WDS

DataScience@Web

第13部分

演员-评论家方法

吴博

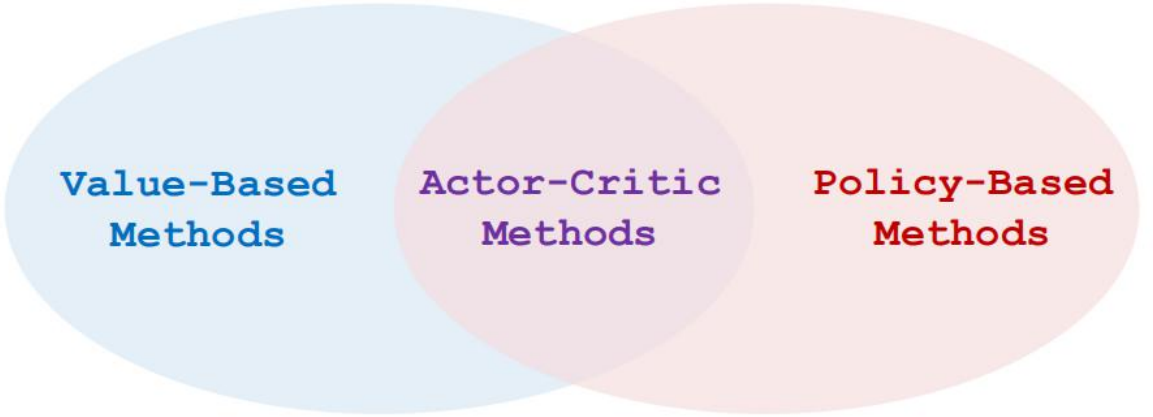
2021.07.27

2021 WDS暑期讨论班

目录

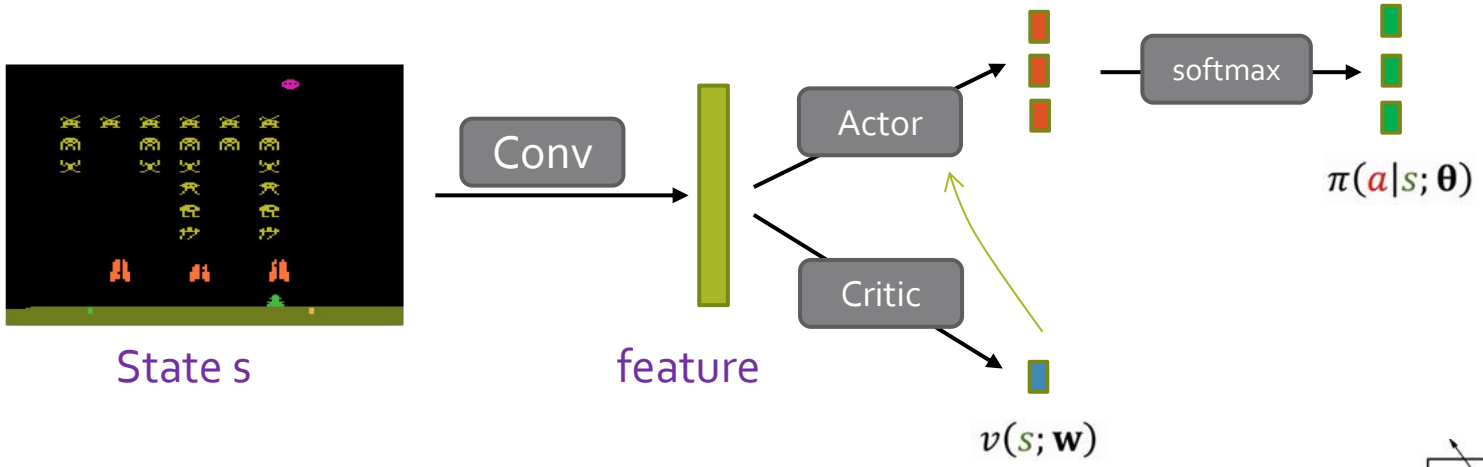
- AC方法概述
- REINFORCE with baseline
- Advantage Actor-Critic method.
- Soft Actor Critic
- 小结

