Group Distributionally Robust Reinforcement Learning with Hierarchical Latent Variables

Mengdi Xu, Peide Huang, Yaru Niu, Visak Kumar, Jielin Qiu, Chao Fang, Kuan-Hui Lee, Xuewei Qi, Henry Lam, Bo Li, Ding Zhao.

Contact: mengdixu@andrew.cmu.edu

Scan for full paper!

Motivation

How to design RL agents that can handle task estimate uncertainties while balancing robustness and performance?

Challenges:

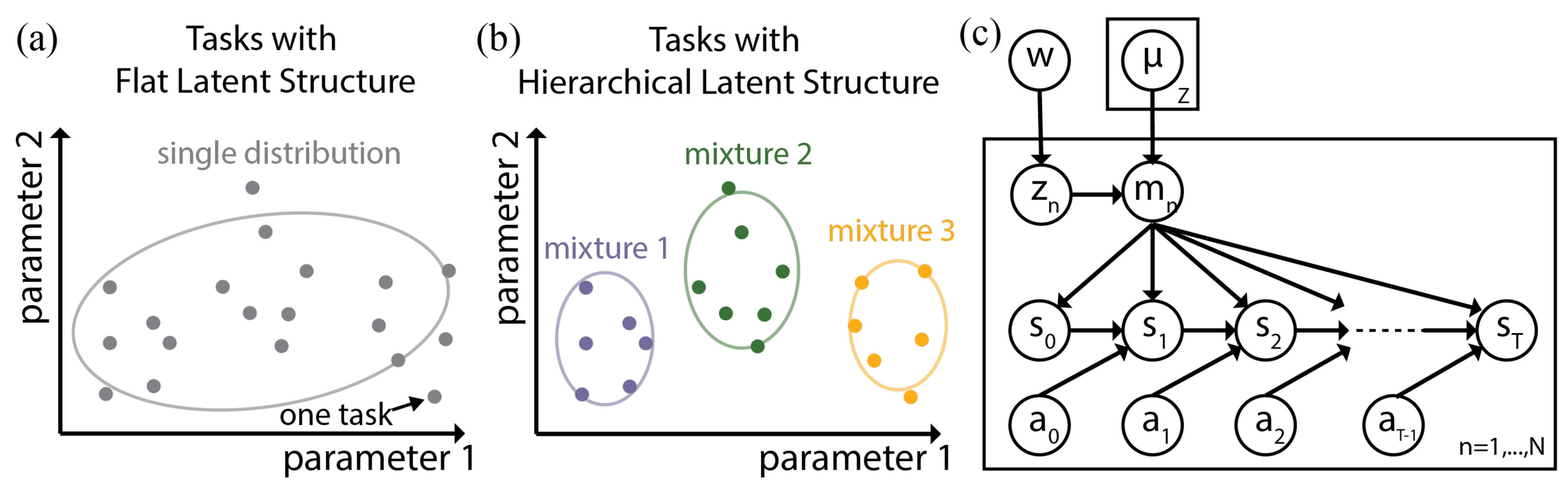
- Reinforcement Learning (RL) agents may only have incomplete information about tasks to solve.
- Robust RL that optimizes over worst-possible tasks, which may generate overly conservative policies.
- Most sequential decision-making formulations assume tasks are i.i.d. sampled from a single distribution and overlook the existence of task subpopulations.

Related subcommunities:

- Distributionally robust optimization.
- Partially observable Markov Decision Processes (MDP).

Group Distributionally Robust MDP

Group distributionally robust formulation + Model task subpopulations



Preliminary: Latent MDP

- An episodic Latent MDP can be specified by a tuple (M,T,S,A,μ). T is the episodic length. S and A are the joint state and action spaces. M is a set of joint MDPs. Each MDP is a tuple (T,S,A,P,R,v), where P and R are the transition probability and reward function. v is the initial distribution.

