Reparameterized Policy Learning for Multimodal Trajectory Optimization

Zhiao Huang [1] **Litian Liang** [1] **Zhan Ling** [1] **Xuanlin Li** [1] **Chuang Gan** [2 3] **Hao Su** [1]

[1]UC San Diego [2]MIT-IBM Watson AI Lab [3]UMass Amherst.
Correspondence to: Zhiao Huang <z2huang@ucsd.edu>.

ICML2023

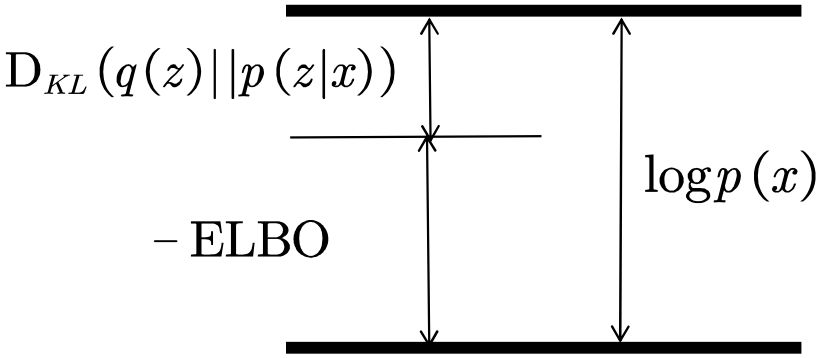
Background

Variational Inference

Purpose: fit a complex and unknown distribution $p(z|x)$ by a simple distribution $q(z)$

$$
\begin{aligned}
q^*(z) &= \arg\min_{q(z)\in Q} \mathrm{D}_{KL}(q(z)||p(z|x)) \\
&= \arg\min_{q(z)\in Q} \mathbb{E}_{q(z)}[\log q(z) - \log p(z|x)] \\
&= \arg\min_{q(z)\in Q} \mathbb{E}_{q(z)}\left[\log q(z) - \log \frac{p(z,x)}{p(x)}\right] \\
&= \arg\min_{q(z)\in Q} \mathbb{E}_{q(z)}[\log q(z) - \log p(z,x) + \log p(x)]
\end{aligned}
$$

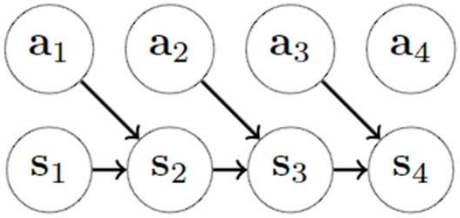
$q(z)$: variational distribution



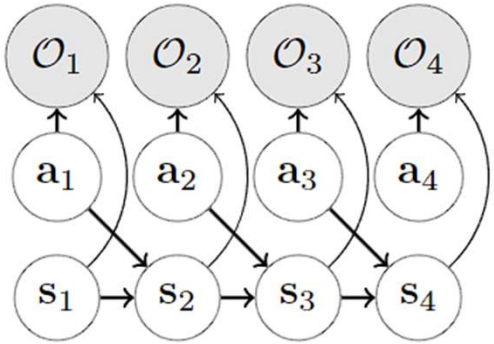
$$
\mathbb{E}_{q(z)}[\log q(z) - \log p(z,x) + \log p(x)] \geq 0 \Rightarrow \log \underbrace{p(x)}_{\text{evidence}} \geq \underbrace{\mathbb{E}_{q(z)}[\log p(z,x) - \log q(z)]}_{\text{ELBO}}
$$

Background

RL as probabilistic inference



(a) graphical model with states and actions



(b) graphical model with optimality variables

$$\tau_{1:T} = \{s_1, a_1, s_2, a_2, \cdots, s_T, a_T\}$$

$$p(\tau) = p(s_1)\,p(a_1)\,p(s_2|s_1,a_1)\cdots p(a_T)\,p(s_{T+1}|s_T,a_T) = p(s_1)\prod_{t=1}^{T} p(s_{t+1}|s_t,a_t)\prod_{t=1}^{T} p(a_t)$$

Optimal variable O_t : $p(O_t = 1\,|s_t, a_t) = \exp(r(s_t, a_t))/\mathcal{T}$

Background

$$p(\tau|O_{1:T}) = p(\tau, O_{1:T})\, p(O_{1:T}) = p(s_1) \prod_{t=1}^{T} p(a_t)\, p(s_{t+1}|s_t, a_t)\, p(O_t|s_t, a_t) \prod_{t=1}^{T} p(O_t)$$

$p(O_t)$ are the optimal variable prior which can be considered as a constant

Optimal variable O_t : $p(O_t|s_t, a_t) = \exp(r(s_t, a_t))$

$$p(\tau|O_{1:T}) \propto p(\tau, O_{1:T}) = \left[p(s_1) \prod_{t=1}^{T} p(a_t)\, p(s_{t+1}|s_t, a_t) \right] \exp\left(\sum_{t=1}^{T} r(s, a) \right)$$