AC方法概述

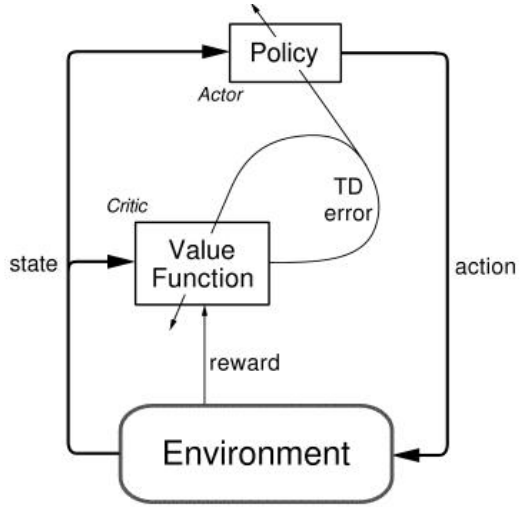


- Value Based：动作价值函数 $Q_\pi(s_t, a_t) = \mathbb{E}(U_t|s_t, a_t)$
- Policy based：策略网络 $\pi(a|s;\theta)$
- Actor critic：结合了策略梯度和时序差分的方法

AC方法概述



- Actor：策略网络$\pi(a|s;\theta)$, 负责选择动作。
- Critic：价值网络$v(s_t;\omega)$，负责评估当前的状态，基于当前的状态评估我们操作的得分，参与到策略网络更新中。
- 训练目标：让actor的累加的reward期望值尽可能大，让critic的评分越来越准。



目录

PG with baseline

- 状态价值函数：$V_\pi(s) = \mathbb{E}_{A\sim\pi}[Q_\pi(s,A)] = \sum_a \pi(a|s;\theta)\cdot Q_\pi(s,a)$
- 策略梯度：$\frac{\partial V_\pi(s)}{\partial \theta} = \mathbb{E}_{A\sim\pi}[\frac{\partial \ln \pi(A|s;\theta)}{\partial \theta}\cdot Q_\pi(s,A)]$
- Baseline b：如果baseline b是独立于A的

$$\begin{aligned}
&\bullet\ \mathbb{E}_{A\sim\pi}\left[b\cdot\frac{\partial \ln \pi(A\mid s;\mathbf{\theta})}{\partial\,\mathbf{\theta}}\right] = b\cdot\mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A\mid s;\mathbf{\theta})}{\partial\,\mathbf{\theta}}\right]\\
&= b\cdot\sum_a \pi(a|s;\mathbf{\theta})\cdot\left[\frac{1}{\pi(a\mid s;\mathbf{\theta})}\cdot\frac{\partial\,\pi(a\mid s;\mathbf{\theta})}{\partial\,\mathbf{\theta}}\right]\\
&= b\cdot\sum_a \frac{\partial\,\pi(a\mid s;\mathbf{\theta})}{\partial\,\mathbf{\theta}}\\
&= b\cdot\frac{\partial\sum_a \pi(a\mid s;\mathbf{\theta})}{\partial\,\mathbf{\theta}}\\
&= b\cdot\frac{\partial\,1}{\partial\,\mathbf{\theta}} = 0.
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial\,V_\pi(s)}{\partial\,\mathbf{\theta}}\\
&= \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A\mid s;\mathbf{\theta})}{\partial\,\mathbf{\theta}}\cdot Q_\pi(s,A)\right] - \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A\mid s;\mathbf{\theta})}{\partial\,\mathbf{\theta}}\cdot b\right]\\
&= \mathbb{E}_{A\sim\pi}\left[\frac{\partial \ln \pi(A\mid s;\mathbf{\theta})}{\partial\,\mathbf{\theta}}\cdot\left(Q_\pi(s,A)-b\right)\right].
\end{aligned}$$

PG with baseline

- 蒙特卡洛近似：

  $=g(\mathrm{A}_t)$

  - $\frac{\partial V_\pi(s)}{\partial \theta} = \mathbb{E}_{\mathrm{A}_t \sim \pi}\left[\frac{\partial \ln \pi(\mathrm{A}_t|s;\theta)}{\partial \theta} \cdot (Q_\pi(s,\mathrm{A}_t) - \mathrm{b})\right]$
  - 随机抽样：$a_t \sim \pi(\cdot|s_t;\theta)$ ，计算$\mathrm{g}(a_t)=\frac{\partial \ln \pi(a_t|s;\theta)}{\partial \theta} \cdot (Q_\pi(s,a_t) - \mathrm{b})$
  - 随机梯度上升：$\theta \leftarrow \theta + \beta \cdot \mathrm{g}(a_t)$

