**强化学习理论学习与代码实现**

# 确定性策略梯度（Deterministic Policy Gradient Algorithms，DPG）

## 介绍

策略梯度算法广泛应用于具有连续\*\*动作空间\*\*的强化学习问题，其基本思想是通过参数化的概率分布$\pi\_{\theta}(a | s)=\mathbb{P}[a | s ; \theta]$来表达策略，该策略会根据参数$\theta$随机地选择状态$s$下的动作$a$，也就是说随机性策略梯度最终得到的是动作的\*\*概率分布\*\*，然后通过采样该随机策略，并朝着最大化累积奖励的方向优化策略参数。

而确定性策略梯度会\*\*确定性地\*\*选择一个动作$a=\mu\_{\theta}(s)$,之前人们认为确定性策略梯度不存在，或者只能在使用模型时才能得到，然而David Silver证明了确定性策略确实存在，并且它具有简单的无模型的形式——仅仅遵循动作值函数的梯度。实际上，确定性策略梯度是随机策略梯度在\*\*策略方差趋于0\*\*时的一种极限情况。

在随机情况下，策略梯度需要收集状态空间和动作空间，而确定性情况，却仅仅收集状态空间，所以随机策略梯度可能会需要更多次的采样，特别是当动作空间维度较大时。

为了保证确定性策略梯度能够和随机策略梯度一样，探索整个状态空间和动作空间（\*\*确定性策略只输出贪婪动作，导致可能有些动作永远都无法执行\*\*），有必要引入<font color='red'>离策略学习算法</font>，基本思想就是\*\*根据随机行为策略（保证充分的探索）选择动作，但最终学习的是确定性策略（利用确定性策略的效率）\*\*。通过确定性策略梯度可以推导出离策略演员-评论家算法（off-policy actorcritic algorithm ），并通过一可微分值函数逼近器估计动作值函数，然后沿着近似动作值函数梯度的方向更新策略参数。还引入了确定性策略梯度的\*\*相容函数近似\*\*(compatible function approximation)的概念，以确保该近似不会对策略梯度产生偏差。

## 背景

### 基础

MDP（马尔可夫决策过程）$\langle\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma\rangle$

奖励函数（reward function）$r: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$

策略（policy）$\pi\_{\theta}: \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$

轨迹（trajectory）$s\_{1}, a\_{1}, r\_{1} \dots, s\_{T}, a\_{T}, r\_{T}$

回报（return）$r\_{t}^{\gamma}=\sum\_{k=t}^{\infty} \gamma^{k-t} r\left(s\_{k}, a\_{k}\right)$

状态价值函数（state-value function）$V^{\pi}(s)=\mathbb{E}\left[r\_{1}^{\gamma} | S\_{1}=s ; \pi\right]$

动作价值函数（action-value function）$Q^{\pi}(s,a)=\mathbb{E}\left[r\_{1}^{\gamma} | S\_{1}=s, A\_{1}=a ; \pi\right]$

转移后的概率分布（density at state $s^{\prime}$ after transitioning from state $s$ ）$p\left(s \rightarrow s^{\prime}, t, \pi\right)$

初始状态分布（an initial state distribution with density） $p\_{1}(s)$

折扣状态分布（discounted state distribution）$\rho^{\pi}\left(s^{\prime}\right):=$ $\int\_{\mathcal{S}} \sum\_{t=1}^{\infty} \gamma^{t-1} p\_{1}(s) p\left(s \rightarrow s^{\prime}, t, \pi\right) \mathrm{d} s$

性能目标函数（performance objective）

$$\begin{aligned}

J\left(\pi\_{\theta}\right) &=\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}\tag{1}$$

### Stochastic Policy Gradient 理论

策略梯度背后的思想是沿着目标梯度$\nabla\_{\theta} J\left(\pi\_{\theta}\right)$的方向调整策略参数$\theta$，\*\*最基本的策略梯度理论\*\*如下：

$$\begin{aligned}

\nabla\_{\theta} J\left(\pi\_{\theta}\right) &=\int\_{\mathcal{S}} \rho^{\pi}(s) \int\_{\mathcal{A}} \nabla\_{\theta} \pi\_{\theta}(a | s) Q^{\pi}(s, a) \mathrm{d} a \mathrm{d} s \\