Background

$$\hat{p}^{\pi}(\tau) = p(s_1)\prod_{t=1}^{T} p(a_t)\,p(s_{t+1}|s_t,a_t)\,\frac{\pi(a_t|s_t)}{p(a_t)} = \left[p(s_1)\prod_{t=1}^{T} p(a_t)\,p(s_{t+1}|s_t,a_t)\right]\prod_{t=1}^{T}\frac{\pi(a_t|s_t)}{p(a_t)}$$

$$p(\tau|O_{1:T}) \propto p(\tau, O_{1:T}) = \left[p(s_1)\prod_{t=1}^{T} p(a_t)\,p(s_{t+1}|s_t,a_t)\right]\exp\left(\sum_{t=1}^{T} r(s,a)\right)$$

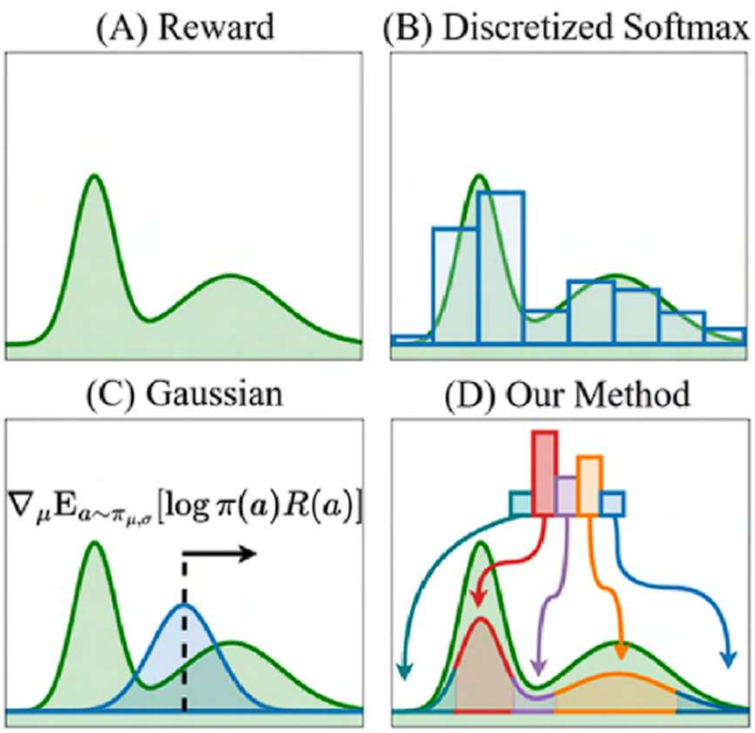
$$\min \mathrm{D}_{KL}\left(\hat{p}^{\pi}(\tau)\,||\,p(\tau|O_{1:T})\right) \Leftrightarrow \max -\mathbb{E}_{\tau\sim\hat{p}(\tau,\pi)}\left[\log\hat{p}(\tau,\pi) - \log p(\tau, O_{1:T})\right]$$

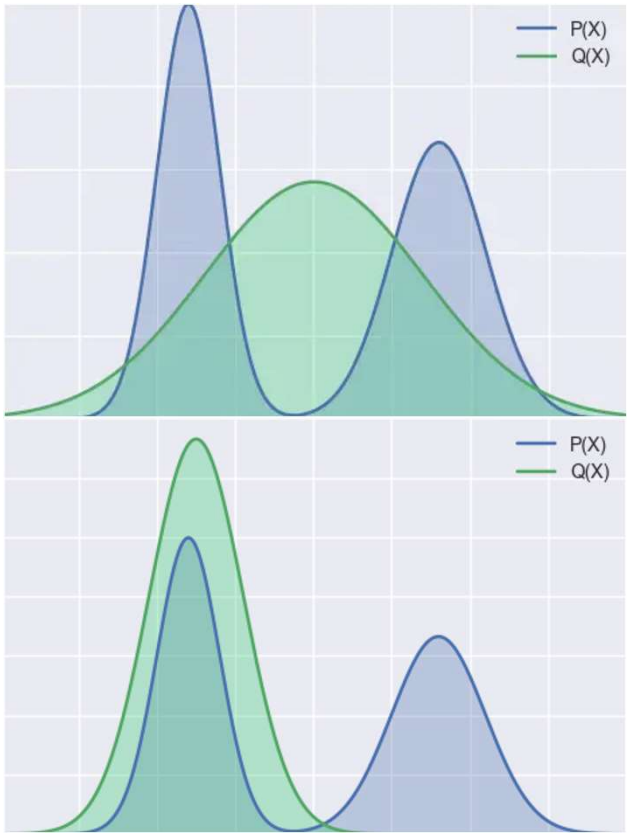
$$= \max\ \mathbb{E}_{\tau\sim\hat{p}(\tau,\pi)}\left[\sum_{t=1}^{T}\left(r(s_t,a_t) - \log\pi(a_t|s_t) + \log p(a_t)\right)\right]$$