- baseline的选取

  - b= $V_\pi(s_t) = \mathbb{E}_{\mathrm{A}_t}[Q_\pi(s_t,A_t)]$ ：s_t是已经观测到的值，所以b是独立于A_t的，$V_\pi(s_t)$和$Q_\pi(s_t,A_t)$比较接近

REINFORCE with baseline

- 随机策略梯度： $g(a_t)=\frac{\partial \ln \pi(a_t|s_t;\theta)}{\partial \theta}\cdot(Q_\pi(s_t,a_t)-V_\pi(s_t))$
- 蒙特卡洛近似得到$Q_\pi(s_t,a_t)$:
  - 观测一条trajectory: $s_t,a_t,r_t,s_{t+1},a_{t+1},r_{t+1},\dots,s_n,a_n,r_n$
  - $Q_\pi(s_t,a_t)\approx U_t=\sum_{i=t}^{n}\gamma^{i-t}\cdot r_i$
- 用价值网络近似得到$V_\pi(s_t;\theta)$： $v_\pi(s_t)\approx V_\pi(s_t;\omega)$
- $g(a_t)\approx=\frac{\partial \ln \pi(a_t|s_t;\theta)}{\partial \theta}\cdot(U_t-v_\pi(s_t;\omega))$

REINFORCE with baseline

- 更新步骤
  - 玩一局游戏，观测一条trajectory: $s_t, a_t, r_t, s_{t+1}, a_{t+1}, r_{t+1}, \dots, s_n, a_n, r_n$
  - 计算： $\mathcal{U}_t = \sum_{i=t}^{n} \gamma^{i-t} \cdot r_i$ 和 $\delta_t = v(s_t; \omega) - \mathcal{U}_t$
  - 更新策略网络： $\theta \leftarrow \theta - \beta \cdot \delta_t \cdot \frac{\partial \ln \pi(a_t|s_t;\theta)}{\partial \theta}$
  - 更新价值网络： $\omega \leftarrow \omega - \partial \cdot \delta_t \cdot \frac{\partial v(s_t;\omega)}{\partial \omega}$

重复n次：t=1,2,3,…,n

目录

Advantage Actor-Critic(A2C)

$$g(a_t)=\frac{\partial \ln \pi(a_t|s_t;\theta)}{\partial \theta}\cdot(Q_\pi(s_t,a_t)-V_\pi(s_t))$$

观测一条trajectory: $s_t, a_t, r_t, s_{t+1}, a_{t+1}, r_{t+1}, \dots, s_n, a_n, r_n$

$Q_\pi(s_t,a_t) \approx U_t = \sum_{i=t}^{n} \gamma^{i-t} \cdot r_i$

用价值网络近似得到$V_\pi(s_t;\theta)$： $v_\pi(s_t) \approx V_\pi(s_t;\omega)$

- $V_\pi(s_t) = \mathbb{E}_{A\sim\pi}[Q_\pi(s_t,A)]$

- $Q_\pi(s_t,a_t) = \mathbb{E}(U_t|s_t,a_t) = \mathbb{E}_{s_{t+1},A_{t+1}}[R_t+\gamma\cdot Q_\pi(S_{t+1},A_{t+1})]$

  $= \mathbb{E}_{s_{t+1}}[R_t+\gamma\cdot \mathbb{E}_{A_{t+1}}[Q_\pi(S_{t+1},A_{t+1})]]$

  $= \mathbb{E}_{s_{t+1}}[R_t+\gamma\cdot V_\pi(S_{t+1})]$

观测一次Transition (s_t, a_t, r_t, s_{t+1})

$Q_\pi(s_t,a_t) \approx r_t+\gamma\cdot V_\pi(s_{t+1})$

- $V_\pi(s_t) = \mathbb{E}_{A\sim\pi}[Q_\pi(s_t,A)] = \mathbb{E}_{A_t,S_{t+1}}[R_t+\gamma\cdot V_\pi(S_{t+1})]$

Advantage Actor-Critic(A2C)

- $g(a_t)=\frac{\partial \ln \pi(a_t|s_t;\theta)}{\partial \theta}\cdot\left(r_t+\gamma\cdot v\left(s_{t+1};\omega\right)-v\left(s_t;\omega\right)\right)$

