RL——Deep Reinforcement Learning amidst Continual/Lifelong Structured Non-Stationarity

作者：凯鲁嘎吉 - 博客园 http://www.cnblogs.com/kailugaji/

这篇博客简要回顾论文"Deep Reinforcement Learning amidst Continual/Lifelong Structured Non-Stationarity"，并记录一下阅读笔记，大多数强化学习算法都用于求解无限阶段且平稳的马尔可夫决策过程(MDP)，求解有限阶段非平稳的MDP的强化学习算法很少见到。论文主要研究求解有限时段非平稳的MDP的强化学习算法，这篇文章可理解为变分自编码器(VAE)框架下的软演员评论员(Soft Actor-Critic, SAC)异策略(off-policy)强化学习算法。看这篇博文的前提需要了解变分推断与变分自编码器(许多推导用到了变分推断的推导过程)，强化学习中的演员-评论员算法。主要讲解有关变分推断的基础知识(涉及到贝叶斯公式)，Jensen不等式(推导证据下界时用到)，平均场理论(RL推导中涉及的因式分解用到这个变分理论)，将强化学习形式化为变分推断问题，非平稳性形式化为概率模型，基于变分推断的联合表示和强化学习(对论文中的公式进行推导，其中，L_{KL}项与原文推导有些许出入，但无非是z采样问题，并无大碍)，以及Lifelong Latent Actor-Critic (LILAC)算法流程。

作为人类，我们的目标和环境在我们的一生中不断变化，这是基于我们的经验、行动以及内在和外在的驱动力。相比之下，经典的强化学习问题设置考虑的决策过程是跨回合的平稳过程。我们能否提出一种强化学习算法来应对先前更现实的问题设置的持续变化?虽然同策略(on-policy)的算法，如策略梯度，在原则上可以扩展到非平稳设置，但更有效的异策略(off-policy)算法(在学习时回放过去的经验)却不能这样。这篇论文形式化了这个问题设置，并借鉴在线学习和概率推理文献的思想，得出了一个异策略强化学习算法，该算法可以推理和处理这种终身非平稳性。所提方法利用潜在变量模型从当前和过去的经验中学习环境的表示，并使用该表示执行异策略强化学习。论文进一步介绍了几个具有终生非平稳性的模拟环境，并根据经验发现，所提方法大大优于那些不考虑环境变化的方法。

1. 前提基础

包括参数估计、贝叶斯估计、变分推断、Jensen不等式、平均场理论，以及软演员评论员(Soft Actor-Critic, SAC)算法。

➢前提基础

- 参数估计

  根据样本所提供的信息，对总体分布中的未知参数θ进行估值

- 贝叶斯估计

极大似然估计

最大后验估计

$$p(\theta \mid X) = \frac{p(X \mid \theta)p(\theta)}{p(X)} = \frac{p(X \mid \theta)p(\theta)}{\int p(X,\theta)d\theta}$$

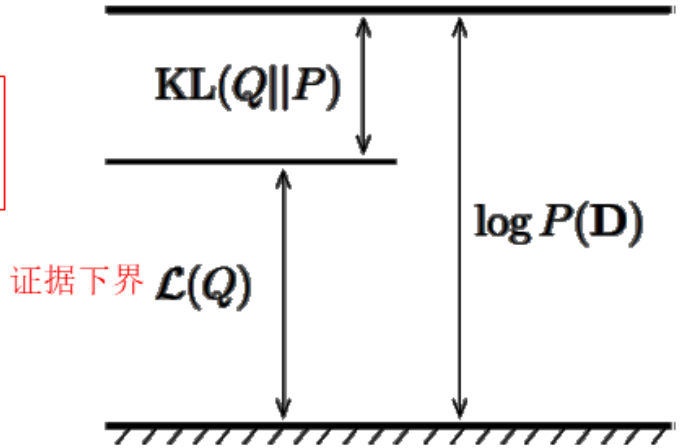
- 变分推断

- 贝叶斯估计中分母$p(X)$往往很难求，于是找一个简单的函数$q(\theta) \approx p(\theta|X)$
- 如何评价$q(\theta)$与$p(\theta|X)$之间的近似程度呢？——Kullback-Leibler散度
- 目标函数：min $KL(q(\theta)\|p(\theta|X))$