$$\Leftrightarrow \max\ \sum_{t=1}^{T}\mathbb{E}_{s_t\sim\hat{p}(s_t),\,a_t\sim\pi(\cdot\,|s_t)}\left[r(s_t,a_t) - \log\pi(a_t|s_t)\right]$$

or $\log p(O_{1:T}) \geqslant \mathbb{E}_{\tau\sim\pi}\left[\log p(\tau, O_{1:T}) - \log\hat{p}^{\pi}(\tau)\right] = \mathbb{E}_{\tau\sim\pi}\left[\log p(O_{1:T}|\tau) + \log p(\tau) - \log\hat{p}^{\pi}(\tau)\right]$

Motivation





$$\mathrm{D}_{KL}(P(x)||Q(x))$$

$$\mathrm{D}_{KL}(Q(x)||P(x))$$

As the distribution of rewards (or Q function) can be multimodal, it is hard to evade local optima if the policy is modeled as a unimodal distribution

Method

Reparameterized Policy $\pi_\theta(z,\tau) = p(s_1)\pi_\theta(z|s_1)\prod_{t=1}^{T} p(s_{t+1}|s_t,a_t)\pi_\theta(a_t|z,s_t)$

$$\log p\left(O_{1:T}\right) \geqslant \mathbb{E}_{q(z)}\left[\log p\left(\tau, O_{1:T}\right) - \log \hat{p}^{\pi}\left(\tau\right)\right] = \mathbb{E}_{q(z)}\left[\log p\left(O_{1:T}|\tau\right) + \log p\left(\tau\right) - \log \hat{p}^{\pi}\left(\tau\right)\right]$$

$$\begin{aligned}
&\log p(O) \\
&= \underbrace{E_{z,\tau\sim\pi_\theta}\left[\log p_\phi(O,z,\tau) - \log \pi_\theta(z,\tau)\right]}_{\text{ELBO}} \\
&+ D_{KL}(\pi_\theta(z,\tau)||p_\phi(z,\tau|O)) \\
&\geq E_{z,\tau\sim\pi_\theta}\left[\log p_\phi(O,\tau,z) - \log \pi_\theta(z,\tau)\right] \qquad p_\phi(O,z,\tau) = p(O|z,\tau)\,p(z|\tau)\,p(\tau) = p(O|\tau)\,p(z|\tau)\,p(\tau) \\
&= E_{z,\tau\sim\pi_\theta}\left[\log p(O,\tau) + \log p_\phi(z|\tau) - \log \pi_\theta(z,\tau)\right] \\
&= E_{z,\tau}\left[\underbrace{\log p(O|\tau)}_{\text{reward}} + \underbrace{\log p(\tau)}_{\text{prior}} + \underbrace{\log p_\phi(z|\tau)}_{\text{cross entropy}} - \underbrace{\log \pi_\theta(z,\tau)}_{\text{entropy}}\right]
\end{aligned}$$

auxiliary distribution

Method

$$E_{z,\tau}\left[\underbrace{\log p(O|\tau)}_{\text{reward}}+\underbrace{\log p(\tau)}_{\text{prior}}+\underbrace{\log p_\phi(z|\tau)}_{\text{cross entropy}}-\underbrace{\log \pi_\theta(z,\tau)}_{\text{entropy}}\right]$$

reward term: $p(O|s,a)=\exp(r(s,a))/\mathcal{T}$

prior term: a constant

cross entropy term: model the posterior of z for τ sampled from π and distinguish different trajectories by z

$$I(\tau,z)=H(z)-H(z|\tau)=\mathbb{E}_{\tau,z}[\log p(z|\tau)-\log p(\tau)]=\mathbb{E}_{\tau,z}\left[\log\frac{p(z|\tau)}{p_\phi(z|\tau)}+\log p_\phi(z|\tau)-\log p(\tau)\right]\geq\mathbb{E}_{\tau,z}[\log p_\phi(z|\tau)-\log p(\tau)]$$

entropy term: encourage exploration like maximum entropy RL

$$\log p_\phi(z|\tau) \quad = \quad \textstyle\sum_{t>0}\log p(z|s_t,a_t)$$

final reward:

$$\underbrace{R(s_t,a_t)/\mathcal{T}}_{r_t}\underbrace{-\alpha\log\pi_\theta(a_t|s_t,z)+\beta\log p_\phi(z|s_t,a_t)}_{r'_t}$$