Our non-robust formulation: Hierarchical Latent MDP

- An episodic Hierarchical Latent MDP can be specified by a tuple (Z,M, T,S,A,w), where Z is a set of Latent MDPs and w is the categorical distribution over Latent MDPs.

Our robust formulation: Group Distributionally Robust MDP

- An episodic GDR-MDP is defined by an 8-tuple (C,Z,M,T,S,A,w,SE). C is the belief ambiguity set. SE is the belief updating rule.
- GDR-MDP maintains a belief over the mixture z and aims to find a history-dependent policy that obtains the optimal value as:

$$V^{\star} = \max_{\pi \in \Pi} \min_{\substack{\hat{b}_{0:T} \\ \in \mathcal{C}_{\triangle Z-1}}} \mathbb{E}_{\hat{b}_{0:T}(z)} \mathbb{E}_{\mu_z(m)} \mathbb{E}^{\pi}_{m} \Big[\sum_{t=1}^{T} \gamma^t r_t \Big]$$

Properties of GDR-MDP

Convergence in infinite-horizon case

- Take the sufficient statistics as (b,s) where b is the belief distribution.
- The Bellman expectation equation, Bellman optimality equation exist.
- The contraction operator exists.
- Convergence exists in infinite-horizon case.

Robustness guarantee: The benefit of (1) distributionally robust formulation and (2) the hierarchical structure

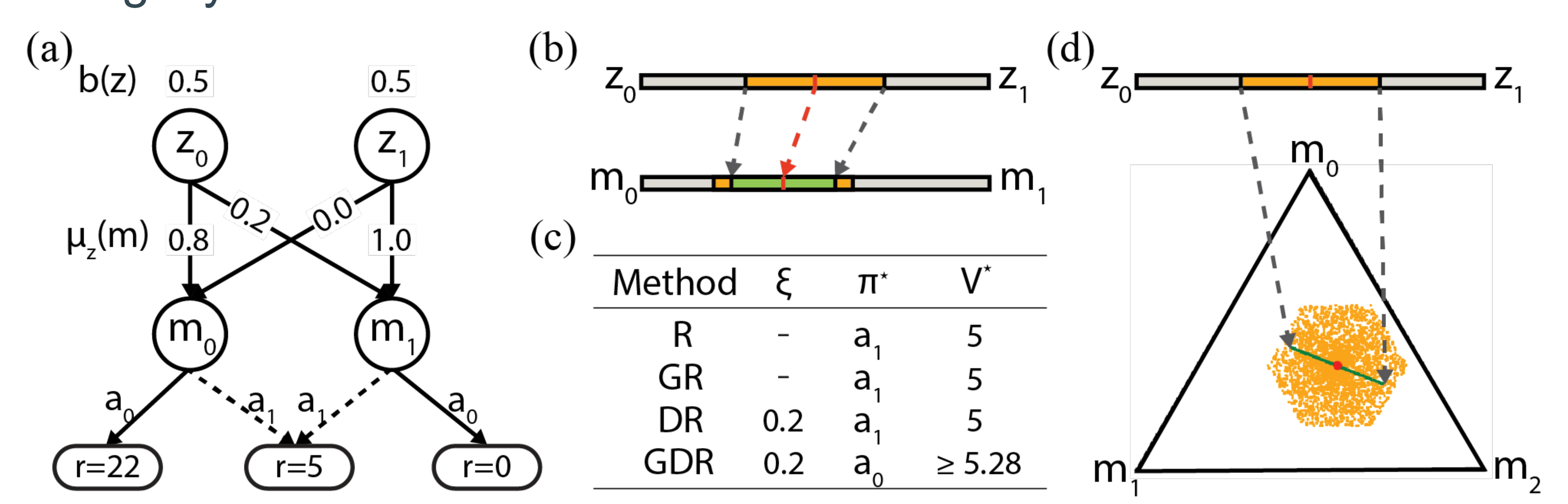
- We compare the optimal values of different robust formulations:

$$V_{GDR}(\pi) = \min_{\hat{b}(z) \in \mathcal{C}_{b(z), d_{TV}, \xi}(z)} \mathbb{E}_{\hat{b}(z)} \mathbb{E}_{\mu_z(m)} [U_m(\pi)], \qquad V_{GR}(\pi) = \min_{z \in [Z]} \mathbb{E}_{\mu_z(m)} [U_m(\pi)],$$

$$V_{DR}(\pi) = \min_{\hat{b}(m) \in \mathcal{C}_{b(m), d_{TV}, \xi}(m)} \mathbb{E}_{\hat{b}(m)} [U_m(\pi)], \qquad V_{R}(\pi) = \min_{m \in [M]} [U_m(\pi)].$$

We have the following inequalities hold: $V_{GDR}(\pi) \geq V_{GR}(\pi) \geq V_R(\pi)$ *and* $V_{GDR}(\pi) \geq V_{DR}(\pi)$.

- We achieve the comparison by studying the relationships between ambiguity sets.



(c)

Method	ξ	π*	V*
R	-	a_1	5
GR	-	a_1	5
DR	0.2	a_1	5
GDR	0.2	a_0	≥ 5.28

Our Algorithms

Algorithm 1: GDR-MDP Trajectory Rollout

Input: Mixing weights $w(z)$ and $\mu_z(m)$, episode index n, episode length T, belief update function SE, rollout policy $\pi_\theta(b(z), s)$, exploration ϵ

Initialize episodic history $h = \{\}$;

Sample mixture $z_n \sim w(z)$;

Sample MDP $m_n \sim \mu_{z_n}(m)$;

Initialize belief $b_0(z)$ as a uniform distribution ;

for $t = 0$ ***to*** T **do**

  Sample action a_t with the ϵ-greedy method and rollout in MDP m;

  $b_{t+1}(z) = SE(b_t(z), s_{t+1})$;

  Append the most recent data pair $d = \{(b_t, s_t), a_t, r_t, (b_{t+1}, s_{t+1})\}$ to h ;

Return: history h, episode return

Algorithm 2: Group Distributionally Robust Training for GDR-DQN and GDR-SAC

Input: Q-net $Q_\theta(b(z), s, a)$, ambiguity set $\mathcal{C}_{\cdot, d_{TV}, \xi}$, training episodes N,

Initialize data buffer $\mathcal{D}$;

for $n = 0$ ***to*** N **do**

  Rollout one episode with Algorithm 1 and append data pairs to $\mathcal{D}$;

  if *Update Q-net parameters* **then**

    Sample batch data from $\mathcal{D}$;

    for *Each* d_i *in the batch* **do**

      Get $b^{adv} \in \mathcal{C}_{b'(z), d_{TV}, \xi}$ with modified FGSM;

    Update Q-net $\theta \leftarrow \theta - \alpha_\theta \nabla_\theta \mathcal{L}_{Q_\theta}$;

Return: Q-net Q_θ

Group distributionally robust training methods:

- **GDR-DQN, GDR-SAC:** Robust value-based

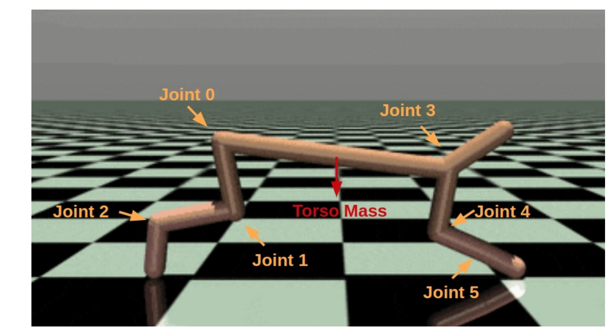
$$\mathcal{L}_{Q_\theta} = \sum_d \Big(r + \min_{p(z) \in \mathcal{C}_{b'(z), d_{TV}, \xi}} \sum_{a \in \mathcal{A}} Q_\theta(p(z), s', a) - Q_\theta(b(z), s, a) \Big)^2$$