Advantege

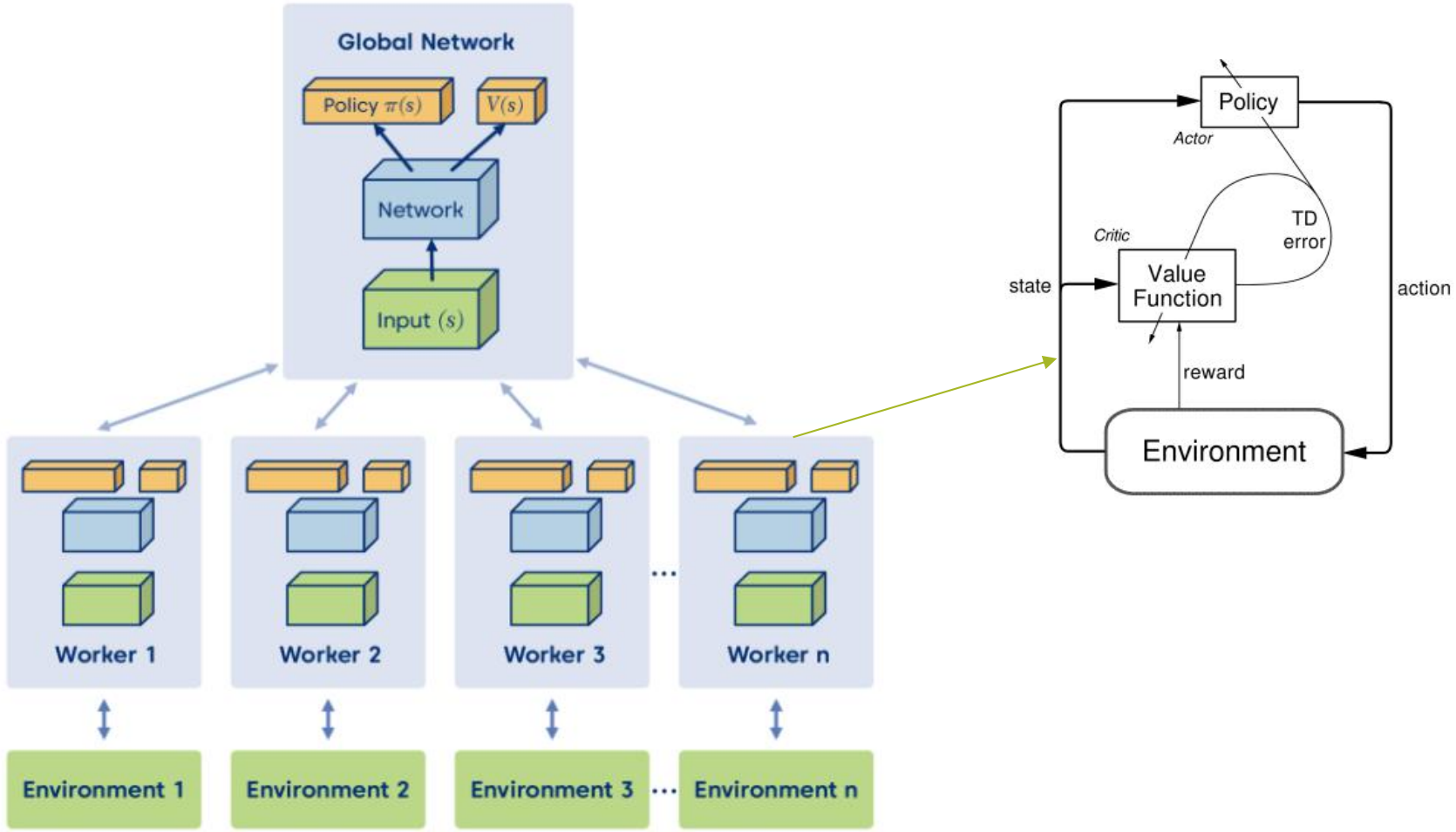
- A2c更新步骤：
  - 观测一条Transition:：(s_t,a_t,r_t,s_{t+1})
  - 计算TD target：$y_t=r_t+\gamma\cdot v\left(s_{t+1};\omega\right)$
  - 计算TD error：$\delta_t=v\left(s_t;\omega\right)-y_t$
  - 更新策略网络：$\theta\leftarrow\theta-\beta\cdot\delta_t\cdot\frac{\partial \ln \pi(a_t|s_t;\theta)}{\partial\theta}$
  - 更新价值网络：$\omega\leftarrow\omega-\partial\cdot\delta_t\cdot\frac{\partial v(s_t;\omega)}{\partial\omega}$

