

目录

1	DQN 算法	2
2	SoftQ 算法	3

1 DQN 算法

DQN 算法

初始化策略网络参数 θ
复制参数到目标网络 $\hat{Q} \leftarrow Q$
初始化经验回放 D
for 回合数 = 1, M **do**
 重置环境, 获得初始状态 s_t
 for 时步 = 1, t **do**
 根据 $\varepsilon - greedy$ 策略采样动作 a_t
 环境根据 a_t 反馈奖励 s_t 和下一个状态 s_{t+1}
 存储 transition 即 (s_t, a_t, r_t, s_{t+1}) 到经验回放 D 中
 更新环境状态 $s_{t+1} \leftarrow s_t$
 更新策略:
 从 D 中采样一个 batch 的 transition
 计算实际的 Q 值, 即 $y_j = \begin{cases} r_j & \text{对于终止状态 } s_{j+1} \\ r_j + \gamma \max_{a'} Q(s_{j+1}, a'; \theta) & \text{对于非终止状态 } s_{j+1} \end{cases}$
 对损失 $(y_j - Q(s_j, a_j; \theta))^2$ 关于参数 θ 做随机梯度下降
 end for
 每 C 个回合复制参数 $\hat{Q} \leftarrow Q$ (此处也可像原论文中放到小循环中改成每 C 步, 但没有每 C 个回合稳定)
end for

2 SoftQ 算法

SoftQ 算法

初始化参数 θ 和 ϕ

复制参数 $\bar{\theta} \leftarrow \theta, \bar{\phi} \leftarrow \phi$

初始化经验回放 D

for 回合数 = 1, M **do**

for 时步 = 1, t **do**

 根据 $\mathbf{a}_t \leftarrow f^\phi(\xi; \mathbf{s}_t)$ 采样动作, 其中 $\xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

 环境根据 a_t 反馈奖励 s_t 和下一个状态 s_{t+1}

 存储 transition 即 (s_t, a_t, r_t, s_{t+1}) 到经验回放 D 中

 更新环境状态 $s_{t+1} \leftarrow s_t$

 更新 soft Q 函数参数:

 对于每个 $s_{t+1}^{(i)}$ 采样 $\{\mathbf{a}^{(i,j)}\}_{j=0}^M \sim q_{\mathbf{a}'}$

 计算 empirical soft values $V_{\text{soft}}^\theta(\mathbf{s}_t) = \alpha \log \mathbb{E}_{q_{\mathbf{a}'}} \left[\frac{\exp(\frac{1}{\alpha} Q_{\text{soft}}^\theta(s_t, \mathbf{a}'))}{q_{\mathbf{a}'}(\mathbf{a}')} \right]$

 计算 empirical gradient $J_Q(\theta) = \mathbb{E}_{\mathbf{s}_t \sim q_{s_t}, \mathbf{a}_t \sim q_{\mathbf{a}_t}} \left[\frac{1}{2} \left(\hat{Q}_{\text{soft}}^\theta(\mathbf{s}_t, \mathbf{a}_t) - Q_{\text{soft}}^\theta(\mathbf{s}_t, \mathbf{a}_t) \right)^2 \right]$

 根据 $J_Q(\theta)$ 使用 ADAM 更新参数 θ

 更新策略:

 对于每个 $s_t^{(i)}$ 采样 $\{\xi^{(i,j)}\}_{j=0}^M \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

 计算 $\mathbf{a}_t^{(i,j)} = f^\phi(\xi^{(i,j)}, \mathbf{s}_t^{(i)})$

 使用经验估计计算 $\Delta f^\phi(\cdot; \mathbf{s}_t) = \mathbb{E}_{\mathbf{a}_t \sim \pi^\phi} \left[\kappa(\mathbf{a}_t, f^\phi(\cdot; \mathbf{s}_t)) \nabla_{\mathbf{a}'} Q_{\text{soft}}^\theta(\mathbf{s}_t, \mathbf{a}') \Big|_{\mathbf{a}'=\mathbf{a}_t} + \alpha \nabla_{\mathbf{a}'} \kappa(\mathbf{a}', f^\phi(\cdot; \mathbf{s}_t)) \Big|_{\mathbf{a}'=\mathbf{a}_t} \right]$

 计算经验估计 $\frac{\partial J_\pi(\phi; \mathbf{s}_t)}{\partial \phi} \propto \mathbb{E}_\xi \left[\Delta f^\phi(\xi; \mathbf{s}_t) \frac{\partial f^\phi(\xi; \mathbf{s}_t)}{\partial \phi} \right]$, 即 $\hat{\nabla}_\phi J_\pi$

 根据 $\hat{\nabla}_\phi J_\pi$ 使用 ADAM 更新参数 ϕ

end for

 每 C 个回合复制参数 $\bar{\theta} \leftarrow \theta, \bar{\phi} \leftarrow \phi$

end for