Method

Model-based RL

$$s_t = f_\psi(o_t)$$

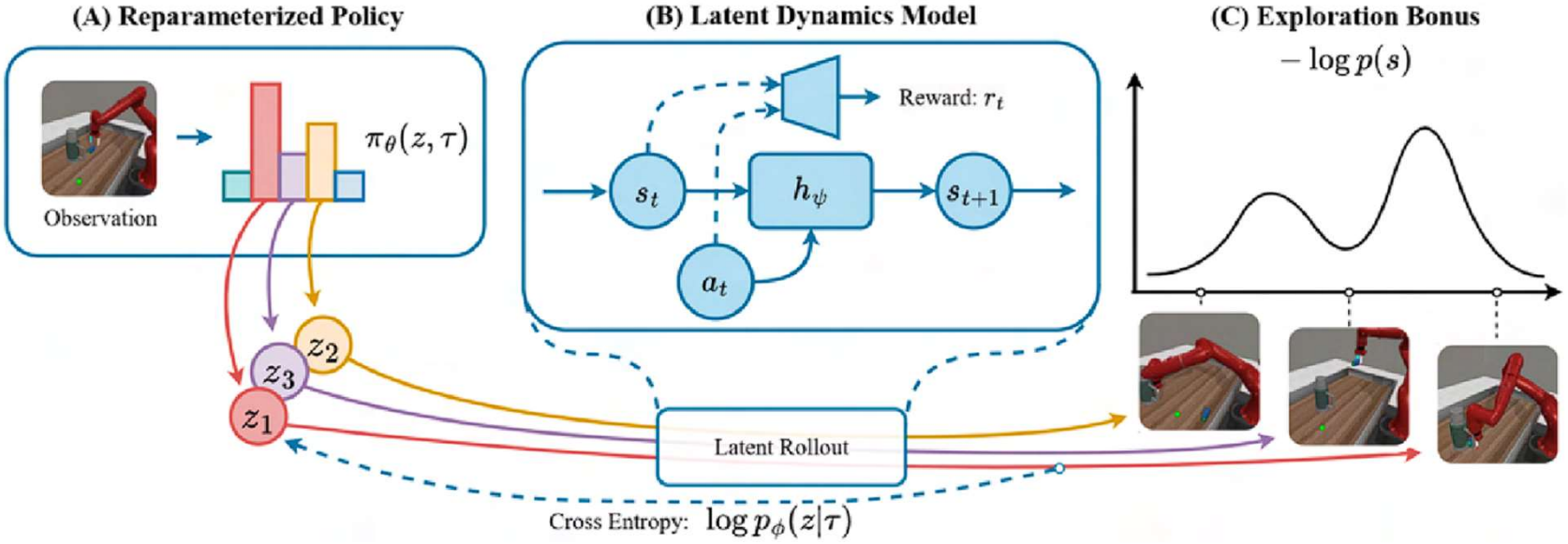
$$r_t = R_\psi(s_t, a_t)$$

$$Q_t = Q_\psi(s_t, a_t)$$

$$s_{t+1} = h_\psi(s_t, a_t)$$

$$V_{\text{est}}(o_{t_0}, z) \approx \gamma^K (Q_{t_0+K} + r'_{t_0+K}) + \sum_{t=t_0}^{t_0+K-1} \gamma^{t-t_0}(r_t + r'_t)$$

$$L_\psi(\tau) = \sum_{t=t_0}^{t_0+K-1} L_1 \|s_{t+1} - \mathbf{ng}(f_\psi(o_{t+1}))\|^2 + L_2(r_t - r_t^{gt})^2 + L_3(Q_t - \mathbf{ng}(r_t^{gt} + \gamma V_{\text{est}}(o_{t+1}, z)))^2 \quad (4)$$



Experiment

Algorithm 1 Model-based Reparameterized Policy Gradient

Input: $p_\phi, \pi_\theta, h_\psi, R_\psi, f_\psi, Q_\psi$ and an optional density estimator g_θ
Initialize p_ϕ, π_θ, construct the replay buffer $\mathcal{B}$.
while time remains **do**
Sample start state o_1 and encode it as $s_1 = f_\psi(o_1)$. Select z from $\pi_\theta(z|s_1)$.
Execute the policy $\pi_\theta(a|s, z)$ and store transitions into the replay buffer $\mathcal{B}$.
Sample a batch of trajectory segment of length K $\{\tau^i_{t:t+K}, z\}$ from the buffer $\mathcal{B}$.
Optional: update and estimate the density estimator g_θ and relabel transitions with the negative density as the intrinsic reward.
Optimize ψ using Equation 4.
Optimize $\pi_\theta(a|s, z)$ with gradient descent to maximize the value estimate in Equation 3 for s, z sampled from the buffer.
Optimize $\pi_\theta(z|s_1)$ with policy gradient to maximize $V_{\text{estimate}}(s_1, z) - \alpha \log \pi_\theta(z|s_1)$ for s_1 sampled from the buffer.
Optimize α, β if necessary .
end while

Experiment

Can multimodal policies help escape local optima?