➢前提基础

$$KL(q \| p) = \int q(\theta) \ln \frac{q(\theta)}{p(\theta \mid X)} d\theta = \int q(\theta) \ln \frac{q(\theta)}{p(X,\theta)} d\theta + \ln p(X)$$

$$\ln p(X) = KL(q \| p) + \int q(\theta) \ln \frac{p(X,\theta)}{q(\theta)} d\theta = KL(q \| p) + L(q)$$

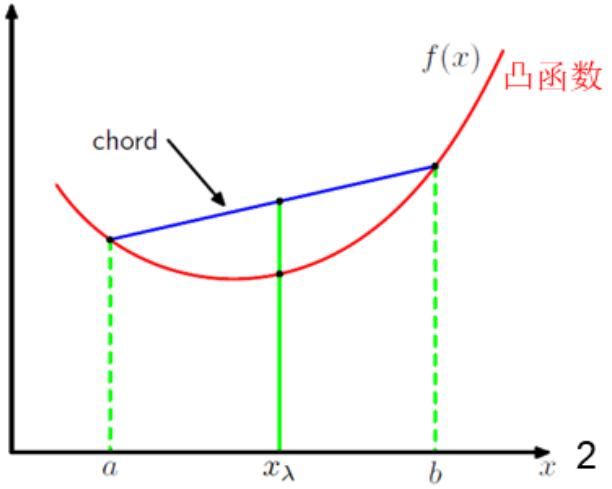


- $\ln p(X) = KL(q\|p) + L(q)$, 而$\ln p(X)$是与$\theta$无关的常量, $\min KL(q(\theta)\|p(\theta|X)) \Leftrightarrow \max L(q)$
- Jensen不等式

若$f(x)$在$[a,b]$上为凸函数, 对$\forall x_i \in [a,b]$, $\lambda_i > 0 (i = 1,\ldots,N)$, $\sum_{i=1}^{N} \lambda_i = 1$

则$f\left(\sum_{i=1}^{N} \lambda_i x_i\right) \le \sum_{i=1}^{N} \lambda_i f(x_i)$, 即$f(\mathbb{E}[x]) \le \mathbb{E}\left[f(x)\right]$.

若$f(x)$为凹函数, 则相反，例如$f(x)$为对数函数，则$f(\mathbb{E}[x]) \ge \mathbb{E}\left[f(x)\right]$.



➢前提基础

- 平均场理论

根据平均场理论，变分分布$q(\theta)$可以因式分解为M个互不相交的部分

$$q(\theta)=\prod_{i=1}^{M}q_i(\theta_i)$$

$$\max L(q)=\int q\ln p(X,\theta)d\theta-\int q\ln q d\theta=\int q_j\left\{\int \ln p(X,\theta)\prod_{i\neq j}q_i d\theta_i\right\}d\theta_j-\int q_j\ln q_j d\theta_j+c$$

$$=\int q_j\ln\tilde{p}(X,\theta_j)d\theta_j-\int q_j\ln q_j d\theta_j+c=-KL(q_j\,\|\,\tilde{p}(X,\theta_j))+c$$

$$\text{其中}\ln\tilde{p}(X,\theta_j)=E_{i\neq j}[\ln p(X,\theta)]+c=\int\ln p(X,\theta)\prod_{i\neq j}q_i d\theta_i+c$$

$$q_j=\tilde{p}(X,\theta_j),L(q)\text{最大}\qquad\therefore\ln q_j^{*}=\ln\tilde{p}(X,\theta_j)=E_{i\neq j}[\ln p(X,\theta)]+c$$

$$q_j^{*}(\theta_j)=\frac{\exp\left(E_{i\neq j}[\ln p(X,\theta)]\right)}{\int\exp\left(E_{i\neq j}[\ln p(X,\theta)]\right)d\theta_j}$$

