On the Design of Variational RL Algorithms

Joseph Marino[1], Alexandre Piché[2], Yisong Yue[1]
[1]California Institute of Technology, [2]Mila, Université de Montréal

Contributions:
- Discuss variational EM framework containing several recent algorithms.
- Improve soft actor-critic (SAC) using iterative amortized inference.

variational RL

Set-Up

Markov Decision Process

state	$\mathbf{s}_t \in \mathcal{S}$	dynamics	$p_{\text{env}}(\mathbf{s}_{t+1}\vert\mathbf{s}_t, \mathbf{a}_t)$
action	$\mathbf{a}_t \in \mathcal{A}$	agent	$p_\theta(\mathbf{a}_t\vert\mathbf{s}_t)$
reward	$r_t = r(\mathbf{s}_t, \mathbf{a}_t)$	return	$\mathcal{R}(\tau) = \sum_t \gamma^t r_t$
trajectory	$\tau = (\mathbf{s}_1, \mathbf{a}_1, \dots)$		

trajectory dist. $p(\tau) = \rho(\mathbf{s}_1) \prod_{t=1}^{T} p_{\text{env}}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) p_\theta(\mathbf{a}_t|\mathbf{s}_t)$

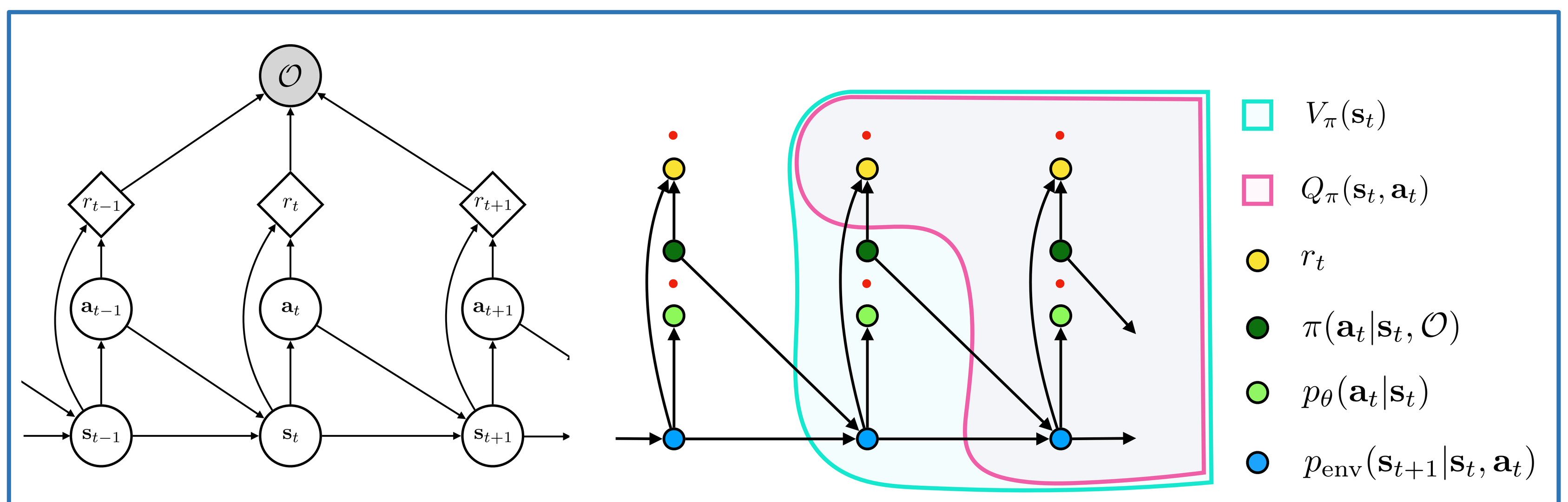
Variational RL

optimality $p(\mathcal{O} = 1|\tau) \propto \exp\left(\mathcal{R}(\tau)/\alpha\right)$

