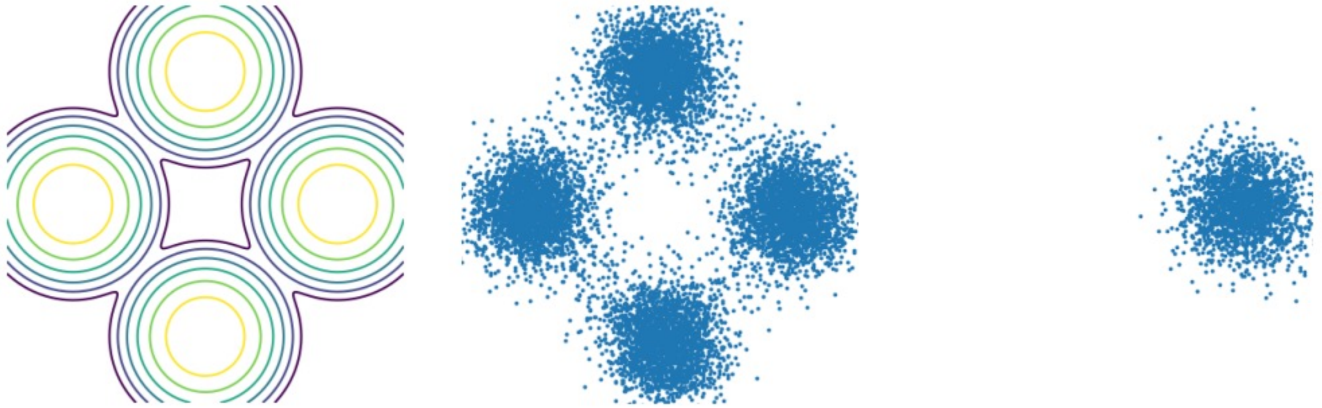
Latent State Marginalization as a Low-cost Approach for Improving Exploration

Dinghuai Zhang, Aaron Courville, Yoshua Bengio,
Qinqing Zheng, Amy Zhang, Ricky T. Q. Chen.

ICLR 2023
https://arxiv.org/abs/2210.00999

Motivation: multi-modes exploration

- Existing RL methods use factorized Gaussian for policy $\pi(a|x)$
  - Single mode behavior is limited
- A latent variable model for policy: $\pi(a|x) = \int \pi(a|s)p(s|x)ds$ will be more flexible for exploration

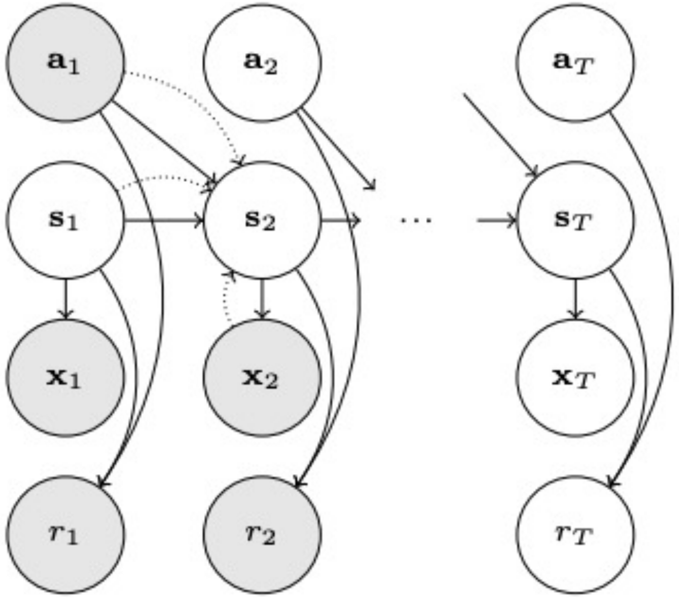


(a) Target $\left(\exp\left\{\frac{r(\mathbf{x},\mathbf{a})}{\alpha}\right\}\right)$ (b) Latent variable policy (c) Gaussian policy

Figure 7: Optimizing a latent variable policy for a one-step multi-modal MaxEnt RL objective.