⇒ 采样得到参数θ

➢前提基础

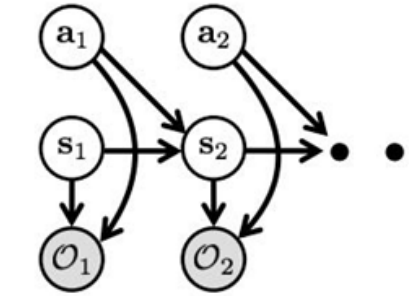
- 软演员评论员(Soft Actor-Critic, SAC)算法

1. Q-function update

Update Q-function to evaluate current policy:

$$Q(\mathbf{s},\mathbf{a}) \leftarrow r(\mathbf{s},\mathbf{a}) + \mathbb{E}_{\mathbf{s}'\sim p_{\mathbf{s}},\ \mathbf{a}'\sim\pi}\left[Q(\mathbf{s}',\mathbf{a}') - \log\pi(\mathbf{a}'|\mathbf{s}')\right]$$

This converges to Q^{π}.



update messages

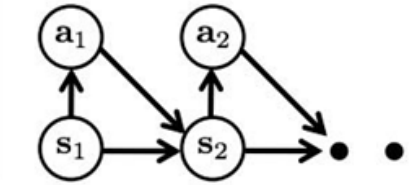
fit variational distribution

2. Update policy

Update the policy with gradient of information projection:

$$\pi_{\text{new}} = \arg\min_{\pi'} \mathrm{D}_{\mathrm{KL}}\left(\pi'(\cdot|\mathbf{s}) \,\middle\|\, \frac{1}{Z}\exp Q^{\pi_{\text{old}}}(\mathbf{s},\cdot)\right)$$

In practice, only take one gradient step on this objective

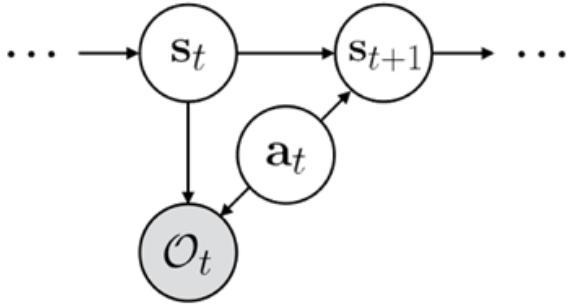


3. Interact with the world, collect more data

Haarnoja, Zhou, Hartikainen, Tucker, Ha, Tan, Kumar, Zhu, Gupta, Abbeel, L. **Soft Actor-Critic Algorithms and Applications**. '18

2. 强化学习作为变分推断

➢强化学习作为变分推断



状态s_t, 动作a_t, 最优性变量$\mathcal{O}_t$, 奖励$p(\mathcal{O}_t=1\mid s_t,a_t)=e^{r(s_t,a_t)}$

轨迹$(s_1,a_1,s_2,\ldots,s_T,a_T)$, 推断后验分布$p(s_{1:T},a_{1:T}\mid\mathcal{O}_{1:T}=1)$

假设动作的先验分布是均匀分布，则最优轨迹分布为

$$p(s_{1:T},a_{1:T}\mid\mathcal{O}_{1:T}=1)=\frac{p(s_{1:T},a_{1:T},\mathcal{O}_{1:T}=1)}{p(\mathcal{O}_{1:T}=1)}\propto p(s_{1:T},a_{1:T},\mathcal{O}_{1:T}=1)$$

$$=p(s_1)\prod_{t=1}^{T}e^{r(s_t,a_t)}p(s_{t+1}\mid s_t,a_t)$$

$$p(\theta\mid X)=\frac{p(X,\theta)}{p(X)},\ \ln p(X)=KL(q\,\|\,p)+L(q)$$

$$L(q)=\int q(\theta)\ln\frac{p(X,\theta)}{q(\theta)}d\theta=\mathbb{E}_q\left[\ln\frac{p(X,\theta)}{q(\theta)}\right]$$