approx. posterior $\pi(\tau|\mathcal{O}) = \rho(\mathbf{s}_1) \prod_{t=1}^{T} p_{\text{env}}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) \pi(\mathbf{a}_t|\mathbf{s}_t, \mathcal{O})$

variational lower bound $\mathcal{J}(\pi, \theta) = \mathbb{E}_{\substack{\mathbf{s}_t, r_t \sim p_{\text{env}} \\ \mathbf{a}_t \sim \pi}} \left[\sum_{t=1}^{T} r_t - \alpha \log \frac{\pi(\mathbf{a}_t|\mathbf{s}_t, \mathcal{O})}{p_\theta(\mathbf{a}_t|\mathbf{s}_t)} \right]$

optimal policy (Boltzmann), *non-parametric* $\pi^{\mathcal{B}}(\mathbf{a}_t|\mathbf{s}_t, \mathcal{O}) = \frac{p_\theta(\mathbf{a}_t|\mathbf{s}_t) \exp(Q_\pi(\mathbf{s}_t, \mathbf{a}_t)/\alpha)}{\int p_\theta(\mathbf{a}_t|\mathbf{s}_t) \exp(Q_\pi(\mathbf{s}_t, \mathbf{a}_t)/\alpha) d\mathbf{a}_t}$



Graphical Model *Computation Graph*

Variational EM for RL

E-Step: infer/approximate the posterior, $\pi(\tau|\mathcal{O})$
- *non-parametric:* estimate $\pi^{\mathcal{B}}(\tau|\mathcal{O})$
- *parametric:* maximize $\mathcal{J}$ w.r.t. $\pi(\tau|\mathcal{O})$

M-Step: learn the prior, $p(\tau)$, by maximizing $\mathcal{J}$ w.r.t. θ

Prior		Conditional Likelihood	Approximate Posterior: Parametric	Approximate Posterior: Non-Parametric
	Uniform	Model-Free	*SAC* [Haarnoja et al., 2018], [Schulman et al., 2017]	*SQL* [Haarnoja et al., 2017]
		Model-Based	-	*SMCP* [Piché et al., 2017]
	Learned	Model-Free	*MPO* [Abdolmaleki et al., 2018a], [Galashov et al., 2019], [Tirumala et al., 2019]	*MPO* [Abdolmaleki et al., 2018a], *RERPI* [Abdolmaleki et al., 2018b], *V-MPO* [Song et al., 2019], *Distral* [Teh et al., 2017]
		Model-Based	-	-

3 Main Design Choices
- Prior (Uniform / Learned)
- Approximate Posterior (Parametric / Non-Parametric)
- Conditional Likelihood Estimate (Model-Free / Model-Based)

SAC + iterative inference

Soft actor-critic (SAC) [Haarnoja et al., 2018] is a state-of-the-art off-policy deep RL algorithm. We replace the ***direct*** amortized inference procedure in SAC with an ***iterative*** amortized inference procedure [Marino et al., 2018].

Policy Parameterization:

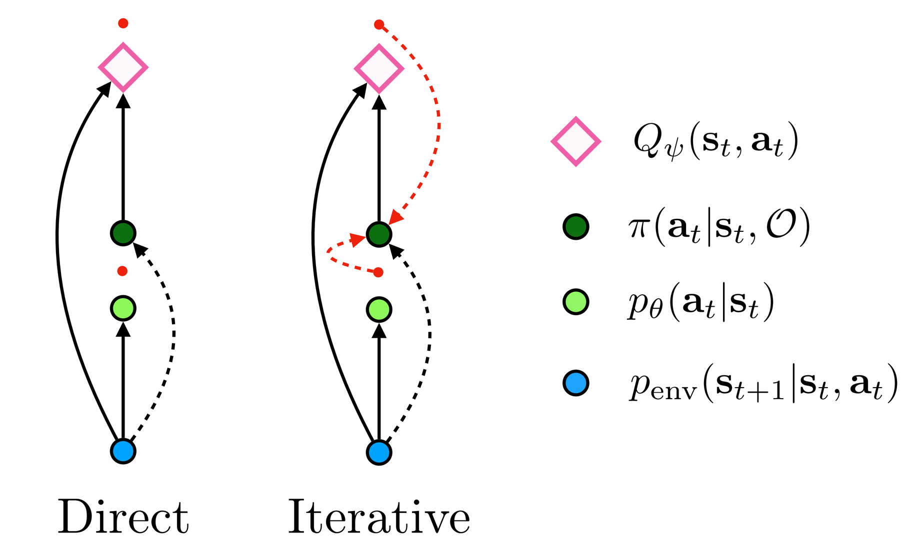
$\mathbf{u}_t \sim \mathcal{N}(\mathbf{u}_t; \boldsymbol{\mu}_\mathbf{u}, \text{diag}(\boldsymbol{\sigma}_\mathbf{u}^2))$