Motivation: Partially Observed MDP

- In POMDP settings, we want to infer the true (latent) state from the observation
- Usually trained together with a world model



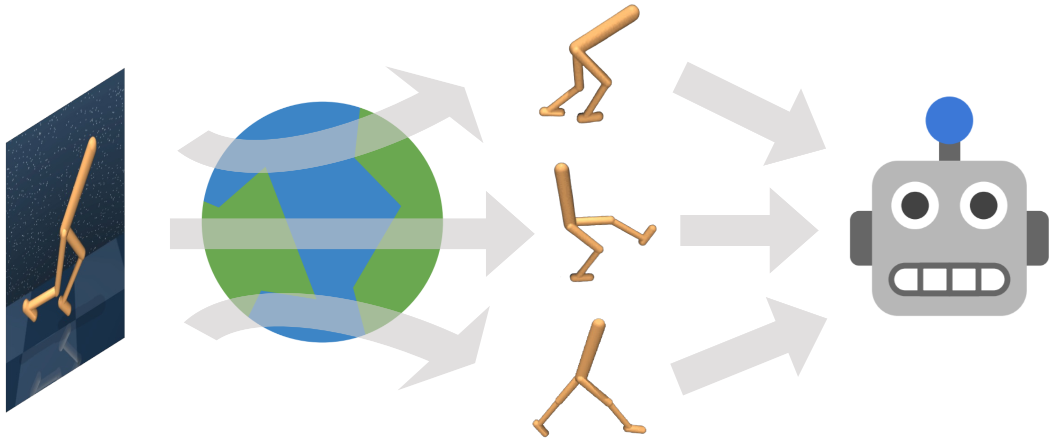
$$\log p(\mathbf{x}_{1:T}|\mathbf{a}_{1:T}) \geq \mathbb{E}_q \left[\sum_{t=1}^{T} \log p(\mathbf{x}_t|\mathbf{s}_t) - D_{\mathrm{KL}}(q(\mathbf{s}_t|\mathbf{s}_{t-1}, \mathbf{a}_{t-1}, \mathbf{x}_t) \| p(\mathbf{s}_t|\mathbf{s}_{t-1}, \mathbf{a}_{t-1})) \right]$$

Motivation: Partially Observed MDP

- Previous works:
  - Extract deterministic feature: $s = f(x)$, making decision conditioned on s: $\pi(a|s)$
  - Modeling the belief of true state $q(s|x)$ with a world model, but only use one sample or take mean of $q(s|x)$
- Information is lost! We should take the whole distribution into account

Latent State Marginalization

- We propose to marginalize out all the possible latent in belief distribution: $\pi(a|h) = \int \pi(a|s)q(s|h)ds$
  - from here we use h for all history obs, instead of x for single obs
  - $q(s|h)$ is from a world model, or an unstructured prior



MaxEnt RL

Formulation: $\mathcal{J}(\pi) = \sum_{t=0}^{T} \mathbb{E}_{(\mathbf{o}_t, \mathbf{a}_t) \sim \rho_\pi} \left[r(\mathbf{o}_t, \mathbf{a}_t) + \alpha \mathcal{H}(\pi(\cdot \mid \mathbf{o}_t)) \right]$

Soft Actor-Critic (SAC; Haarnoja et al. (2018)) algorithm:

Needs entropy estimation

$$\mathcal{J}_\pi(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{o} \sim \mathcal{D}} \mathbb{E}_{\mathbf{a} \sim \pi(\cdot|\mathbf{o})} \left[\alpha \log \pi_{\boldsymbol{\theta}}(\mathbf{a}|\mathbf{o}) - Q_{\boldsymbol{\theta}}(\mathbf{o}, \mathbf{a}) \right],$$