- 观测由t时刻到t+m-1的一条轨迹

  $\{(s_{t+i},a_{t+i},r_{t+i},s_{t+i+1})\}_{i=0}^{m-1}$

- $y_t=\sum_{i=0}^{m-1}\gamma^i r_{t+i}+\gamma^m\cdot v\left(s_{t+m};\omega\right)$

Asynchronous Advantage Actor-Critic(A3C)



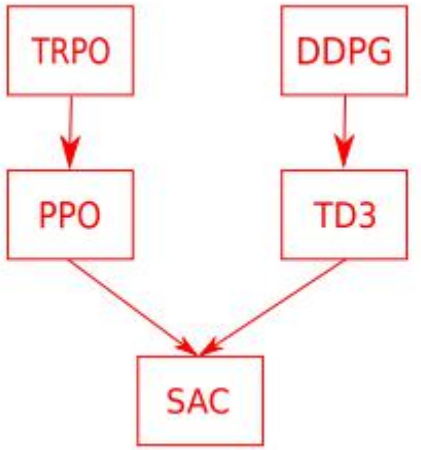
Asynchronous Advantage Actor-Critic(A3C)

Algorithm S3 Asynchronous advantage actor-critic - pseudocode for each actor-learner thread.

// Assume global shared parameter vectors θ *and* θ_v *and global shared counter* $T = 0$
// Assume thread-specific parameter vectors θ' *and* θ'_v
Initialize thread step counter $t \leftarrow 1$
repeat
  Reset gradients: $d\theta \leftarrow 0$ and $d\theta_v \leftarrow 0$.
  Synchronize thread-specific parameters $\theta' = \theta$ and $\theta'_v = \theta_v$
  $t_{start} = t$
  Get state s_t
  repeat
    Perform a_t according to policy $\pi(a_t|s_t;\theta')$
    Receive reward r_t and new state s_{t+1}
    $t \leftarrow t + 1$
    $T \leftarrow T + 1$
  until terminal s_t **or** $t - t_{start} == t_{max}$
  $R = \begin{cases} 0 & \text{for terminal } s_t \\ V(s_t, \theta'_v) & \text{for non-terminal } s_t \text{// Bootstrap from last state} \end{cases}$
  for $i \in \{t-1, \ldots, t_{start}\}$ **do**
    $R \leftarrow r_i + \gamma R$
    Accumulate gradients wrt θ': $d\theta \leftarrow d\theta + \nabla_{\theta'} \log \pi(a_i|s_i;\theta')(R - V(s_i;\theta'_v))$
    Accumulate gradients wrt θ'_v: $d\theta_v \leftarrow d\theta_v + \partial \left(R - V(s_i;\theta'_v)\right)^2 / \partial \theta'_v$
  end for
  Perform asynchronous update of θ using $d\theta$ and of θ_v using $d\theta_v$.
until $T > T_{max}$

目录

Soft Actor Critic



- A3C、TRPO和PPO方法： 随机策略, On-Policy , low sample efficiency, stable
- DDPG和TD3方法：确定性策略, replay buffer, *better Sample efficiency* ,unstable
- SAC： 随机策略, replay buffer + entropy regularization, stable and sample efficient

Soft Actor Critic

Theorem 1 (Soft Policy Iteration). *Repeated application of soft policy evaluation and soft policy improvement from any* $\pi \in \Pi$ *converges to a policy* π^* *such that* $Q^{\pi^*}(\mathbf{s}_t, \mathbf{a}_t) \geq Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t)$ *for all* $\pi \in \Pi$ *and* $(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{S} \times \mathcal{A}$, *assuming* $|\mathcal{A}| < \infty$.

- Soft : entropy

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

$$J(\pi) = \sum_{t=0}^{\infty} \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_\pi} \left[\sum_{l=t}^{\infty} \gamma^{l-t} \, \mathbb{E}_{\mathbf{s}_l \sim p, \mathbf{a}_l \sim \pi} \left[r(\mathbf{s}_t, \mathbf{a}_t) + \alpha \mathcal{H}(\pi(\cdot|\mathbf{s}_t)) | \mathbf{s}_t, \mathbf{a}_t \right] \right].$$

