texttt{AnnealBeta}\,(n, \beta)$
  $\pi_\beta \leftarrow \beta\pi_\theta + (1-\beta)\hat{\pi}_\psi$
  $\mathcal{T} = \{\tau_i\}_{i=1}^{\mathcal{I}} \sim q_{\pi_\beta}(\tau)$
  $D \leftarrow \texttt{UpdateBuffer}\,(D, \mathcal{T})$
  $V^{\pi_\beta} \leftarrow \beta V_{\nu_m}^{\pi_\theta} + (1-\beta) V_{\nu_p}^{\hat{\pi}_\psi}$
  $\theta, \nu_m, \nu_p \leftarrow \texttt{RLStep}\,(\mathcal{T}, V^{\pi_\beta}, \pi_\beta)$
  $\psi \leftarrow \texttt{AILStep}\,(D, \pi_\theta, \hat{\pi}_\psi)$
end for

Mixing networks can accelerate convergence.

Blue indicates differences with original DAgger Algorithm.

A2D Algorithm

Additional modifications:

1. Control variates / baselines
2. Generalized advantage estimation (GAE) [sch2016a]
3. Monte Carlo estimation (asymmetric case)
4. Critic approximation

Algorithm 1 Adaptive Asymmetric DAgger (A2D)

Input: MDP $\mathcal{M}_\Theta$, POMDP $\mathcal{M}_\Phi$, Annealing schedule $\texttt{AnnealBeta}(n, \beta)$.
Return: Variational policy parameters ψ.
$\theta, \psi, \nu_m, \nu_p, \leftarrow \texttt{InitNets}\,(\mathcal{M}_\Theta, \mathcal{M}_\Phi)$
$\beta \leftarrow 1,\ D \leftarrow \emptyset$
for $n = 0, \ldots, N$ **do**
  $\beta \leftarrow \texttt{AnnealBeta}\,(n, \beta)$
  $\pi_\beta \leftarrow \beta\pi_\theta + (1-\beta)\hat{\pi}_\psi$
  $\mathcal{T} = \{\tau_i\}_{i=1}^{\mathcal{I}} \sim q_{\pi_\beta}(\tau)$
  $D \leftarrow \texttt{UpdateBuffer}\,(D, \mathcal{T})$
  $V^{\pi_\beta} \leftarrow \beta V_{\nu_m}^{\pi_\theta} + (1-\beta) V_{\nu_p}^{\hat{\pi}_\psi}$
  $\theta, \nu_m, \nu_p \leftarrow \texttt{RLStep}\,(\mathcal{T}, V^{\pi_\beta}, \pi_\beta)$
  $\psi \leftarrow \texttt{AILStep}\,(D, \pi_\theta, \hat{\pi}_\psi)$
end for

Step 1: Rollout under trainee

Step 2: A2D gradient step

Step 3: AIL update step

Mixing networks can accelerate convergence.

Blue indicates differences with original DAgger Algorithm.

A2D Algorithm

Additional modifications:

1. Control variates / baselines
2. Generalized advantage estimation (GAE) [sch2016a]
3. Monte Carlo estimation (asymmetric case)
4. Critic approximation

Step 1: Rollout under trainee

Step 2: A2D gradient step

Step 3: AIL update step

Step 4: Repeat to convergence

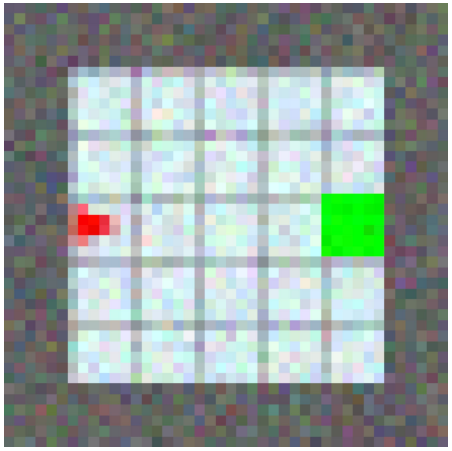
Algorithm 1 Adaptive Asymmetric DAgger (A2D)