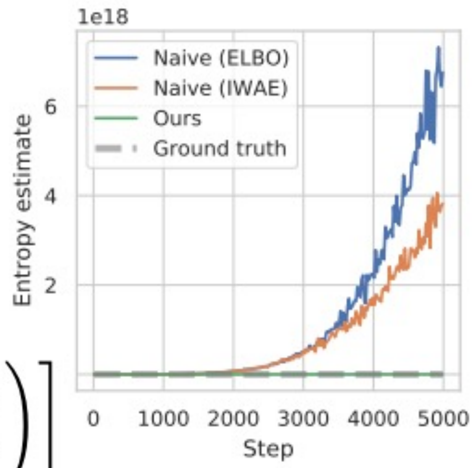
$$\mathcal{J}_Q(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{o}, \mathbf{a}, \mathbf{o}' \sim \mathcal{D}} \left[\frac{1}{2} \left(Q_{\boldsymbol{\theta}}(\mathbf{o}, \mathbf{a}) - r(\mathbf{o}, \mathbf{a}) - \gamma \bar{V}_{\boldsymbol{\theta}}(\mathbf{o}') \right)^2 \right],$$

where the value function $V_{\boldsymbol{\theta}}(\mathbf{o}') = \mathbb{E}_{\mathbf{a}' \sim \pi_{\boldsymbol{\theta}}(\cdot|\mathbf{o}')} \left[Q_{\boldsymbol{\theta}}(\mathbf{o}', \mathbf{a}') - \alpha \log \pi_{\boldsymbol{\theta}}(\mathbf{a}'|\mathbf{o}') \right]$, and $\bar{V}$ means a stop gradient operator upon V.

- We don’t change the base RL algorithm, only change policy
- Latent variable model (LVM) is easy to sample, but it is hard to estimate entropy (or log marginal probability)

Entropy estimation

- Normal methods such as ELBO, IWAE
  - Estimates lower bound of marginal prob
  - Thus upper bound of entropy term
- We cannot use upper bound of entropy for MaxEnt!
- Looking for a lower bound of entropy?



$$\widetilde{\mathcal{H}}_K(\mathbf{h}_t) \triangleq \mathbb{E}_{\mathbf{a}_t\sim\pi(\mathbf{a}_t|\mathbf{h}_t)}\mathbb{E}_{\mathbf{s}_t^{(0)}\sim\pi(\mathbf{s}_t|\mathbf{a}_t,\mathbf{h}_t)}\mathbb{E}_{\mathbf{s}_t^{(1:K)}\sim q(\mathbf{s}_t|\mathbf{h}_t)}\left[-\log\left(\frac{1}{K+1}\sum_{k=0}^{K}\pi\left(\mathbf{a}_t|\mathbf{s}_t^{(k)}\right)\right)\right]$$

Artem Sobolev et al. Importance weighted hierarchical variational inference. NeurIPS 2019.

- Variance reduction w/ multi-level Monte Carlo

$$\widetilde{\mathcal{H}}_K^{\mathrm{MLMC}} = \sum_{\ell=0}^{\lfloor\log_2(K)\rfloor}\Delta\widetilde{\mathcal{H}}_{2^\ell}, \quad \text{where } \Delta\widetilde{\mathcal{H}}_{2^\ell} = \begin{cases}\widetilde{\mathcal{H}}_1 & \text{if } \ell=0,\\ \widetilde{\mathcal{H}}_{2^\ell} - \frac{1}{2}\left(\widetilde{\mathcal{H}}_{2^{\ell-1}}^{(a)} + \widetilde{\mathcal{H}}_{2^{\ell-1}}^{(b)}\right) & \text{otherwise.}\end{cases}$$

Estimating marginal critic

- From ”control as inference” framework

$$Q(\mathbf{s}_t, \mathbf{a}_t) = \log p(\mathcal{O}_{t:T} | \mathbf{s}_t, \mathbf{a}_t)$$

$$p(\mathcal{O}_t = 1 | \mathbf{s}_t, \mathbf{a}_t) \propto \exp(r(\mathbf{s}_t, \mathbf{a}_t))$$