&=\mathbb{E}\_{s \sim \rho^{\pi}, a \sim \pi\_{\theta}}\left[\nabla\_{\theta} \log \pi\_{\theta}(a | s) Q^{\pi}(s, a)\right]

\end{aligned}\tag{2}$$

策略梯度理论的重要价值在于，其将目标梯度的计算简化为一个期望值的计算，通过构造对该期望的基于样本的估计，可以推导出各种不同的策略梯度算法（详见。。。。。）。\*\*但是这些算法必须要解决的问题是如何估计动作值函数$Q^{\pi}(s,a)$\*\*，这其中最简单的形式就是使用采样回报$r^{\gamma}\_t$来估计$Q^{\pi}(s\_t,a\_t)$，这就是REINFORCE算法，具体的细节在之前的文章中都有所介绍。

### 随机演员-评论家算法

actor-critic是在策略梯度中广泛使用的架构，其包括两个组件：演员通过随机梯度下降（式2）调整随机策略$\pi\_\theta(s)$的参数$\theta$，由于真实动作值函数$Q^{\pi}(s,a)$未知，所以使用带参数$w$的$Q^{w}(s,a)$替代；评论家则是使用适当的策略评估算法估计动作值函数$Q^{w}(s, a) \approx Q^{\pi}(s, a)$，如时间差分学习。具体的关于AC算法的细节，可以看这篇文章（。。。。。。。。。。）。

通常情况下，使用$Q^{w}(s,a)$代替真实的$Q^{\pi}(s,a)$可能会引入偏差，为了消除误差，$Q^{w}(s,a)$应满足以下两个条件：

1. $Q^{w}(s, a)=\nabla\_{\theta} \log \pi\_{\theta}(a | s)^{\top} w$

2. $w$需要最小化均方误差$\epsilon^{2}(w)=\mathbb{E}\_{s \sim \rho^{\pi}, a \sim \pi\_{\theta}}\left[\left(Q^{w}(s, a)-Q^{\pi}(s, a)\right)^{2}\right]$

条件1是说兼容函数逼近器是随机策略“特征”$\nabla\_{\theta} \log \pi\_{\theta}(a | s)$的线性函数，条件2是说需要保证参数$w$为线性回归问题（根据特征估计$Q^{\pi}(s,a)$）的解。在实际中，\*\*条件2通常可以放宽\*\*，以支持通过时间差分学习来更有效地估计值函数地策略评估算法。因此如果条件1和条件2都满足，那就等价于根本没有使用评论家，即更像是REINFORCE算法。

### 离策略演员-评论家(Off-Policy Actor-Critic,OffPAC)

通常根据从另一个完全不同的行为策略$\beta(a | s) \neq \pi\_{\theta}(a | s)$中采样的轨迹，来离策略地估计策略梯度是非常有用的，在离策略情况下，需要将目标函数重新定义为\*\*目标策略的值函数在行为策略的状态分布上的平均值\*\*:

$$\begin{aligned}

J\_{\beta}\left(\pi\_{\theta}\right) &=\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}$$

同归对其进行微分，可以得到离策略策略梯度的近似：

$$\begin{align}

\nabla\_{\theta} J\_{\beta}\left(\pi\_{\theta}\right) & \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 \tag {4}\\

&=\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] \tag {5}

\end{align}$$

这个近似去掉了依赖动作值函数梯度的项：$\nabla\_{\theta} Q^{\pi}(s, a)$。离策略演员-评论家算法使用行为策略$\beta(a|s)$生成轨迹，评论家使用梯度时间差分学习从这些轨迹中离策略地估计状态值函数$V^{v}(s) \approx V^{\pi}(s)$，演员也是从这些轨迹中，根据随机梯度上升（公式5）更新策略参数$\theta$。在公式5中不再使用未知的真实动作值函数$Q^{\pi}(s,a)$,而是使用时间差分误差$\delta\_{t}=r\_{t+1}+\gamma V^{v}\left(s\_{t+1}\right)-V^{v}\left(s\_{t}\right)$，可以证明其是真实值的近似值。演员和评论家均采用了重要度采样比率$\frac{\pi\_{\theta}(a | s)}{\beta\_{\theta}(a | s)}$进行调整，因为动作是实际上是根据$\pi$而不是$\beta$选择的。