Input: MDP $\mathcal{M}_\Theta$, POMDP $\mathcal{M}_\Phi$, Annealing schedule $\texttt{AnnealBeta}(n, \beta)$.
Return: Variational policy parameters ψ.
$\theta, \psi, \nu_m, \nu_p, \leftarrow \texttt{InitNets}\,(\mathcal{M}_\Theta, \mathcal{M}_\Phi)$
$\beta \leftarrow 1,\ D \leftarrow \emptyset$
for $n = 0, \ldots, N$ **do**
  $\beta \leftarrow \texttt{AnnealBeta}\,(n, \beta)$
  $\pi_\beta \leftarrow \beta\pi_\theta + (1-\beta)\hat{\pi}_\psi$
  $\mathcal{T} = \{\tau_i\}_{i=1}^{\mathcal{I}} \sim q_{\pi_\beta}(\tau)$
  $D \leftarrow \texttt{UpdateBuffer}\,(D, \mathcal{T})$
  $V^{\pi_\beta} \leftarrow \beta V_{\nu_m}^{\pi_\theta} + (1-\beta) V_{\nu_p}^{\hat{\pi}_\psi}$
  $\theta, \nu_m, \nu_p \leftarrow \texttt{RLStep}\,(\mathcal{T}, V^{\pi_\beta}, \pi_\beta)$
  $\psi \leftarrow \texttt{AILStep}\,(D, \pi_\theta, \hat{\pi}_\psi)$
end for

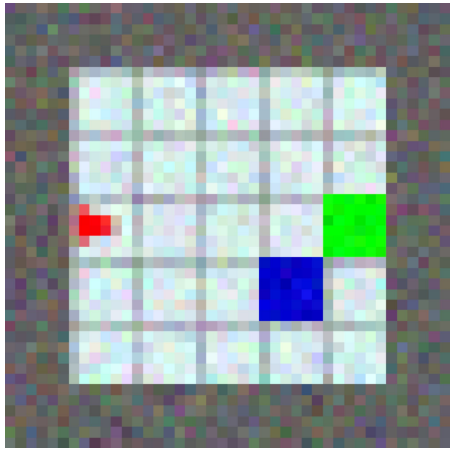
Mixing networks can accelerate convergence.

Blue indicates differences with original DAgger Algorithm.