用$q(s_{1:T},a_{1:T})$近似$p(s_{1:T},a_{1:T}\mid\mathcal{O}_{1:T}=1)$, $q(s_{1:T},a_{1:T})=p(s_1)\prod_{t=1}^{T}p(s_{t+1}\mid s_t,a_t)q(a_t\mid s_t)=p(s_1)\prod_{t=1}^{T}p(s_{t+1}\mid s_t,a_t)\pi(a_t\mid s_t)$

$$\ln p(\mathcal{O}_{1:T}=1)=L(q)+KL\left(q(s_{1:T},a_{1:T})\,\|\,p(s_{1:T},a_{1:T}\mid\mathcal{O}_{1:T}=1)\right)$$

$$\min KL\left(q(s_{1:T},a_{1:T})\,\|\,p(s_{1:T},a_{1:T}\mid\mathcal{O}_{1:T}=1)\right)\Leftrightarrow$$

$$\max L(q)=\mathbb{E}_q\left[\ln\frac{p(s_{1:T},a_{1:T},\mathcal{O}_{1:T}=1)}{q(s_{1:T},a_{1:T})}\right]=\mathbb{E}_q\left[\ln\frac{p(s_1)\prod_{t=1}^{T}e^{r(s_t,a_t)}p(s_{t+1}\mid s_t,a_t)}{p(s_1)\prod_{t=1}^{T}p(s_{t+1}\mid s_t,a_t)\pi(a_t\mid s_t)}\right]=\mathbb{E}_\pi\left[\sum_{t=1}^{T}\left(r(s_t,a_t)-\ln\pi(a_t\mid s_t)\right)\right]$$

Entropy-regularized RL

3. 非平稳环境下的异策略(off-policy)强化学习

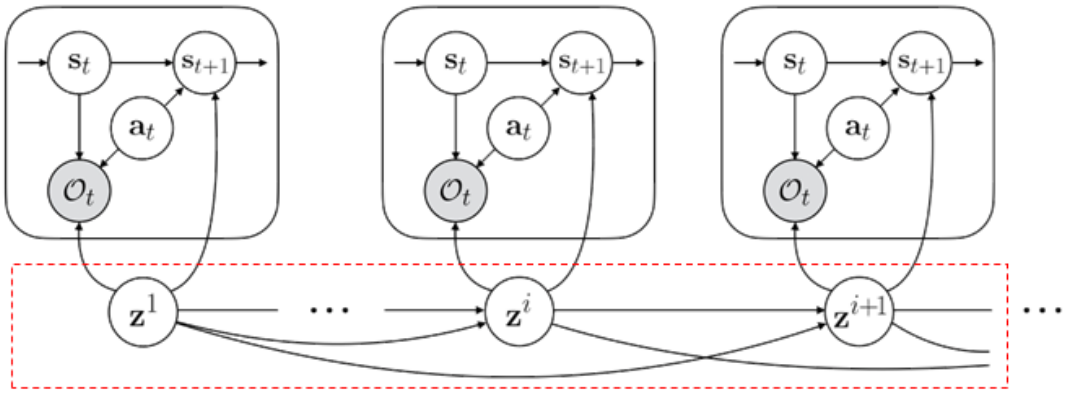
包括非平稳性作为概率模型、基于变分推断的联合表示和强化学习，以及Lifelong Latent Actor-Critic (LILAC)算法流程。

➢非平稳环境下的异策略(off-policy)强化学习

- **平稳策略**：动作的选取只和当前的状态有关，而与时间无关；
- **非平稳策略**：经时间索引后的一系列状态到行动的集合。非平稳策略即使对于同样的状态，在过程的不同时刻，可能会对应不同的行动。

- 非平稳性作为概率模型

  可以将动态参数马尔科夫决策过程(MDP)转换为一个概率层次模型，其中非平稳性发生在每个回合，并且在每个回合都是平稳MDP的一个实例。为此，构建一个双层模型：第一层，将隐变量序列z^i作为马尔可夫链，第二层，对应于每个z^i的马尔可夫决策过程。所提出的动态参数马尔科夫决策过程(DP-MDP)的概率图模型如右图所示。在这个模式下，从每个回合收集的轨迹被单独建模，而不是像上一页那样被摊销化处理。