- **GDR-PPO:** Robust policy-based

$$\hat{A}(b_t, s_t) = \sum_{t'=t}^{T-1} r_t - R_{drop} - V_\theta(b_t, s_t), \text{ where } R_{drop} = V(b_t, s_t) - \min_{p(z) \in \mathcal{C}_{b_t(z), d_{TV}, \xi}} V_\theta(p(z), s_t).$$

Experiments

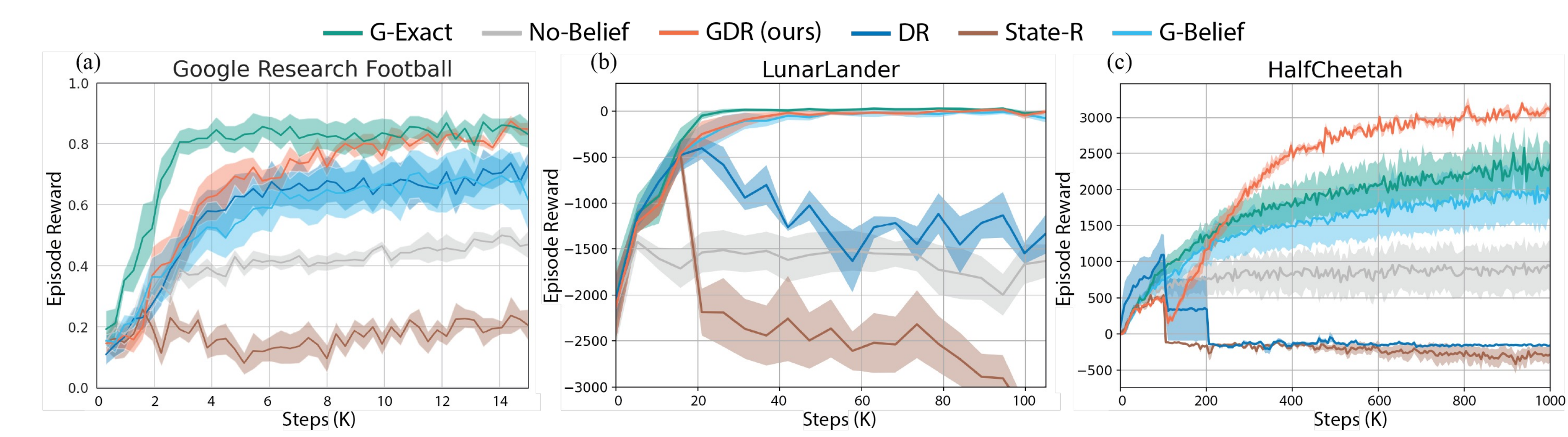
Environments



Environment	GRF (3 vs. 2)	LunarLander	HalfCheetah
Parameter 1 (Mixture)	Player Type {CM vs. CB, CB vs. CM}	Engine Mode {Normal, Flipped}	Failure Joint {0,1,2,3,4,5}
Parameter 2	Player Capability Level {0.9 vs. 0.6, 1.0 vs. 0.7}	Engine Power {3.0, 6.0}	Torso Mass {0.9, 1.0, 1.1}
# Mixtures	2	2	6
w	[0.5, 0.5]	[0.5, 0.5]	$\frac{1}{6}\mathbf{1}^6$
# MDPs	4	4	18
$\mu_z(m)$	$\begin{bmatrix} \frac{1}{2}\mathbf{1}^2 & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{1}^2 \end{bmatrix}$	$\begin{bmatrix} \frac{1}{2}\mathbf{1}^2 & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{1}^2 \end{bmatrix}$	$\begin{bmatrix} E_0(6) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & E_5(6) \end{bmatrix}$

