从DPG到DDPG

> 肖淇文, Apr/20/2024

[!important] 本文仅为大纲

I. DPG (Deterministic Policy Gradient)

易混淆符号规定:

符号及定义	含义
$r_t^\gamma = \sum_{k=t}^{\infty} \gamma^{k-t} r(s_k, a_k)$	discounted reward from time-step t
$J(\pi) = \mathbb{E}[r_1^\gamma \mid \pi]$	performance objective
$p(s \to s', t, \pi)$	density at state s' after t steps from s
$\rho^\pi(s') = \int_{\mathcal{S}} \sum_{t=1}^{\infty} \gamma^{t-1} p_1(s) p(s \to s', t, \pi) \mathrm{d}s$	discounted state distribution
$V^\pi(s) = \mathbb{E}[r_1^\gamma \mid S_1 = s; \pi]$	state value
$Q^\pi(s, a) = \mathbb{E}[r_1^\gamma \mid S_1 = s, A_1 = a; \pi]$	action value

因此有:

$$\begin{aligned} J(\pi_\theta) &= \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \pi_\theta(s, a) r(s, a) \mathrm{d}a \mathrm{d}s \\ &= \mathbb{E}_{s \sim \rho^\pi, a \sim \pi_\theta}[r(s, a)] \end{aligned}$$

I.A. Stochastic Policy Gradient Theorem

使用梯度上升最大化 $J(\pi_\theta)$ 需要求解梯度，根据随机策略梯度定理:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{s \sim \rho^\pi, a \sim \pi_\theta}[\nabla_\theta \pi_\theta(a|s) Q^\pi(s, a)] \tag{1}$$

I.B. Stochastic Actor-Critic

Actor: 由 (1) 式梯度上升更新 θ.

Critic: 参数 w 通过temporal-difference learning近似 $Q^w(s, a) \approx Q^\pi(s, a)$.

会产生bias，除非满足:

1. $Q^w(s, a) = \nabla_\theta \log \pi_\theta(a|s)^\top w$，以此来满足 (1) 式;
2. w 最小化MSE $\epsilon^2(w) = \mathbb{E}_{s \sim \rho^\pi, a \sim \pi_\theta}\left[(Q^w(s, a) - Q^\pi(s, a))^2\right]$.

对1.的证明(Sutton, 1999):

如果达到局部最优，则有

$$\int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \pi_\theta(s, a) \left[Q^\pi(s, a) - Q^w(s, a)\right] \nabla_w Q^w(s, a) \mathrm{d}a \mathrm{d}s = 0$$

此时如果满足

$$Q^w(s,a) = \nabla_\theta \log \pi_\theta(a|s)^\top w$$

则有

$$\int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \pi_\theta(s,a) \left[Q^\pi(s,a) - Q^w(s,a)\right] \nabla_\theta \log \pi_\theta(a|s) \mathrm{d}a\mathrm{d}s = 0$$

可知估计误差 $[Q^\pi(s,a) - Q^w(s,a)]$ 与策略梯度 $\nabla_\theta \pi_\theta(a|s)$ 正交。

由随机策略梯度定理:

$$\begin{aligned}
\nabla_\theta J(\pi_\theta) &= \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} {\color{red}Q^\pi(s,a)} \nabla_\theta \pi_\theta(a|s) \mathrm{d}a\mathrm{d}s \\
&- \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \pi_\theta(s,a) \left[{\color{red}Q^\pi(s,a)} - Q^w(s,a)\right] \nabla_\theta \log \pi_\theta(a|s) \mathrm{d}a\mathrm{d}s \\
&= \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} Q^w(s,a) \nabla_\theta \pi_\theta(a|s) \mathrm{d}a\mathrm{d}s
\end{aligned}$$

更直观地说，估计函数 $Q^w(s,a)$ 与策略梯度在features上呈线性关系。

对于条件2，实际应用中可能不会去满足，而是使用temporal-difference learning来提高效率。

I.C. Off-Policy Actor-Critic

Off-policy: $\beta(a|s) \neq \pi_\theta(a|s)$.