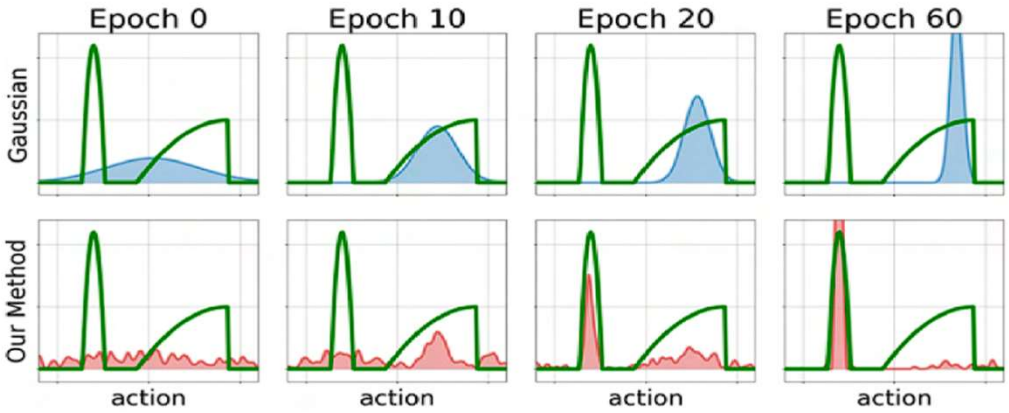


Figure 4. Illustrative experiment on continuous bandit

Can multimodal policies accelerate exploration?

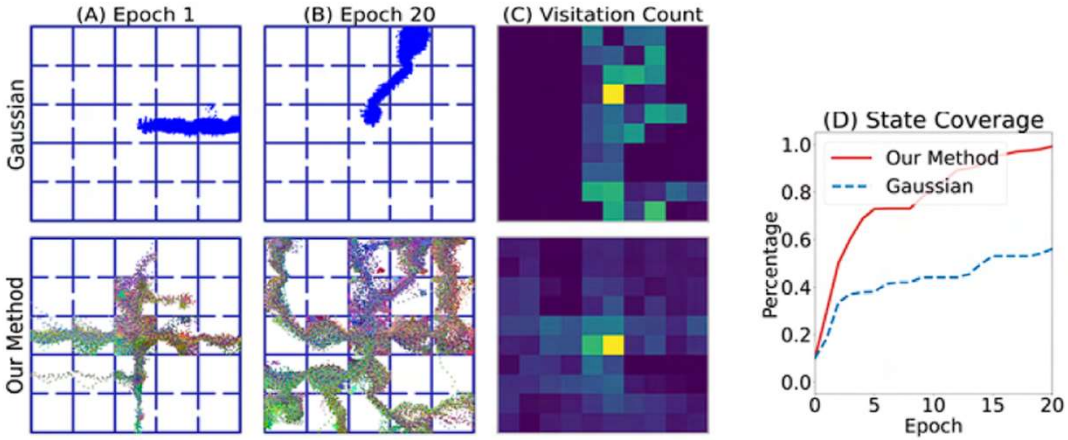
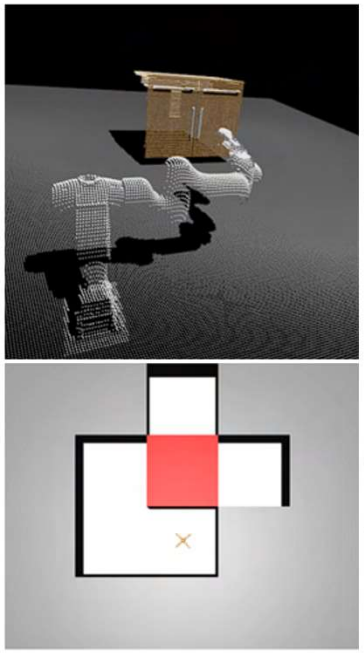


Figure 5. Illustrative experiment on 2D maze navigation problem

Experiment



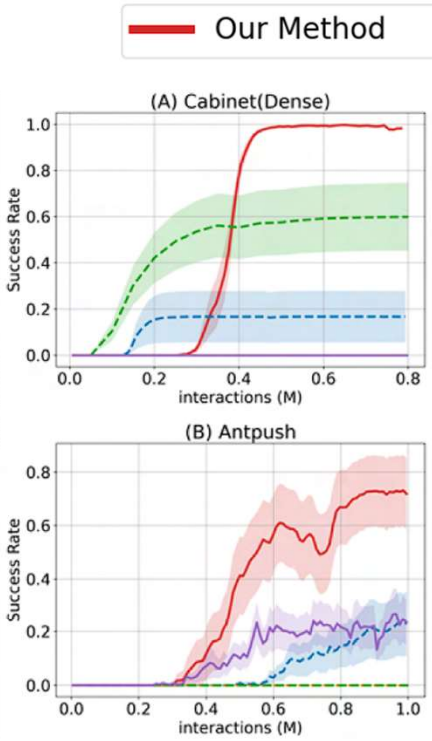


Figure 6. Results on dense reward tasks with local optima (exploration disabled)