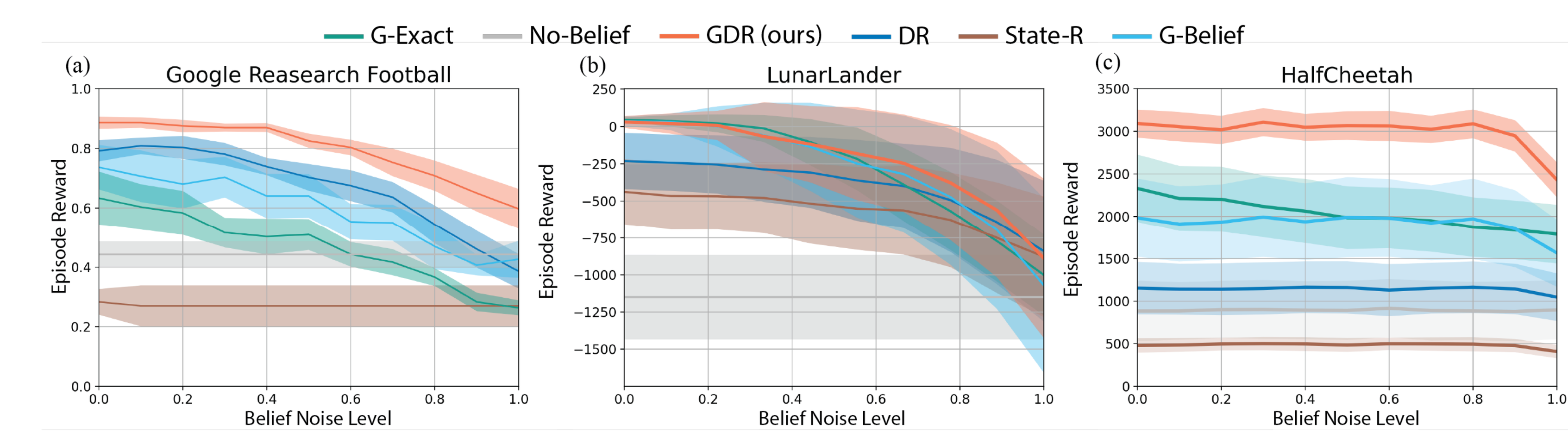
Training Stability

- GDR achieves a higher average return at convergence compared with other robust training baselines, including DR and State-R in all envs.
- DR which maintains a belief b(m) over MDPs induces significant training instability, instead of learning a meaningful conservative policy.



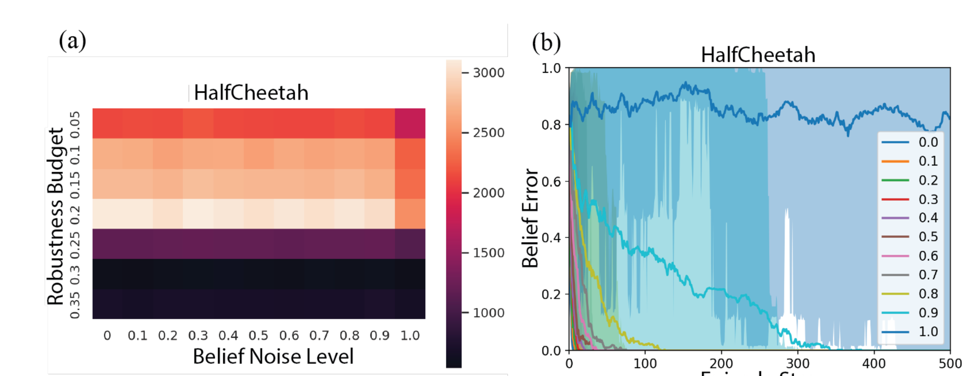
Robustness to belief noise

- In HalfCheetah and Google Research Football, GDR is consistently more robust to belief noise than baselines.



Ablation Study

- For GDR, gradually increasing the ambiguity set size up to 0.2 helps improve the robustness.
- With set of size 0.2 and pretraining for 500000 steps, DR without the mixture information still causes unstable training.



Reference:

Kwon, Jeongyeol, et al. "RL for latent MDPs: Regret guarantees and a lower bound." Advances in Neural Information Processing Systems 34 (2021).