Gridworld

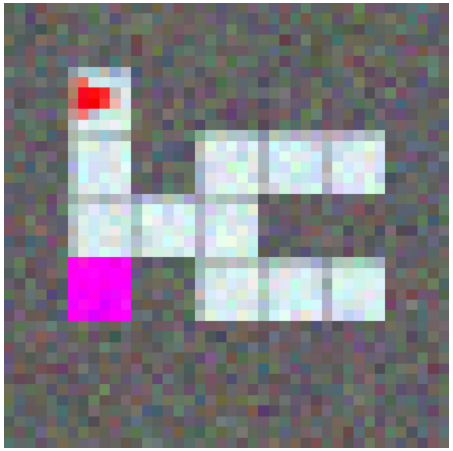


Agent

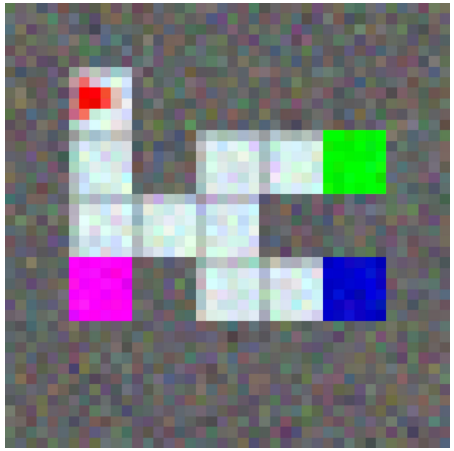


State

Frozen Lake



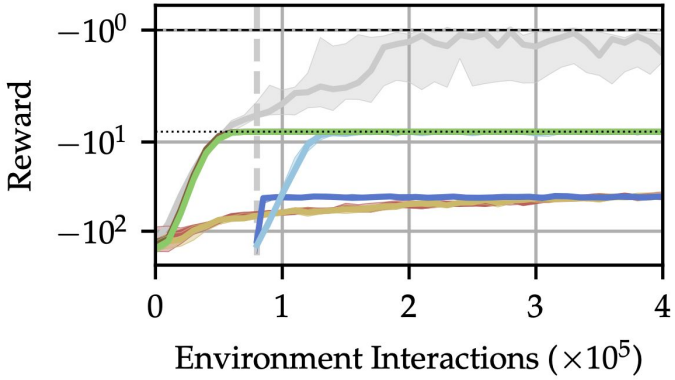
Agent (/ before button)



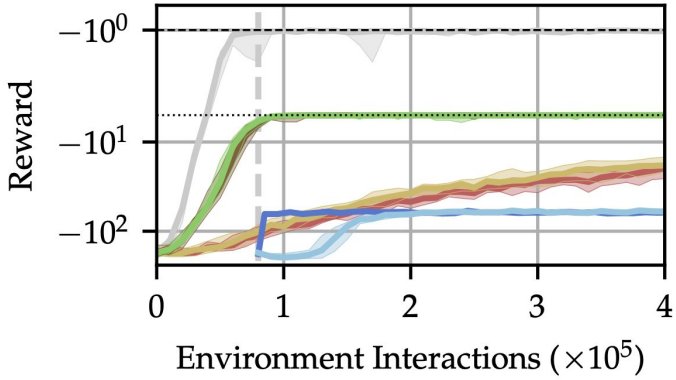
State (/ after button)