- Soft Policy evaluation(Critic):

$$\mathcal{T}^\pi Q(\mathbf{s}_t, \mathbf{a}_t) \triangleq r(\mathbf{s}_t, \mathbf{a}_t) + \gamma \, \mathbb{E}_{\mathbf{s}_{t+1} \sim p} \left[V(\mathbf{s}_{t+1}) \right]$$

$$V(\mathbf{s}_t) = \mathbb{E}_{\mathbf{a}_t \sim \pi} \left[Q(\mathbf{s}_t, \mathbf{a}_t) - \log \pi(\mathbf{a}_t|\mathbf{s}_t) \right]$$

- Soft Policy improvement(Actor):

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

Soft Actor Critic

- State value function: $V_\psi(\mathbf{s}_t)$ soft Q function: $Q_\theta(\mathbf{s}_t, \mathbf{a}_t)$ a tractable policy : $\pi_\phi(\mathbf{a}_t|\mathbf{s}_t)$
- Soft value function update: MSE
  - 目标函数：$J_V(\psi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}} \left[\frac{1}{2} \left(V_\psi(\mathbf{s}_t) - \mathbb{E}_{\mathbf{a}_t \sim \pi_\phi} \left[Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - \log \pi_\phi(\mathbf{a}_t|\mathbf{s}_t) \right] \right)^2 \right]$
  - 无偏梯度估计：$\hat{\nabla}_\psi J_V(\psi) = \nabla_\psi V_\psi(\mathbf{s}_t) \left(V_\psi(\mathbf{s}_t) - Q_\theta(\mathbf{s}_t, \mathbf{a}_t) + \log \pi_\phi(\mathbf{a}_t|\mathbf{s}_t) \right)$
- Soft Q function : soft Bellman residual MSE
  - 目标函数：$J_Q(\theta) = \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \mathcal{D}} \left[\frac{1}{2} \left(Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - \hat{Q}(\mathbf{s}_t, \mathbf{a}_t) \right)^2 \right]$

    $\hat{Q}(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{\mathbf{s}_{t+1} \sim p} \left[V_{\bar{\psi}}(\mathbf{s}_{t+1}) \right]$
  - 无偏梯度估计：$\hat{\nabla}_\theta J_Q(\theta) = \nabla_\theta Q_\theta(\mathbf{a}_t, \mathbf{s}_t) \left(Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - r(\mathbf{s}_t, \mathbf{a}_t) - \gamma V_{\bar{\psi}}(\mathbf{s}_{t+1}) \right)$