## 确定性策略的梯度

下面将讨论如何将策略梯度框架扩展到确定性策略。

### 动作价值梯度

大部分无模型强化学习算法都是基于广义迭代策略：交叉进行策略评估和改进。但是在连续动作空间中，贪婪策略改进是有问题的，因为需要在每一步都进行全局最大化。一种简单且计算方便的方式是\*\*沿着$Q$的梯度方向移动策略\*\*，而不是全局最大化$Q$。特别地，对于每一个观察到的状态$s$，策略参数$\theta^{k+1}$按照梯度$\nabla\_{\theta} Q^{\mu^{k}}\left(s, \mu\_{\theta}(s)\right)$进行比例更新，每个状态都提出一个策略改进的不同方向，这些可以通过对状态分布的期望$\rho^{\mu}(s)$求平均值：

$$\theta^{k+1}=\theta^{k}+\alpha \mathbb{E}\_{s \sim \rho^{\mu^{k}}}\left[\nabla\_{\theta} Q^{\mu^{k}}\left(s, \mu\_{\theta}(s)\right)\right]\tag{6}$$

通过应用链式法则，策略改进可以分解成动作价值关于动作的梯度，以及策略关于策略参数的梯度，于是有：

$$\theta^{k+1}=\theta^{k}+\alpha \mathbb{E}\_{s \sim \rho^{\mu^{k}}}\left[\left.\nabla\_{\theta} \mu\_{\theta}(s) \nabla\_{a} Q^{\mu^{k}}(s, a)\right|\_{a=\mu\_{\theta}(s)}\right]\tag{7}$$

按照惯例，$\nabla\_{\theta} \mu\_{\theta}(s)$为一个雅可比矩阵，其中每一列为策略的第$d$个动作维度关于策略参数$\theta$的梯度$\nabla\_{\theta}\left[\mu\_{\theta}(s)\right]\_{d}$。然而，改变策略的同时，会观察到不同的状态，状态分布$\rho^{\mu}$也会发生变化。因此，如果不考虑分布的变化，这种方法是否保证了改进并不是很明显。但是下面的理论表明，无需计算状态分布的梯度。

### 确定性策略梯度理论

考虑一个确定性策略 $\mu\_{\theta}: \mathcal{S} \rightarrow$ $\mathcal{A},\theta \in \mathbb{R}^{n}$ ，定义一个目标函数$J\left(\mu\_{\theta}\right)=\mathbb{E}\left[r\_{1}^{\gamma} | \mu\right]$，概率分布$p\left(s \rightarrow s^{\prime}, t, \mu\right)$和折扣状态分布$\rho^{\mu}(s)$,仍然将目标函数写成期望的形式：

$$\begin{aligned}

J\left(\mu\_{\theta}\right) &=\int\_{\mathcal{S}} \rho^{\mu}(s) r\left(s, \mu\_{\theta}(s)\right) \mathrm{d} s \\

&=\mathbb{E}\_{s \sim \rho^{\mu}}\left[r\left(s, \mu\_{\theta}(s)\right)\right]

\end{aligned}\tag{8}$$

那么可以推导出确定性策略梯度理论如下：

$$\begin{aligned}

\nabla\_{\theta} J\left(\mu\_{\theta}\right) &=\left.\int\_{\mathcal{S}} \rho^{\mu}(s) \nabla\_{\theta} \mu\_{\theta}(s) \nabla\_{a} Q^{\mu}(s, a)\right|\_{a=\mu\_{\theta}(s)} \mathrm{d} s \\

&=\mathbb{E}\_{s \sim \rho^{\mu}}\left[\left.\nabla\_{\theta} \mu\_{\theta}(s) \nabla\_{a} Q^{\mu}(s, a)\right|\_{a=\mu\_{\theta}(s)}\right]

\end{aligned}\tag{9}$$

### 随机策略梯度的极限

乍一看，确定性策略梯度理论并不太像随机策略梯度理论（式2），然而，对于一类广泛的随机策略，包括许多凹凸函数，确定性策略梯度确实是随机策略梯度的一种特殊(极限)情况。现在通过一个确定性策略$\mu\_{\theta}: \mathcal{S} \rightarrow \mathcal{A}$和方差参数$\sigma$来参数化随机策略$\pi\_{\mu\_{\theta},\sigma}$，当$\sigma=0$，随机策略就等价于确定性策略$\pi\_{\mu\_{\theta}, 0} \equiv \mu\_{\theta}$。