- We propose to estimate marginal Q-function

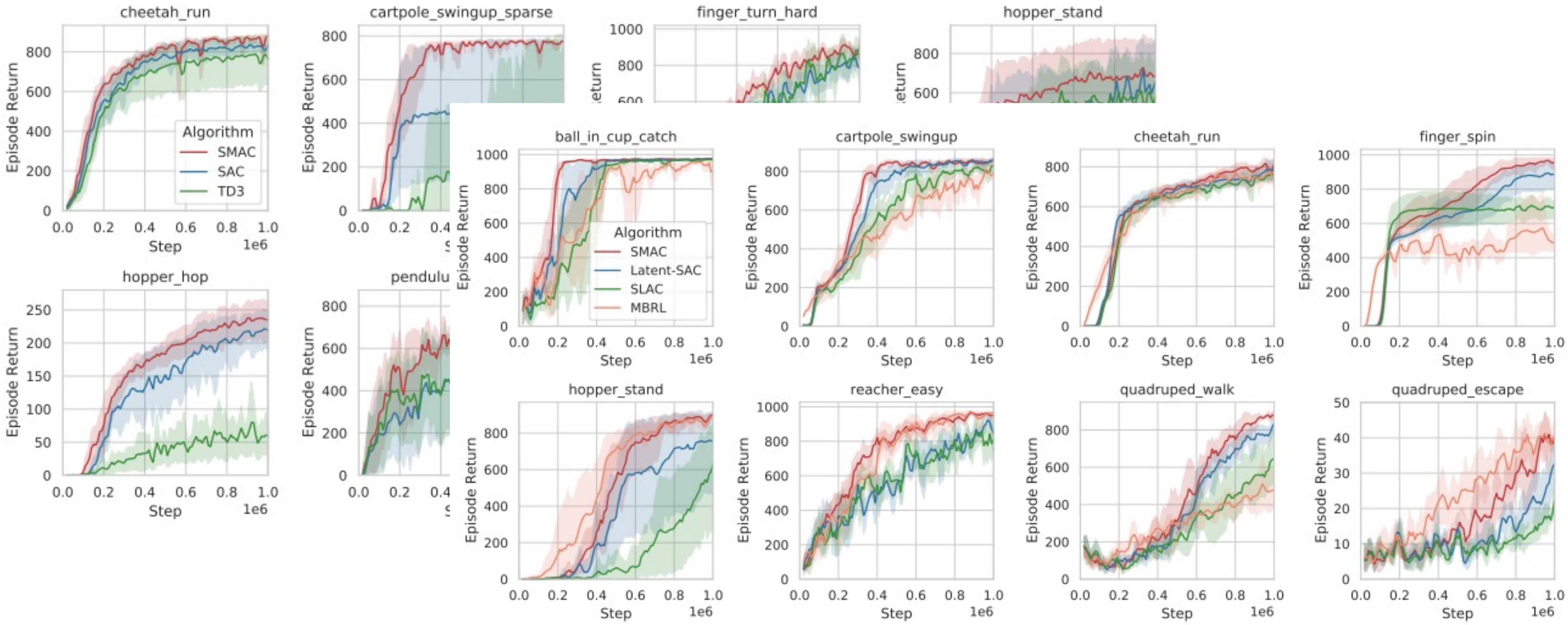
h: history information

$$Q(\mathbf{h}_t, \mathbf{a}_t) = \log \int p(\mathcal{O}_{t:T} | \mathbf{s}_t, \mathbf{a}_t) q(\mathbf{s}_t | \mathbf{h}_t) \, \mathrm{d}\mathbf{s}_t = \log \int \exp\{Q(\mathbf{s}_t, \mathbf{a}_t)\} q(\mathbf{s}_t | \mathbf{h}_t) \, \mathrm{d}\mathbf{s}_t$$

$$Q(\mathbf{h}_t, \mathbf{a}_t) \approx \widetilde{Q}_K(\mathbf{h}_t, \mathbf{a}_t) \triangleq \log \left(\frac{1}{K} \sum_{k=1} \exp\left\{ Q(\mathbf{s}_t^{(k)}, \mathbf{a}_t) \right\} \right), \quad \mathbf{s}_t^{(1:K)} \sim q(\mathbf{s}_t | \mathbf{h}_t)$$

Results

- Conduct experiments on various control tasks (DeepMind Control)



Thanks for listening!