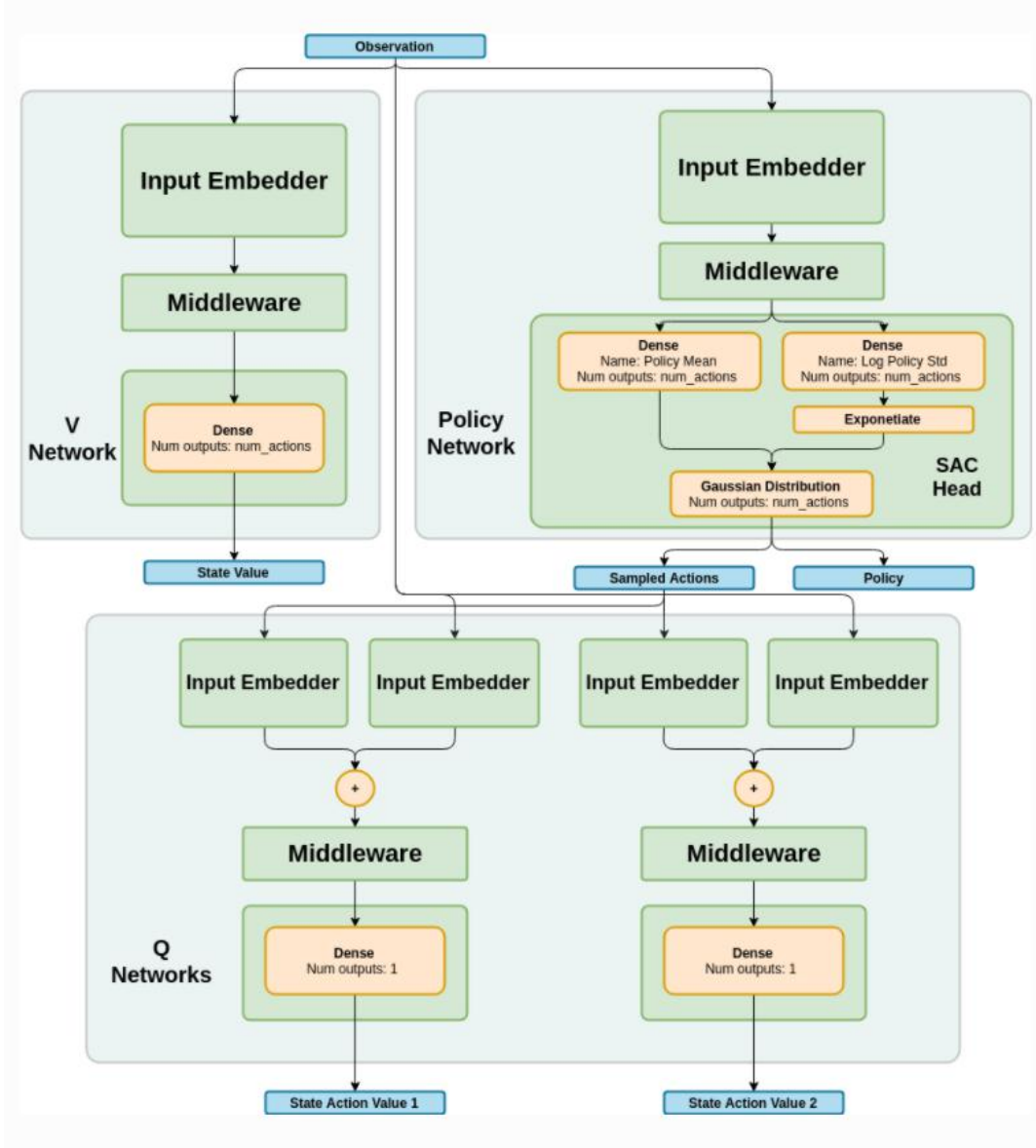
Soft Actor Critic

- Policy improvement (actor):
  - 目标函数： $J_\pi(\phi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}} \left[\mathrm{D}_{\mathrm{KL}} \left(\pi_\phi(\cdot|\mathbf{s}_t) \,\middle\|\, \frac{\exp\left(Q_\theta(\mathbf{s}_t, \cdot)\right)}{Z_\theta(\mathbf{s}_t)} \right) \right]$
  - Action reparameterize trick: $\mathbf{a}_t = f_\phi(\epsilon_t; \mathbf{s}_t),$
  - 目标函数: $J_\pi(\phi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}, \epsilon_t \sim \mathcal{N}} \left[\log \pi_\phi(f_\phi(\epsilon_t; \mathbf{s}_t)|\mathbf{s}_t) - Q_\theta(\mathbf{s}_t, f_\phi(\epsilon_t; \mathbf{s}_t)) \right]$
  - 梯度估计： $\hat{\nabla}_\phi J_\pi(\phi) = \nabla_\phi \log \pi_\phi(\mathbf{a}_t|\mathbf{s}_t) + \left(\nabla_{\mathbf{a}_t} \log \pi_\phi(\mathbf{a}_t|\mathbf{s}_t) - \nabla_{\mathbf{a}_t} Q(\mathbf{s}_t, \mathbf{a}_t) \right) \nabla_\phi f_\phi(\epsilon_t; \mathbf{s}_t)$