## 确定性演员-评论家算法

现在根据确定性策略梯度理论来推导在策略和离策略演员-评论家算法，首先使用简单的Sarsa评论家介绍在策略更新的思想，然后借助Q学习评论家说明离策略的更新思想。

### 在策略确定性演员-评论家

和 随机梯度下降一样，critic 评估动作价值函数，actor 根据式（9）调整$\mu\_{\theta}\left(s\right)$的参数$\theta$ ，并且用$Q^{w}(s, a) \approx Q^{\mu}(s, a)$来逼近真实值。例如，critic 使用Sarsa 更新来评估动作价值函数：

$$\delta\_{t} =r\_{t}+\gamma Q^{w}\left(s\_{t+1}, a\_{t+1}\right)-Q^{w}\left(s\_{t}, a\_{t}\right)\tag{11}$$

$$w\_{t+1} =w\_{t}+\alpha\_{w} \delta\_{t} \nabla\_{w} Q^{w}\left(s\_{t}, a\_{t}\right)\tag{12}$$

$$\theta\_{t+1} =\theta\_{t}+\left.\alpha\_{\theta} \nabla\_{\theta} \mu\_{\theta}\left(s\_{t}\right) \nabla\_{a} Q^{w}\left(s\_{t}, a\_{t}\right)\right|\_{a=\mu\_{\theta}(s)}\tag{13}$$

### 离策略确定性演员-评论家（off-policy deterministic actorcritic，OPDAC）

现在考虑使用离策略的方法，根据任意一个随机行为策略$\pi(s,a)$生成的轨迹学习确定性目标策略$\mu\_{\theta}\left(s\right)$，采用和之前一样方式定义目标函数和策略梯度：

$$\begin{aligned}

J\_{\beta}\left(\mu\_{\theta}\right) &=\int\_{\mathcal{S}} \rho^{\beta}(s) V^{\mu}(s) \mathrm{d} s \\

&=\int\_{\mathcal{S}} \rho^{\beta}(s) Q^{\mu}\left(s, \mu\_{\theta}(s)\right) \mathrm{d} s

\end{aligned}\tag{14}$$

$$\begin{aligned}

\nabla\_{\theta} J\_{\beta}\left(\mu\_{\theta}\right) & \approx \int\_{\mathcal{S}} \rho^{\beta}(s) \nabla\_{\theta} \mu\_{\theta}(a | s) Q^{\mu}(s, a) \mathrm{d} s \\

&=\mathbb{E}\_{s \sim \rho^{\beta}}\left[\left.\nabla\_{\theta} \mu\_{\theta}(s) \nabla\_{a} Q^{\mu}(s, a)\right|\_{a=\mu\_{\theta}(s)}\right]

\end{aligned}\tag{15}$$

和之前一样，这个近似去掉了依赖动作值函数梯度的项：$\nabla\_{\theta} Q^{\pi}(s, a)$。

同样，用一个可微的动作价值函数$Q^{w}(s, a) $来代替真实的动作价值函数$Q^{w}(s, a) \approx Q^{\mu}(s, a)$ ，并且让critic 通过已知策略生成的轨迹来评估它。比如使用 Q-learning：

$$\delta\_{t} =r\_{t}+\gamma Q^{w}\left(s\_{t+1}, \mu\_{\theta}\left(s\_{t+1}\right)\right)-Q^{w}\left(s\_{t}, a\_{t}\right)\tag{16}$$

$$w\_{t+1} =w\_{t}+\alpha\_{w} \delta\_{t} \nabla\_{w} Q^{w}\left(s\_{t}, a\_{t}\right)\tag{17}$$

$$\theta\_{t+1} =\theta\_{t}+\left.\alpha\_{\theta} \nabla\_{\theta} \mu\_{\theta}\left(s\_{t}\right) \nabla\_{a} Q^{w}\left(s\_{t}, a\_{t}\right)\right|\_{a=\mu\_{\theta}(s)}\tag{18}$$