- 给定动作序列$u=a_{1:T}$，轨迹与隐变量的联合概率定义为：

$$p(z^{1:N},\tau^{1:N}\mid u^{1:N})=p(z^1)p(\tau^1\mid z^1,u^1)\prod_{i=1}^{N}p(z^i\mid z^{1:i-1})p(\tau^i\mid z^i,u^i)$$

- 其中，$p(\tau\mid z,u)=\dfrac{p(\tau,u;z)}{p(u;z)}\propto p(\tau,u;z)=p(s_{1:T},\mathcal{O}_{1:T}=1,a_{1:T};z)=p(s_1)\prod_{t=1}^{T}e^{r(s_t,a_t;z)}p(s_{t+1}\mid s_t,a_t;z)$

- 通过此分解，环境的非平稳元素由潜在变量z捕获，并且在每个任务中，奖励(reward)和动力(dynamics)函数必然是平稳的。这表明，学习推断z(相当于用z表示环境的非平稳元素)将使该强化学习设置退化为平稳设置。采用这种方法很有吸引力，因为标准强化学习设置已经存在了大量可利用的算法。下一页将描述如何通过在变分推断框架下推导该模型的证据下界($L(q)$)来近似z上的后验值。

➢非平稳环境下的异策略(off-policy)强化学习

$$L(q)=\int q(\theta)\ln\frac{p(X,\theta)}{q(\theta)}d\theta=\int q(\theta)\ln\frac{p(X\mid\theta)p(\theta)}{q(\theta)}d\theta$$
$$=\int q(\theta)\ln p(X\mid\theta)d\theta+\int q(\theta)\ln\frac{p(\theta)}{q(\theta)}d\theta$$
$$=\mathbb{E}_q\left[\ln p(X\mid\theta)\right]-KL(q(\theta)\|p(\theta))$$

- 基于变分推断的联合表示和强化学习

回合1:i-1的轨迹 $\tau^{1:i-1}=\{\tau^1,\ldots,\tau^{i-1}\}$,其中 $\tau=(s_1,a_1,r_1,\ldots,s_T,a_T,r_T)$

- 给定动作序列$u=a_{1:T}$

$$\ln p(\tau^{1:i-1},\mathcal{O}^i_{1:T}=1\mid u^{1:i-1})=\ln\left(p(\mathcal{O}^i_{1:T}=1\mid\tau^{1:i-1})p(\tau^{1:i-1}\mid u^{1:i-1})\right)=\ln p(\tau^{1:i-1}\mid u^{1:i-1})+\ln p(\mathcal{O}^i_{1:T}=1\mid\tau^{1:i-1})$$

- 对于第一项

$$\ln p(\tau^{1:i-1}\mid u^{1:i-1})=\ln\int p(\tau^{1:i-1},z^{1:i-1}\mid u^{1:i-1})dz=\ln\int q(z^{1:i-1})\frac{p(\tau^{1:i-1},z^{1:i-1}\mid u^{1:i-1})}{q(z^{1:i-1})}dz=\ln\mathbb{E}_q\left[\frac{p(\tau^{1:i-1},z^{1:i-1}\mid u^{1:i-1})}{q(z^{1:i-1})}\right]=\ln\mathbb{E}_q\left[\frac{p(\tau^{1:i-1}\mid u^{1:i-1},z^{1:i-1})p(z^{1:i-1})}{q(z^{1:i-1})}\right]$$

$$\geq\mathbb{E}_q\left[\ln\frac{p(\tau^{1:i-1}\mid u^{1:i-1},z^{1:i-1})p(z^{1:i-1})}{q(z^{1:i-1})}\right]=\mathbb{E}_q\left[\ln p(\tau^{1:i-1}\mid u^{1:i-1},z^{1:i-1})\right]-KL(q(z^{1:i-1})\|p(z^{1:i-1}))$$