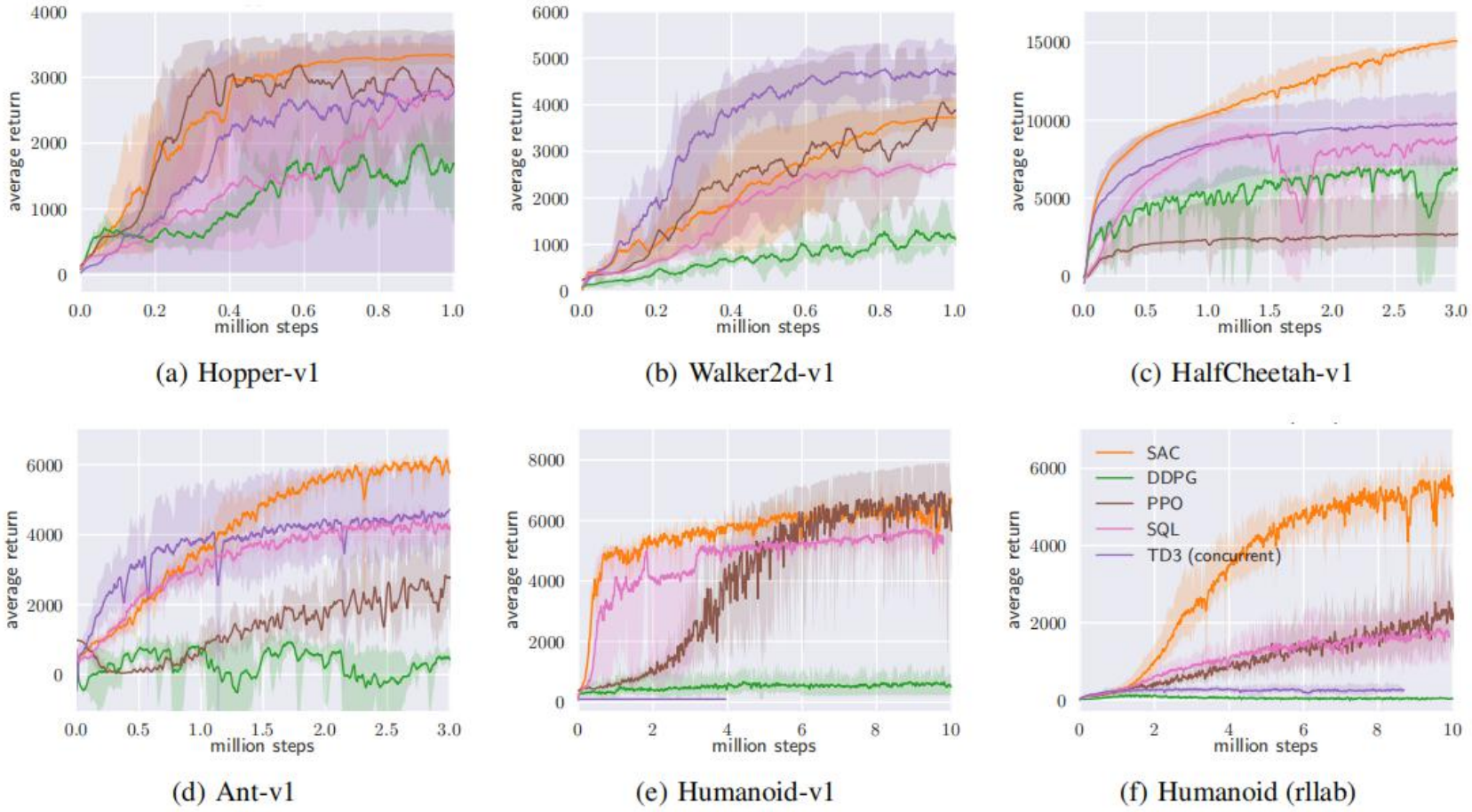
Soft Actor Critic

Algorithm 1 Soft Actor-Critic

Initialize parameter vectors $\psi, \bar{\psi}, \theta, \phi$.
for each iteration **do**
  for each environment step **do**
    $\mathbf{a}_t \sim \pi_\phi(\mathbf{a}_t|\mathbf{s}_t)$
    $\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})\}$
  end for
  for each gradient step **do**
    $\psi \leftarrow \psi - \lambda_V \hat{\nabla}_\psi J_V(\psi)$
    $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$
    $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$
    $\bar{\psi} \leftarrow \tau\psi + (1-\tau)\bar{\psi}$
  end for
end for



Soft Actor Critic



总结：

- REINFORCE with baseline : 是策略梯度的Monte carlo版本，总收益是从整个轨迹中采样的，优势函数是采样得到的总收益和减去*baseline*。通过策略梯度上升训练策略网络。
- Actor-Critic: 是策略梯度的TD版本，优势函数是TD error。通过训练策略网络作为Actor，价值网络作为Critic。Actor 用于做出决策，通过策略梯度方法进行学习。Critic通过计算价值函数来评估参与者的行动。
- A3C：异步的A2C，利用多个CPU开启多个worker，独立对环境进行探索，各自使用计算得到的梯度去异步更新global network。
- Soft actor critic : 引入了最大熵的actor-critic方法，off policy。兼顾了稳定性和样本利用率高的优点。

Thanks!

参考文献：

- 深度强化学习 王树森
  https://www.bilibili.com/video/BV1204y197US?from=search&seid=18123095896934228404
- A3C：https://arxiv.org/pdf/1602.01783.pdf
- A3c李宏毅强化学习笔记：https://datawhalechina.github.io/easy-rl/#/chapter9/chapter9?id=a3c
- SAC：http://arxiv.org/abs/1801.01290
- SAC: https://zhuanlan.zhihu.com/p/70360272