Performance objective一般被更改为目标策略 π 在行为策略 β 的状态分布上的平均价值:

$$\begin{aligned}
J_\beta(\pi_\theta) &= \int_{\mathcal{S}} \rho^\beta(s) V^\pi(s) \mathrm{d}s \\
&= \int_{\mathcal{S}} \int_{\mathcal{A}} \rho^\beta(s) \pi_\theta(a|s) Q^\pi(s,a) \mathrm{d}a\mathrm{d}s
\end{aligned}$$

再求梯度:

$$\begin{aligned}
\nabla_\theta J_\beta(\pi_\theta) &\approx \int_{\mathcal{S}} \int_{\mathcal{A}} \rho^\beta(s) \nabla_\theta \pi_\theta(a|s) Q^\pi(s,a) \mathrm{d}a\mathrm{d}s \\
&= \mathbb{E}_{s\sim\rho^\beta}\left[\int_{\mathcal{A}} \beta_\theta(a|s) \frac{\pi_\theta(a|s)}{\beta_\theta(a|s)} \nabla_\theta \log \pi_\theta(a|s) Q^\pi(s,a) \mathrm{d}a\right] \\
&= \mathbb{E}_{s\sim\rho^\beta, a\sim\beta}\left[\frac{\pi_\theta(a|s)}{\beta_\theta(a|s)} \nabla_\theta \log \pi_\theta(a|s) Q^\pi(s,a)\right]
\end{aligned}$$

丢掉了 $\nabla_\theta Q^\pi(s,a)$（因为难以估计），但是可以保留local optima.

为什么？证明如下（来自Degris et al., Off-Policy Actor-Critic, 2012）:

定理：对于任何策略参数 θ，令

$$\theta' = \theta + \alpha \int_{\mathcal{S}} \int_{\mathcal{A}} \rho^\beta(s) \nabla_\theta \pi_\theta(a|s) Q^\pi(s,a) \mathrm{d}a\mathrm{d}s$$

则存在 $\epsilon > 0$ 使得 $\forall \alpha < \epsilon$,

$$J_\beta(\pi_{\theta'}) \geq J_\beta(\pi_\theta)$$

具体证明过程略。由此可知这种梯度上升方式是可行的。

在Off-Policy Actor-Critic架构中，Critic的目标进而改为估计 $V^v(s) \approx V^\pi(s)$.

由于 $Q^\pi(s,a)$ 未知，使用temporal-difference error $\delta_t = r_{t+1} + \gamma V^v(s_{t+1}) - V^v(s_t)$，可证明是真实梯度的近似。

Actor和Critic都使用importance sampling ratio $\dfrac{\pi_\theta(a|s)}{\beta_\theta(a|s)}$ 来调整使得actions是根据 π 来选择而不是 β.

I.D. Gradients of Deterministic Policies

确定性策略用 μ 表示。

如果直接令 $\mu^{k+1}(s) = \arg\max_a Q^{\mu^k}(s,a)$，在连续动作空间里每一步都需要求最大值，不方便。

替代方案是使策略以 ∇Q 的方向移动，即

$$\theta^{k+1} = \theta^k + \alpha \mathbb{E}_{s\sim\rho^{\mu^k}} \left[\nabla_\theta Q^{\mu^k}(s, \mu_\theta(s)) \right]$$

可以看到对于策略的提升可被分解为action value对action的梯度以及策略对策略参数的梯度：

$$\theta^{k+1} = \theta^k + \alpha \mathbb{E}_{s\sim\rho^{\mu^k}, a=\mu_\theta(s)} \left[\nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu^k}(s, a) \right]$$

DPG定理证明上面的更新就是performance objective的梯度方向，即

$$\nabla_\theta J(\mu_\theta) = \mathbb{E}_{s\sim\rho^{\mu^k}, a=\mu_\theta(s)} \left[\nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu^k}(s, a) \right]$$

另外文章还证明了DP是SP的极限形式。

I.D.1. On-Policy Deterministic Actor-Critic

虽然是deterministic，但是在噪声较大的情况下也能保证exploration.

算法流程：

$$\begin{gathered}
\delta_t = r_t + \gamma Q^w(s_{t+1}, a_{t+1}) - Q^w(s_t, a_t) \\
w_{t+1} = w_t + \alpha_w \delta_t \nabla_w Q^w(s_t, a_t) \\
\theta_{t+1} = \theta_t + \alpha_\theta \nabla_\theta \mu_\theta(s_t) \nabla_a Q^w(s_t, a_t)|_{a=\mu_\theta(s)}
\end{gathered}$$

I.D.2. Off-Policy Deterministic Actor-Critic

跟stochastic情况相似，有

$$\nabla_\theta J_\beta(\mu_\theta) \approx \mathbb{E}_{s\sim\rho^\beta, a=\mu_\theta(s)} \left[\nabla_\theta \mu_\theta(s) \nabla_a Q^\mu(s, a) \right]$$