Jensen不等式得到第一项的证据下界，注意对数函数是凹函数

$$=\mathbb{E}_q\left[\sum_{j=1}^{i}\ln\left(p(s_1^j)\prod_{t=1}^{T}p(s_{t+1}^j,r_t^j\mid s_t^j,a_t^j,z^j)\right)\right]-\mathbb{E}_q\left[\sum_{j=1}^{i}\ln\frac{q(z^j\mid\tau^j)}{p(z^j\mid z^{1:j-1})}\right]$$

$$\Leftrightarrow\mathbb{E}_q\left[\sum_{j=1}^{i}\left(\sum_{t=1}^{T}\ln p(s_{t+1}^j,r_t^j\mid s_t^j,a_t^j,z^j)-\ln\frac{q(z^j\mid\tau^j)}{p(z^j\mid z^{1:j-1})}\right)\right]=\mathcal{L}_{rep}$$

model dynamics & reward

model latent shifts

$$\therefore\max\mathcal{L}_{rep}=-(\mathcal{L}_{dec}+\mathcal{L}_{KL})=-\left(-\mathbb{E}_q\left[\sum_{j=1}^{i}\sum_{t=1}^{T}\ln p(s_{t+1}^j,r_t^j\mid s_t^j,a_t^j,z^j)\right]+\mathbb{E}_q\left[\sum_{j=1}^{i}\ln\frac{q(z^j\mid\tau^j)}{p(z^j\mid z^{1:j-1})}\right]\right)$$

➢非平稳环境下的异策略(off-policy)强化学习

- 基于变分推断的联合表示和强化学习
- 对于第二项

$$\ln p(\mathcal{O}^i_{1:T}=1\,|\,\tau^{1:i-1}) = \ln \int p(\mathcal{O}^i_{1:T}=1, z^i\,|\,\tau^{1:i-1})dz^i = \ln \int p(\mathcal{O}^i_{1:T}=1\,|\,z^i)p(z^i\,|\,\tau^{1:i-1})dz^i$$

$$\geq \mathbb{E}_{p(z^i|\tau^{1:i-1})}\left[\ln p(\mathcal{O}^i_{1:T}=1\,|\,z^i)\right]$$

Jensen不等式

证据下界

$$\geq \mathbb{E}_{p(z^i|\tau^{1:i-1})}\mathbb{E}_{\pi(a_t|s_t,z^i)}\left[\sum_{t=1}^{T}\left(r(s_t,a_t;z^i)-\ln\pi(a_t\,|\,s_t,z^i)\right)\right]=\mathcal{L}_{RL}$$

entropy-regularized RL

- 将最大化$\mathcal{L}_{\text{RL}}$拓展成软演员评论员算法(Soft Actor-Critic, SAC)，实现了最大熵异策略强化学习。

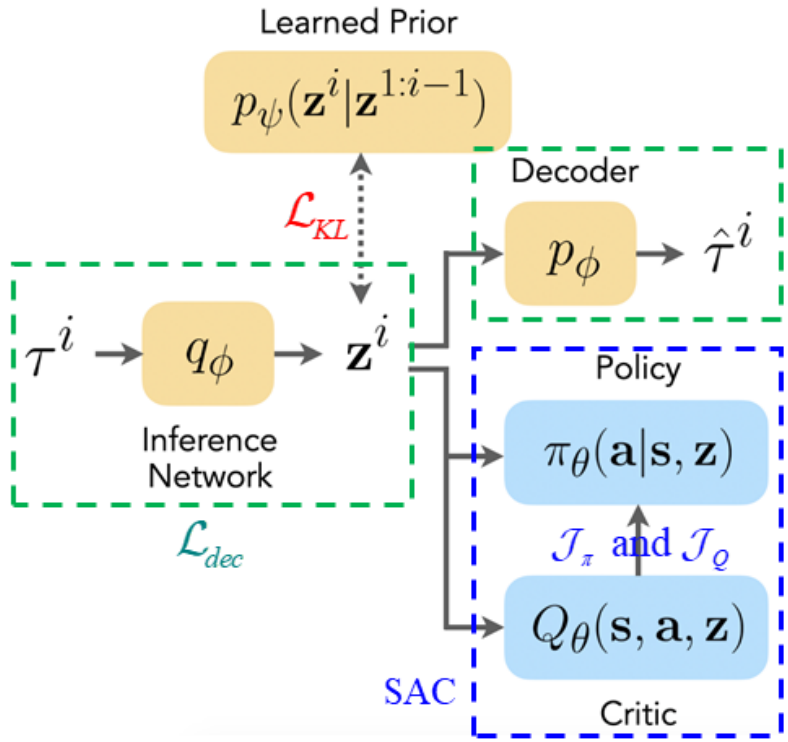
$$\begin{cases} Actor\ Loss: \mathcal{J}_\pi = \mathbb{E}_{\tau\sim\mathcal{D}}\mathbb{E}_{z\sim q(\cdot|\tau)}\left[KL\left(\pi(a\,|\,s,z)\,\|\,\dfrac{\exp(Q(s,a,z))}{Z(s)}\right)\right] \\ Critic\ Loss: \mathcal{J}_Q = \mathbb{E}_{\tau\sim\mathcal{D}}\mathbb{E}_{z\sim q(\cdot|\tau)}\left[\left(Q(s,a,z)-(r+V(s',z))\right)^2\right] \end{cases}$$