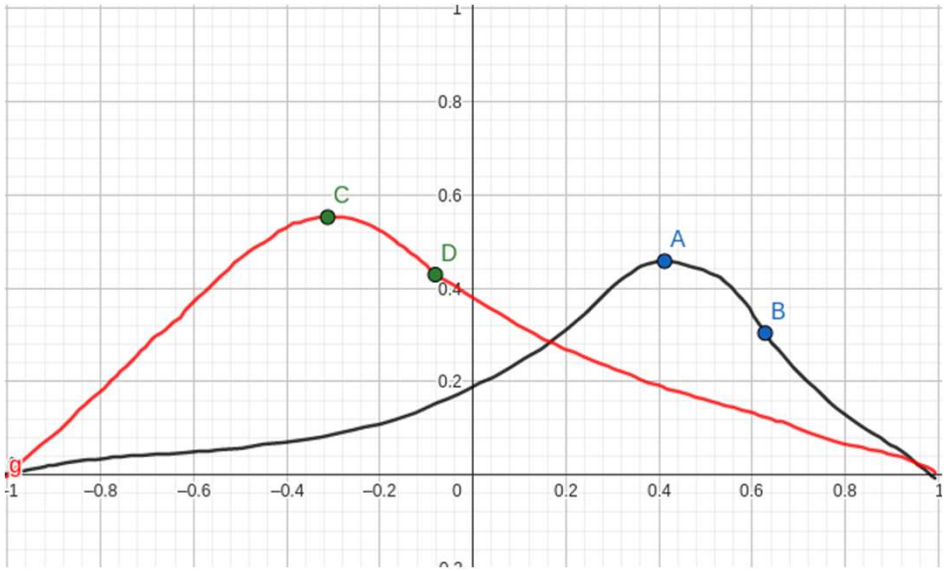
Figure 7. Results on sparse reward tasks

Discussion

PBRL

$$P_\psi[\sigma^1 \succ \sigma^0] = \frac{\exp \sum_t \widehat{r}_\psi(\mathbf{s}_t^1, \mathbf{a}_t^1)}{\sum_{i\in\{0,1\}} \exp \sum_t \widehat{r}_\psi(\mathbf{s}_t^i, \mathbf{a}_t^i)}$$

$$\mathcal{L}^{\texttt{Reward}} = - \mathop{\mathbb{E}}_{(\sigma^0,\sigma^1,y)\sim\mathcal{D}} \Big[y(0) \log P_\psi[\sigma^0 \succ \sigma^1] + y(1) \log P_\psi[\sigma^1 \succ \sigma^0] \Big]$$



$$\begin{aligned}
\log p\left(O_{1:T}\right) &\geqslant \mathbb{E}_{\tau\sim\pi}\left[\log p\left(\tau, O_{1:T}\right) - \log \hat{p}^{\pi}\left(\tau\right)\right] \\
&= \mathbb{E}_{\tau\sim\pi}\left[\log p\left(O_{1:T}|\tau\right) + \log p\left(\tau\right) - \log \hat{p}^{\pi}\left(\tau\right)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{s_t\sim\hat{p}^{\pi}\left(s_t\right), a\sim\pi\left(\cdot|s_t\right)}\left[r\left(s_t, a_t\right) - \log \pi\left(a_t|s_t\right)\right]
\end{aligned}$$

Discussion

最小化与较好轨迹的KL散度，最大化与较差轨迹的KL散度

$$\mathrm{D}_{KL}\left(\hat{p}^{\pi}(\tau)||p(\tau|O_{1:T})\right) \Leftrightarrow \max \sum_{t=1}^{T} \mathbb{E}_{s_t\sim\hat{p}^{\pi}(s_t),a_t\sim\pi(\cdot|s_t)}\left[r(s_t,a_t)-\log\pi(a_t|s_t)\right]$$

$$\Leftrightarrow \max \sum_{t=1}^{T} \mathbb{E}_{s_t\sim\hat{p}^{\pi}(s_t),a_t\sim\pi(\cdot|s_t)}\left[r(s_t,a_t)\right] + \sum_{t=1}^{T} \mathbb{E}_{s_t\sim\hat{p}^{\pi}(s_t)}\left[H\left(\pi(\cdot|s_t)\right)\right]$$