$\mathbf{a}_t = \tanh(\mathbf{u}_t)$

Direct Amortization (SAC):

$\boldsymbol{\mu}_\mathbf{u} = f_{\phi,\boldsymbol{\mu}}(\mathbf{s}_t)$

$\boldsymbol{\sigma}_\mathbf{u} = f_{\phi,\boldsymbol{\sigma}}(\mathbf{s}_t)$



Iterative Amortization:

$$\boldsymbol{\mu}_\mathbf{u} \leftarrow \boldsymbol{\omega}_{\phi,\boldsymbol{\mu}} \odot \boldsymbol{\mu}_\mathbf{u} + (1 - \boldsymbol{\omega}_{\phi,\boldsymbol{\mu}}) \odot f_{\phi,\boldsymbol{\mu}}(\mathbf{s}_t, \nabla_{\boldsymbol{\mu}_\mathbf{u}} \tilde{\mathcal{J}})$$

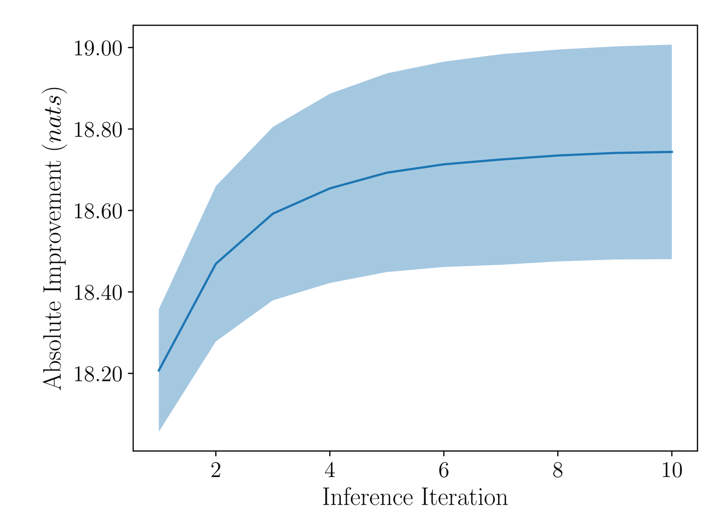
$$\boldsymbol{\sigma}_\mathbf{u} \leftarrow \boldsymbol{\omega}_{\phi,\boldsymbol{\sigma}} \odot \boldsymbol{\sigma}_\mathbf{u} + (1 - \boldsymbol{\omega}_{\phi,\boldsymbol{\sigma}}) \odot f_{\phi,\boldsymbol{\sigma}}(\mathbf{s}_t, \nabla_{\boldsymbol{\sigma}_\mathbf{u}} \tilde{\mathcal{J}})$$