- 总体目标函数

$$\min\ \mathcal{L}=\mathcal{L}_{dec}+\mathcal{L}_{KL}+\mathcal{J}_\pi+\mathcal{J}_Q$$

➢非平稳环境下的异策略(off-policy)强化学习

- Lifelong Latent Actor-Critic (LILAC)算法流程

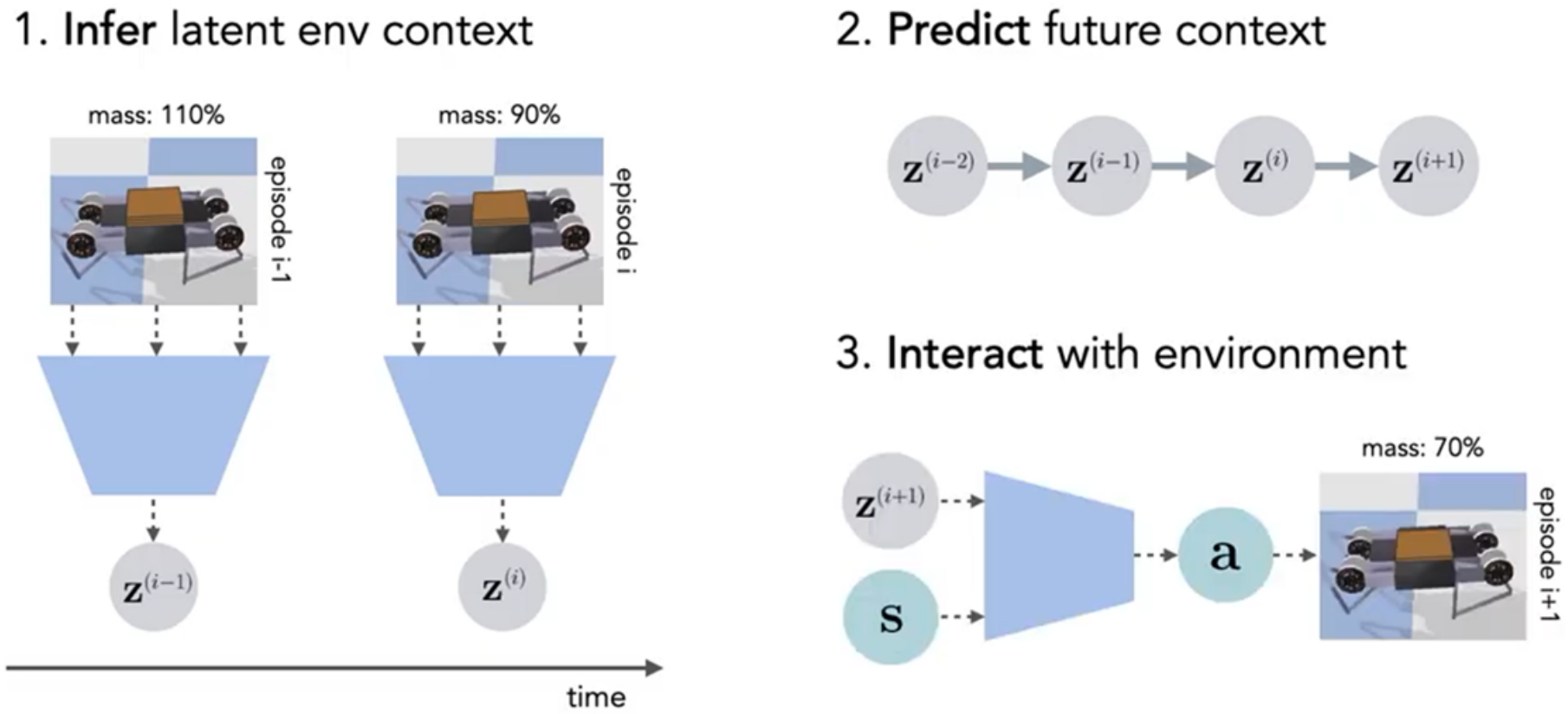


方法由演员π、评论员Q、推断网络q、解码器网络p_ψ和潜在嵌入的学习先验p_ϕ组成。每个部分都由一个神经网络实现。在执行时，演员与评论员都将之前学到的潜在变量z作为输入。重构损失和先验学习为推断网络提供了额外的学习监督。

Input: env, α_Q, α_π, α_{enc}, α_{dec}, α_ψ
Randomly initialize θ_Q, θ_π, ϕ_{enc}, ϕ_{dec}, and ψ
Initialize empty replay buffer $\mathcal{D}$
Assign $\mathbf{z}^1 \leftarrow \vec{0}$
for i = 1, 2, . . . **do**
  Sample $\mathbf{z}^i \sim p_\psi(\mathbf{z}^i|\mathbf{z}^{1:i-1})$
  Collect trajectory τ^i from env with $\pi_\theta(\mathbf{a}|\mathbf{s},\mathbf{z})$
  Update replay buffer $\mathcal{D}[i] \leftarrow \tau^i$
  for j = 1, 2, . . . , N **do**
    Sample a batch of episodes E from $\mathcal{D}$
    ▷ Update actor and critic
    $\theta_Q \leftarrow \theta_Q - \alpha_Q \nabla_{\theta_Q} \mathcal{J}_Q$
    $\theta_\pi \leftarrow \theta_\pi - \alpha_\pi \nabla_{\theta_\pi} \mathcal{J}_\pi$
    ▷ Update inference network