注意到$\hat{p}^{\pi}(\tau)=p(s_1)\prod_{t=1}^{T}p(a_t)\,p(s_{t+1}|s_t,a_t)\,\dfrac{\pi(a_t|s_t)}{p(a_t)}=\left[p(s_1)\prod_{t=1}^{T}p(a_t)\,p(s_{t+1}|s_t,a_t)\right]\prod_{t=1}^{T}\dfrac{\pi(a_t|s_t)}{p(a_t)}$，其中第一项与轨迹$\tau$的转移动态一致，故假设$\hat{p}^{\pi}(s_t)=p^{\tau}(s_t)$

$$\mathrm{D}_{KL}\left(\hat{p}^{\pi}(\tau)||p(\tau_w|O_{1:T})\right)-\mathrm{D}_{KL}\left(\hat{p}^{\pi}(\tau)||p(\tau_l|O_{1:T})\right) \Leftrightarrow \max \sum_{t=1}^{T} \mathbb{E}_{(s_t,a_t)\sim\tau_w}\frac{\pi(a_t|s_t)}{\pi_{\tau_w}(a_t|s_t)}\left[r(s_t,a_t)\right] + \sum_{t=1}^{T} \mathbb{E}_{(s_t,a_t)\sim\tau_w}\left[H\left(\pi(\cdot|s_t)\right)\right]$$

$$- \sum_{t=1}^{T} \mathbb{E}_{(s_t,a_t)\sim\tau_l}\frac{\pi(a_t|s_t)}{\pi_{\tau_l}(a_t|s_t)}\left[r(s_t,a_t)\right] - \sum_{t=1}^{T} \mathbb{E}_{(s_t,a_t)\sim\tau_l}\left[H\left(\pi(\cdot|s_t)\right)\right]$$

Discussion

对于PBRL，目标为最大化$\exp\left(\sum_{t=1}^{T_{\mathrm{w}}}\mathbb{E}_{\tau\sim p(\tau_{\mathrm{w}})}[r(s_t,a_t)]-\sum_{t=1}^{T_{\mathrm{l}}}\mathbb{E}_{\tau\sim p(\tau_{\mathrm{w}})}[r(s_t,a_t)]\right)$

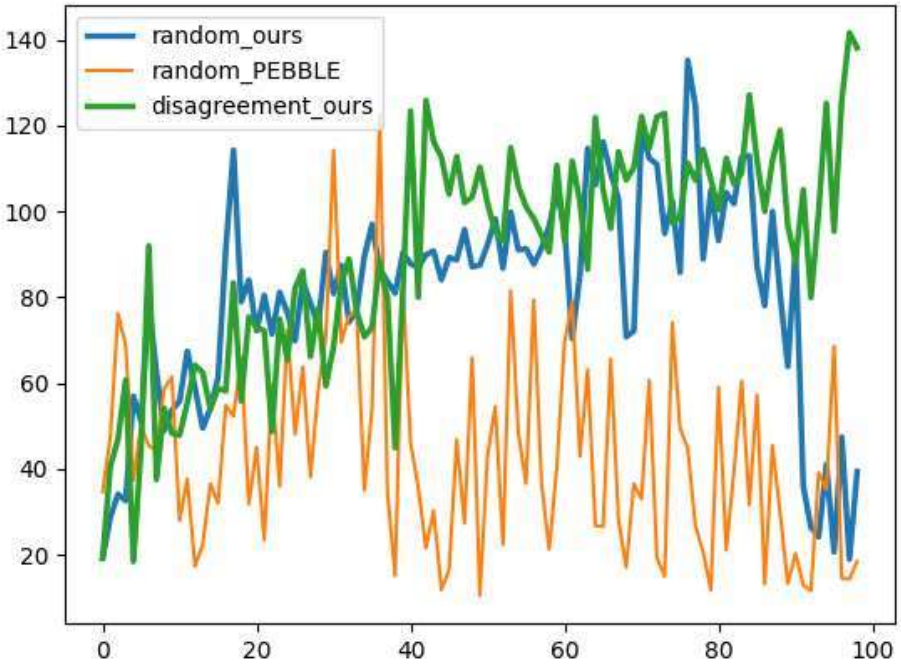
对比而言，应该最大化$\exp\left(-\mathrm{D}\left(\hat{p}^{\pi}(\tau)||p(\tau_{\mathrm{w}}|O_{1:T})\right)+\mathrm{D}\left(\hat{p}^{\pi}(\tau)||p(\tau_{\mathrm{l}}|O_{1:T})\right)\right)$