Tiger Door

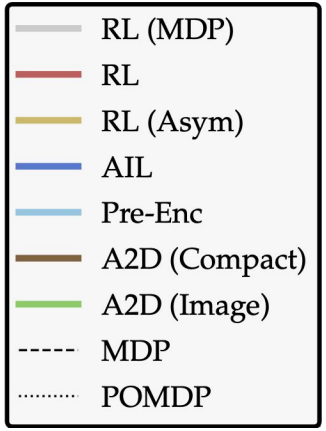
Gridworld



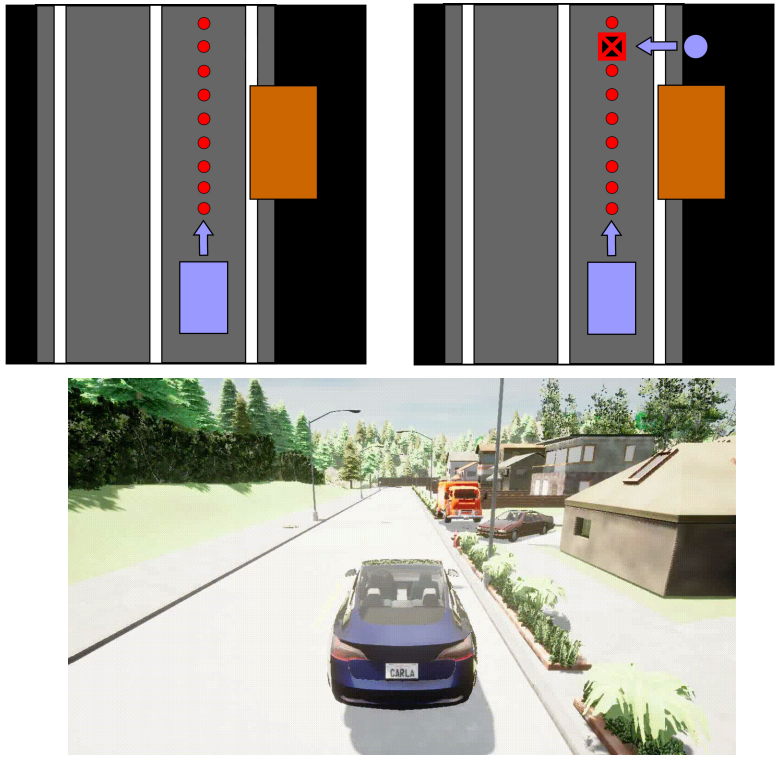
Frozen Lake



Tiger Door



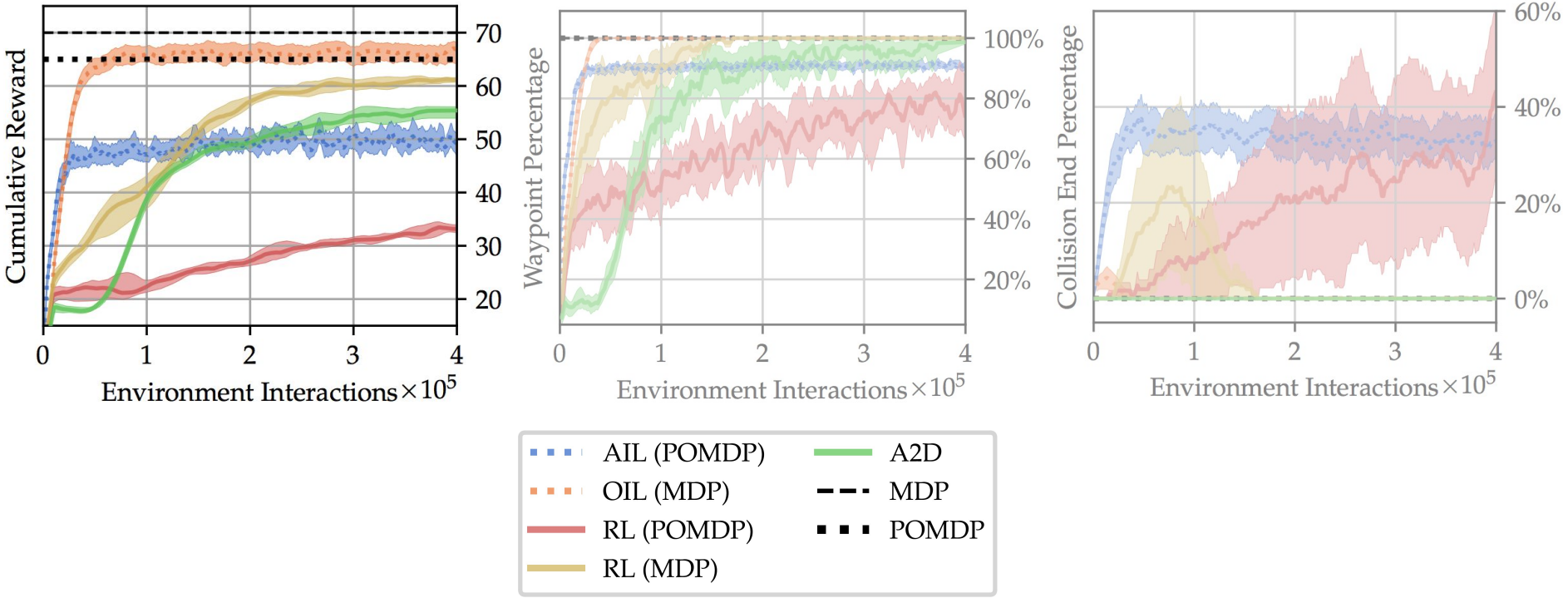
Autonomous Vehicles Scenario