    $\phi_{\text{enc}} \leftarrow \phi_{\text{enc}} - \alpha_{\text{enc}} \nabla_{\phi_{\text{enc}}} (\mathcal{J}_{\text{dec}} + \mathcal{J}_{\text{KL}} + \mathcal{J}_Q)$
    ▷ Update model
    $\phi_{\text{dec}} \leftarrow \phi_{\text{dec}} - \alpha_{\text{dec}} \nabla_{\phi_{\text{dec}}} \mathcal{J}_{\text{dec}}$
    $\psi \leftarrow \psi - \alpha_\psi \nabla_\psi \mathcal{J}_{\text{KL}}$
  end for
end for

➢非平稳环境下的异策略(off-policy)强化学习

- Lifelong Latent Actor-Critic (LILAC)算法流程



4. 参考文献

[1] Annie Xie, James Harrison, Chelsea Finn. Deep Reinforcement Learning amidst Continual/Lifelong Structured Non-Stationarity. ICML, 2021.

Paper and Code: http://proceedings.mlr.press/v139/xie21c.html

Video: https://icml.cc/virtual/2021/spotlight/10468 or https://slideslive.com/38931572/deep-reinforcement-learning-amidst-lifelong-nonstationarity?ref=speaker-38044-latest or https://www.youtube.com/watch?v=FyKPO7LV06I

Experiments: https://sites.google.com/stanford.edu/lilac/

[2] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, Sergey Levine. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. ICML, 2018. http://proceedings.mlr.press/v80/haarnoja18b