算法流程：

$$\begin{gathered}
\delta_t = r_t + \gamma Q^w(s_{t+1}, \mu_\theta(s_{t+1})) - Q^w(s_t, a_t) \\
w_{t+1} = w_t + \alpha_w \delta_t \nabla_w Q^w(s_t, a_t) \\
\theta_{t+1} = \theta_t + \alpha_\theta \nabla_\theta \mu_\theta(s_t) \nabla_a Q^w(s_t, a_t)|_{a=\mu_\theta(s)}
\end{gathered}$$

I.D.3. Compatible Function Approximation

相似地，$Q^w(s,a)$ 的选择需要满足下列两个条件：

1. $\nabla_a Q^w(s,a)|_{a=\mu_\theta(s)} = \nabla_\theta \mu_\theta(s)^\top w$;
2. w 最小化 $MSE(\theta, w) = \mathbb{E}[\epsilon(s;\theta,w)^\top \epsilon(s;\theta,w)]$，其中 $\epsilon(s;\theta,w) = \nabla_a Q^w(s,a)|_{a=\mu_\theta(s)} - \nabla_a Q^\mu(s,a)|_{a=\mu_\theta(s)}$.

对条件1的证明类似于stochastic的情况。

对条件2的证明：

达到最小值时，梯度为0，即

$$\begin{aligned}\nabla_w MSE(\theta, w) &= 0 \\ \mathbb{E}\left[\nabla_\theta \mu_\theta(s) \nabla_a Q^w(s,a)|_{a=\mu_\theta(s)}\right] &= \mathbb{E}\left[\nabla_\theta \mu_\theta(s) \nabla_a Q^\mu(s,a)|_{a=\mu_\theta(s)}\right] \\ &= \nabla_\theta J_{(\beta)}(\mu_\theta)\end{aligned}$$

对于任意确定性策略 $\mu_\theta(s)$，总存在approximator，形式为

$$Q^w(s,a) = (a - \mu_\theta(s))^\top \nabla_\theta \mu_\theta(s)^\top w + V^v(s)$$

其中，$V^v(s)$ 可以是任何可导且独立于 a 的基准函数，比如 $V^v(s) = v^\top \phi(s)$，$\phi(s)$ 为state features.

可以大致理解为

$$s\text{下}a\text{的动作价值} = \underbrace{\text{选}a\text{而不是}\mu_\theta(s)\text{的优势}}_{A^w(s,a)=\phi(s,a)^\top w=(a-\mu_\theta(s))^\top \nabla_\theta \mu_\theta(s)^\top w} + s\text{的状态价值}$$

这里作者都选择用线性approximator：

> We note that a linear function approximator is not very useful for predicting action-values globally, since the action value diverges to ±∞ for large actions. However, it can still be highly effective as a *local* critic.

总结下来：

- Critic 是一个线性approximator来估计action value
- Actor向Critic给出的action value梯度方向更新参数

II. DQN

用神经网络（Q-network）拟合action-value.

DQN最初提出时仍然是stochastic的情况。

Q-network在每个iteration i 的训练目标是最小化

$$L_i(\theta_i) = \mathbb{E}_{s,a\sim\rho(\cdot)}\left[(y_i - Q(s,a;\theta_i))^2\right]$$

其中

$$y_i = \mathbb{E}_{s\sim\mathcal{S}}\left[r + \gamma \max_{a'} Q(s', a'; \theta_{i-1}) | s, a\right]$$

即iteration i 的target.

需要注意到一点是训练目标包含上一个迭代的参数。

求导可得：

$$\nabla_{\theta_i} L_i(\theta_i) = \mathbb{E}_{s,a\sim\rho(\cdot);s'\in\mathcal{S}}\left[r + \gamma \max_{a'} Q(s', a'; \theta_{i-1}) - Q(s,a;\theta_i)\nabla_{\theta_i} Q(s,a;\theta_i)\right]$$

实际应用中不会去计算这个期望，而是每个时间步都根据单样本来更新。

DQN的创新点Experience Replay：可以避免strong correlation.

> By using experience replay the behavior distribution is averaged over many of its previous states, smoothing out learning and avoiding oscillations or divergence in the parameters.

III. DDPG

DDPG结合了DPG和DQN.

DQN的问题在于连续动作空间内 $\max_{a'} Q(s', a'; \theta_{i-1})$ 不易求得，所以把DQN放进deterministic的框架中。

References

[DDPG] https://arxiv.org/abs/1509.02971

[DPG] https://proceedings.mlr.press/v32/silver14.pdf

[Off-Policy AC] https://arxiv.org/pdf/1205.4839.pdf

[Func Approx] https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf

[DQN] https://www.cs.toronto.edu/~vmnih/docs/dqn.pdf