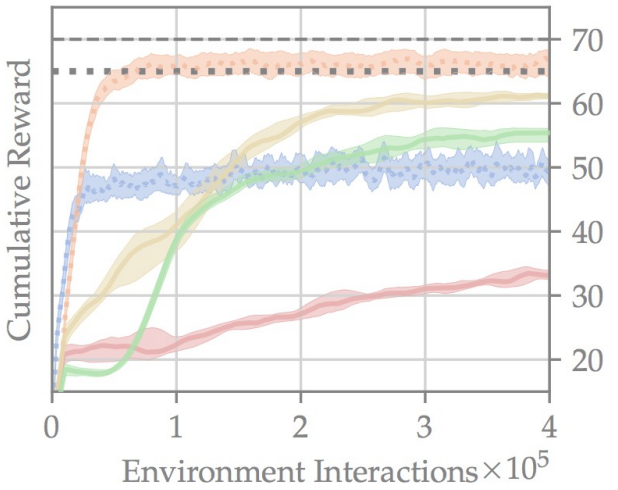
Autonomous Vehicles Scenario: Inputs

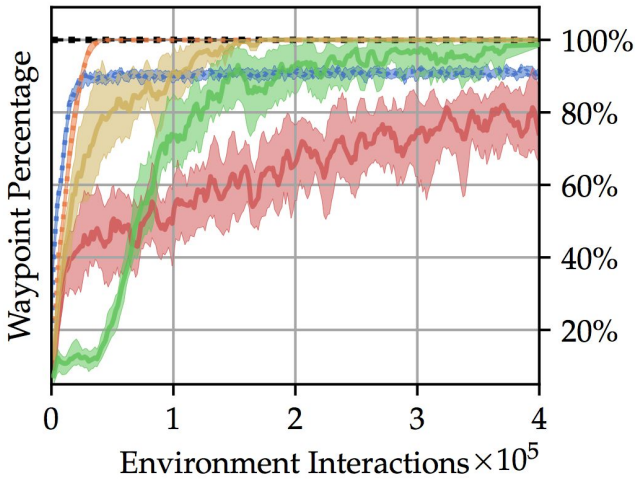
$$\begin{vmatrix} x_0 & x_1 & x_2 \\ y_0 & y_1 & y_2 \\ z_0 & z_1 & z_2 \end{vmatrix}$$