results

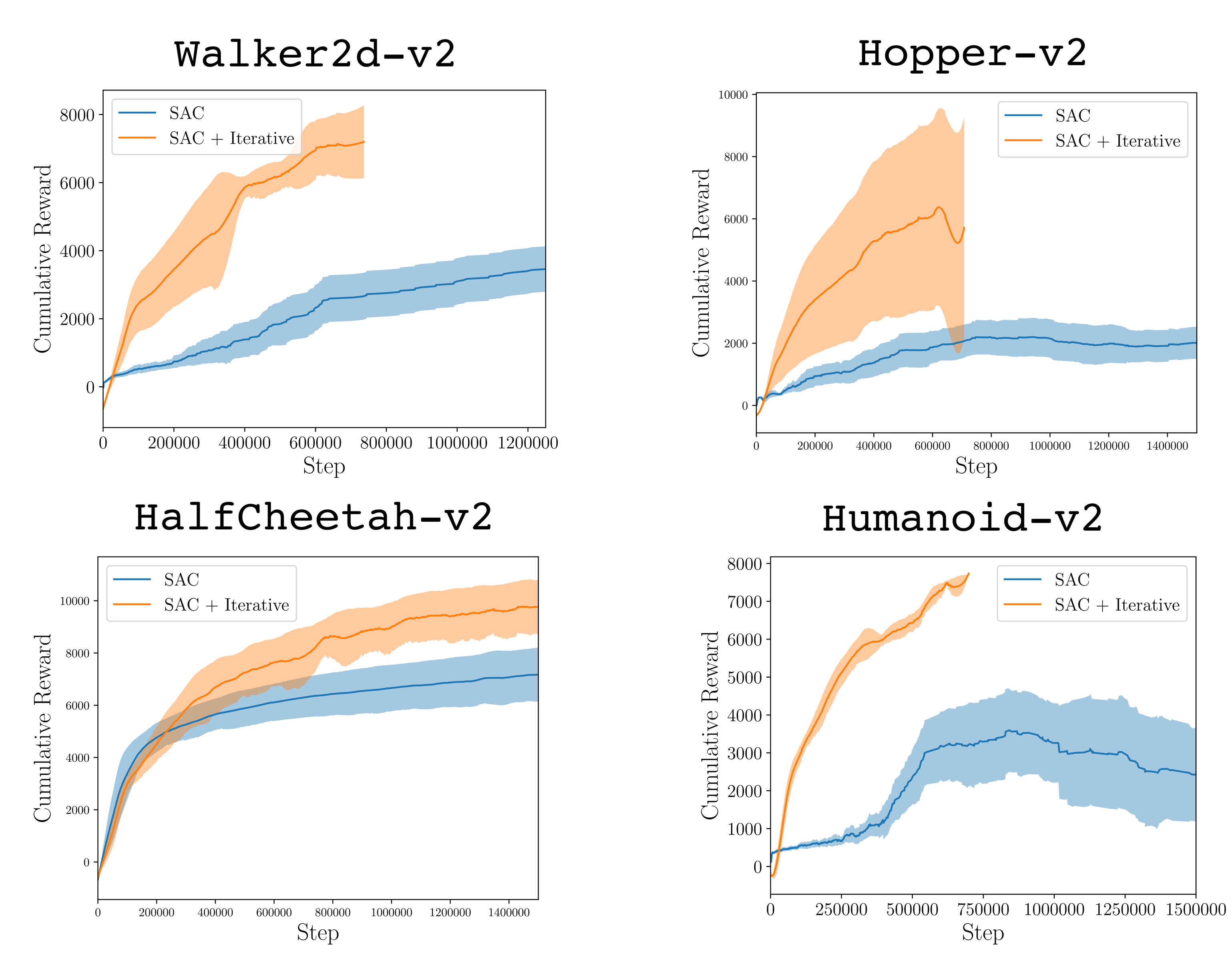
Inference Improvement

→ *with 10 inference iterations / step*

- performance increases across iterations
- diminishing improvement at each iteration



Performance Comparison on Continuous Control Tasks



discussion

Variational EM provides a common framework for many recent algorithms.

Variational EM also provides clear directions for improving current methods: more expressive priors, approximate posteriors, and conditional likelihoods, as well as better inference and learning procedures.

Iterative amortized inference results in significant performance improvements over direct amortized methods employed in current policy-based methods.