### 兼容函数近似

这部分重点关于为什么可以用一个可微的函数近似器$Q^{w}(s, a)$来代替真实的动作价值函数$Q^{\mu}(s, a)$。和前面"Stochastic Actor-Critic算法"部分类似，如果近似器$Q^{w}(s, a)$与$\mu\_{\theta}\left(s\right)$和$\nabla\_{\theta} J\_{\beta}(\theta)=\mathbb{E}\left[\left.\nabla\_{\theta} \mu\_{\theta}(s) \nabla\_{a} Q^{w}(s, a)\right|\_{a=\mu\_{\theta}(s)}\right]$是兼容的，应满足如下两个条件：

1. $\left.\nabla\_{a} Q^{w}(s, a)\right|\_{a=\mu\_{\theta}(s)}=\nabla\_{\theta} \mu\_{\theta}(s)^{\top} w$

2. $w$最小化均方误差$MSE(\theta, w)=$ $\mathbb{E}\left[\epsilon(s ; \theta, w)^{\top} \epsilon(s ; \theta, w)\right] \quad$，其中$\epsilon(s ; \theta, w)=\left.\nabla\_{a} Q^{w}(s, a)\right|\_{a=\mu\_{\theta}(s)}-\left.\nabla\_{a} Q^{\mu}(s, a)\right|\_{a=\mu\_{\theta}(s)}$

一个compatible off-policy deterministic actor-critic (COPDAC) 算法有两部分组成：Critic 是一个线性函数近似器，通过特征$\phi(s, a)=a^{\top} \nabla\_{\theta} \mu\_{\theta}(s)$来评估动作价值函数。这个可以通过离线地学习已知策略的样本得到，比如用 Q-learning 或者 gradient Q-learning。Actor 则朝着 critic 的动作价值函数的梯度方向更新参数。例如，下列COPDAC-Q 算法中 critic 使用了简单 Q-learning：

$$\delta\_{t} =r\_{t}+\gamma Q^{w}\left(s\_{t+1}, \mu\_{\theta}\left(s\_{t+1}\right)\right)-Q^{w}\left(s\_{t}, a\_{t}\right)\tag{19}$$

$$\theta\_{t+1} =\theta\_{t}+\alpha\_{\theta} \nabla\_{\theta} \mu\_{\theta}\left(s\_{t}\right)\left(\nabla\_{\theta} \mu\_{\theta}\left(s\_{t}\right)^{\top} w\_{t}\right)\tag{20}$$

$$w\_{t+1} =w\_{t}+\alpha\_{w} \delta\_{t} \phi\left(s\_{t}, a\_{t}\right)\tag{21}$$

$$v\_{t+1} =v\_{t}+\alpha\_{v} \delta\_{t} \phi\left(s\_{t}\right)\tag{22}$$

或者 gradient Q-learning critic：

$$\delta\_{t} =r\_{t}+\gamma Q^{w}\left(s\_{t+1}, \mu\_{\theta}\left(s\_{t+1}\right)\right)-Q^{w}\left(s\_{t}, a\_{t}\right) \tag{23}$$

$$\theta\_{t+1} =\theta\_{t}+\alpha\_{\theta} \nabla\_{\theta} \mu\_{\theta}\left(s\_{t}\right)\left(\nabla\_{\theta} \mu\_{\theta}\left(s\_{t}\right)^{\top} w\_{t}\right)\tag{24}$$

$$w\_{t+1} =w\_{t}+\alpha\_{w} \delta\_{t} \phi\left(s\_{t}, a\_{t}\right)

\\ -\alpha\_{w} \gamma \phi\left(s\_{t+1}, \mu\_{\theta}\left(s\_{t+1}\right)\right)\left(\phi\left(s\_{t}, a\_{t}\right)^{\top} u\_{t}\right) \tag{25}$$

$$v\_{t+1} =v\_{t}+\alpha\_{v} \delta\_{t} \phi\left(s\_{t}\right) \\ -\alpha\_{v} \gamma \phi\left(s\_{t+1}\right)\left(\phi\left(s\_{t}, a\_{t}\right)^{\top} u\_{t}\right)\tag{26}$$

$$ u\_{t+1} =u\_{t}+\alpha\_{u}\left(\delta\_{t}-\phi\left(s\_{t}, a\_{t}\right)^{\top} u\_{t}\right) \phi\left(s\_{t}, a\_{t}\right)\tag{27}$$