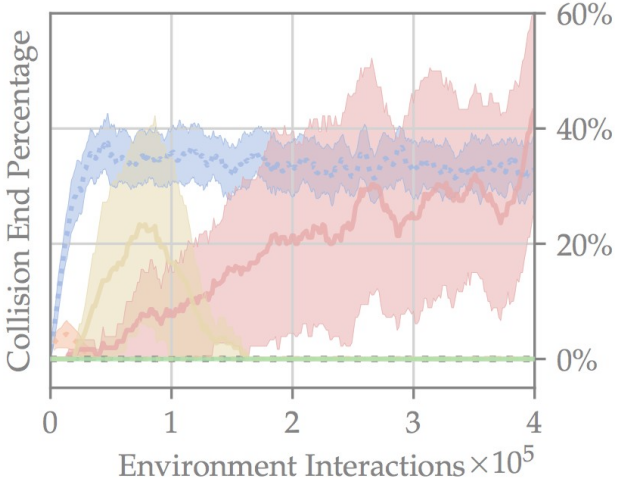
Autonomous Vehicles Results



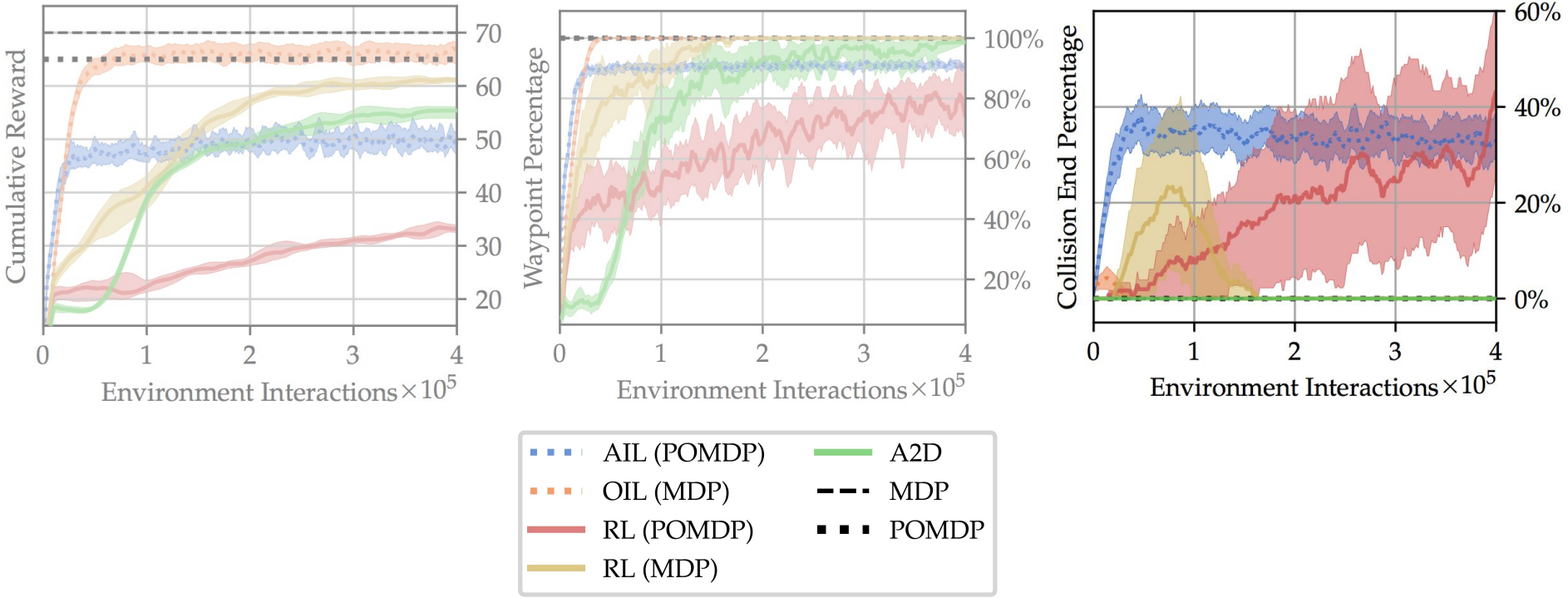
Autonomous Vehicles







Autonomous Vehicles



Autonomous Vehicles

Child

No child

RL (MDP)

OIL (MDP)

Autonomous Vehicles

Child



No child



RL (POMDP) AIL (POMDP) A2D

Summary

- Imitation learning and asymmetric information are powerful tools
- Using both can lead to unsafe policies
- We propose *adaptive asymmetric DAgger* (A2D)

Summary

- Imitation learning and asymmetric information are powerful tools
- Using both can lead to unsafe policies
- We propose *adaptive asymmetric DAgger* (A2D)

Summary

- Imitation learning and asymmetric information are powerful tools
- Using both can lead to unsafe policies
- We propose *adaptive asymmetric DAgger* (A2D)

Summary

- Imitation learning and asymmetric information are powerful tools
- Using both can lead to unsafe policies
- We propose *adaptive asymmetric DAgger* (A2D)

References & Resources

Our paper:

[war21a] Warrington, A.*, Lavington, J. W.*, Ścibior, A., Schmidt, M., & Wood, F. (2021). Robust Asymmetric Learning in POMDPs. *To appear in International Conference on Machine Learning 2021, arXiv preprint arXiv:2012.15566*. (paper)

Related Work:

[mur2000a] Murphy, K. P. (2000). A survey of POMDP solution techniques. *Environment, 2:X3*. (paper)

[pin2017a] Pinto, L., Andrychowicz, M., Welinder, P., Zaremba, W., & Abbeel, P. (2017). Asymmetric actor critic for image-based robot learning. *Robotics: Science and Systems XIV*. (paper)

[ros2011a] Ross, S., Gordon, G., & Bagnell, D. (2011). A reduction of imitation learning and structured prediction to no-regret online learning. *AISTATS. JMLR*. (paper)

[sch2015a] Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). Trust region policy optimization. *International conference on machine learning. PMLR*. (paper)

[sch2016a] Schulman, J., Moritz, P., Levine, S., Jordan, M., & Abbeel, P. (2016). High-Dimensional Continuous Control Using Generalized Advantage Estimation. *International Conference on Learning Representations (ICLR)*. (paper)

[wei2020a] Weihs, L., Jain, U., Salvador, J., Lazebnik, S., Kembhavi, A., & Schwing, A. (2020). Bridging the imitation gap by adaptive insubordination. *arXiv preprint arXiv:2007.12173*. (paper)

Paper: https://arxiv.org/pdf/2012.15566.pdf

Code, poster, talk slides and additional materials: https://github.com/plai-group/a2d

Thank you for listening!

Come talk to us live at the conference!