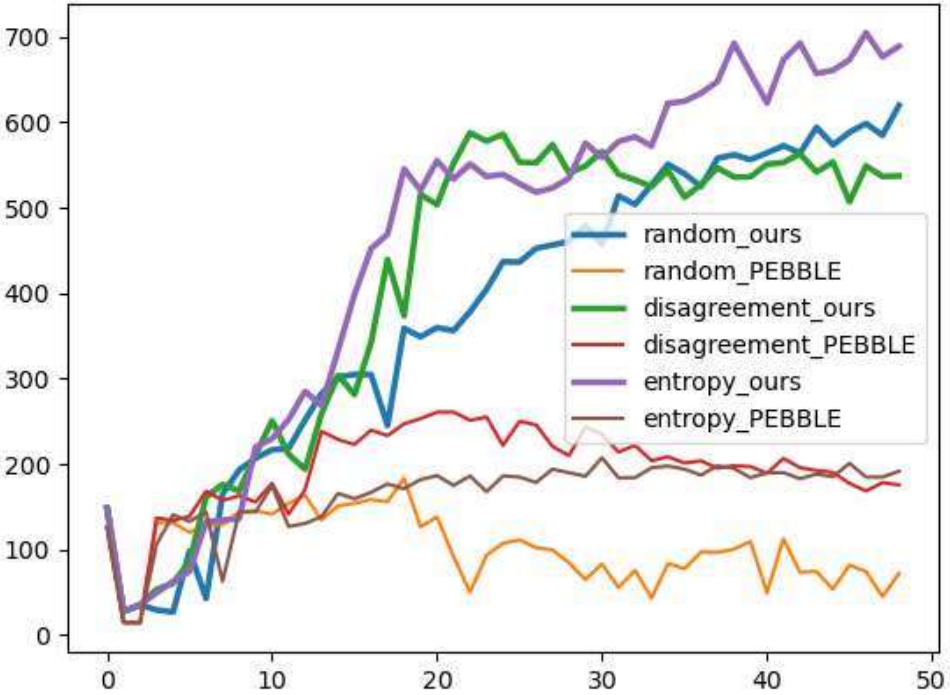
$$=\exp\left(\sum_{t=1}^{T_{\mathrm{w}}}\mathbb{E}_{\tau\sim p(\tau_{\mathrm{w}})}\frac{\pi(a_t|s_t)}{\pi_{\tau_w}(a_t|s_t)}[r(s_t,a_t)]-\sum_{t=1}^{T_{\mathrm{l}}}\mathbb{E}_{\tau\sim p(\tau_{\mathrm{l}})}\frac{\pi(a_t|s_t)}{\pi_{\tau_l}(a_t|s_t)}[r(s_t,a_t)]\right)$$

由于PBRL中较差轨迹和较好轨迹都是以前策略采样动作交互得到的，那么可以在策略采样动作记录$\log\pi_b(a|s)$，

在学习奖励模型的时候即可计算重要性采样比$\frac{\pi(a|s)}{\pi_b(a|s)}=\exp(\log\pi(a|s)-\log\pi_b(a|s))$，作为不同状态动作对下的奖励的权重

Discussion


random_ours
random_PEBBLE
disagreement_ours
140
120
100
80
60
40
20
0
20
40
60
80
100


random_ours
random_PEBBLE
disagreement_ours
disagreement_PEBBLE
entropy_ours
entropy_PEBBLE
700
600
500
400
300
200
100
0
0
10
20
30
40
50

Discussion

$$L(\phi) = -\log\left(\frac{\exp\left(\sum r_\phi(s_t^1,a_t^1)\right)}{\exp\left(\sum r_\phi(s_t^1,a_t^1)\right)+\exp\left(\sum r_\phi(s_t^2,a_t^2)\right)}\right)$$

$$\frac{\partial L}{\partial r_\phi(s_t^1,a_t^1)} = -\left(1 - \frac{\exp(\sum r_\phi(s_t^1,a_t^1))}{\exp(\sum r_\phi(s_t^1,a_t^1))+\exp(\sum r_\phi(s_t^2,a_t^2))}\right)$$

$$\frac{\partial L}{\partial r_\phi(s_t^2,a_t^2)} = 1 - \frac{\exp(\sum r_\phi(s_t^1,a_t^1))}{\exp(\sum r_\phi(s_t^1,a_t^1))+\exp(\sum r_\phi(s_t^2,a_t^2))}$$

$$L(\phi) = -\log\left(\frac{\exp\left(\sum A(s_t^1,a_t^1) r_\phi(s_t^1,a_t^1)\right)}{\exp\left(\sum A(s_t^1,a_t^1) r_\phi(s_t^1,a_t^1)\right)+\exp\left(\sum A(s_t^2,a_t^2) r_\phi(s_t^2,a_t^2)\right)}\right)$$

$$\frac{\partial L}{\partial r_\phi(s_t^1,a_t^1)} = -A(s_t^1,a_t^1)\left(1 - \frac{\exp(\sum A(s_t^1,a_t^1) r_\phi(s_t^1,a_t^1))}{\exp(\sum A(s_t^1,a_t^1) r_\phi(s_t^1,a_t^1))+\exp(\sum A(s_t^2,a_t^2) r_\phi(s_t^2,a_t^2))}\right)$$

$$\frac{\partial L}{\partial r_\phi(s_t^2,a_t^2)} = A(s_t^2,a_t^2)\left(1 - \frac{\exp(\sum A(s_t^1,a_t^1) r_\phi(s_t^1,a_t^1))}{\exp(\sum A(s_t^1,a_t^1) r_\phi(s_t^1,a_t^1))+\exp(\sum A(s_t^2,a_t^2) r_\phi(s_t^2,a_t^2))}\right)$$

梯度$\propto |A(s,a)(1-P)|$，最大化梯度以最大化每次更新能够获得的信息