为什么 PPO 需要重要性采样，而 DDPG 这个 off-policy 算法不需要

PPO 视角

在以 TRPO 和 PPO 这类算法中，为了计算目标函数的梯度 $J(\tilde{\pi})$，需要计算第二项中关于更新策略 $\tilde{\pi}(a|s)$ 的**期望**。这时候 $s \sim \rho_\pi$ 而 $a \sim \tilde{\pi}$，那么对于每一个从 ρ_π 采样得到的状态，为了利用蒙特卡洛法估计 $E_{a\sim\tilde{\pi}}(A_\pi(s,a))$，需要对 $a \sim \tilde{\pi}(a|s)$ 进行大量采样并计算 $A_\pi(s,a)$，且对 $\tilde{\pi}$ 更新一次后便需要重新采样计算，这样做对数据的利用非常低效且方差会很大。

所以我们转而使用**重要性采样 Importance Sampling**，将其转化为对更新前策略函数 $\pi(a|s)$ 的期望：

$$\Sigma\rho_\pi\Sigma\tilde{\pi}(\Sigma r^t A_\pi(s,a)) = E_{s\sim\rho_\pi,a\sim\tilde{\pi}}(A_\pi(s,a)) = E_{s\sim\rho_\pi,a\sim\pi}(\frac{\tilde{\pi}(a|s)}{\pi(a|s)}A_\pi(s,a))$$

如此一来，$s \sim \rho_\pi, a \sim \pi$，那么 π 与环境交互所得的状态和动作服从上述分布，可以直接用于估计：

$$E_{s\sim\rho_\pi,a\sim\pi}(\frac{\tilde{\pi}(a|s)}{\pi(a|s)}A_\pi(s,a))$$

这样对数据的利用便高效很多，不过依然容易出现估计的方差过大的问题。但是 PPO 中通过对重要性采样系数（importance sampling ratio）$\frac{\tilde{\pi}(a|s)}{\pi(a|s)}$ 的 clip 操作，使得估计的方差得到有效控制。

DDPG视角

DDPG(Timothy et al., 2015) 算法的核心，是 DPG(Silver et al., 2014)原论文中推导出的 **off-policy 版的确定性策略梯度定理**。此定理与**策略梯度定理**最大的区别在于，一方面用于更新当前策略 μ 的(s,a)分布服从行为策略 β，另一方面策略 μ 是确定性策略。此定理的目标函数如下：

$$J_\beta(\mu_\theta) = E_{s\sim\rho_\beta}V^\mu(s) = \int_S \rho_\beta(s)V^\mu(s)ds = \int_S \rho_\beta(s)Q^\mu(s,\mu_\theta(s))ds$$

注：此处目标函数定义为 $E_{s\sim\rho_\beta}V^\mu(s)$ 而不是 $E_{s\sim\rho_\beta}\frac{\rho_\mu}{\rho_\beta}V^\mu(s)$，其原因可以参考2012年提出的首个off-policy Actor Critic算法：The Off-Policy Actor-Critic algorithm (Degris et al., 2012b)。

有了目标函数后，我们对策略 μ 的参数 θ 求梯度便可得到 **off-policy 版的确定性策略梯度定理**，具体如下所示：

$$
\begin{aligned}
\nabla_\theta J_\beta(\mu_\theta) &= \int_S \rho_\beta(s)(\nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a) + \nabla_\theta Q^{\mu_\theta}(s,a))|_{a=\mu(s)} ds \\
&\approx \int_S \rho_\beta(s) \nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a) ds \\
&= E_{s\sim\rho_\beta}[\nabla_\theta \mu_\theta(s) \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu(s)}]
\end{aligned}
$$

> 上文公式中近似的合理性，在 DPG 原文中是这样描述的：
>
> Analogous to the stochastic case, we have dropped a term that depends on $\nabla_\theta Q^{\mu_\theta}(s,a)$; justification similar to Degris et al. (2012b) can be made in support of this approximation.
>
> 在 Degris et al. (2012b)一文中， 作者对 off-policy 版的随机性策略梯度定理的这一步近似进行了严谨证明，且定量分析得出这样的近似效果并不差。但是，上述说法对于确定性策略是否还成立，DPG 作者并未给出严谨证明，目前是一个比较模糊的状态。不过原文中证明了在某些条件下：确定性策略等价于方差趋于零的随机性策略，因此在实践中也就暂且认同这个结论。

总结

所以对于本小节提出的问题，一句话总结就是：**DDPG 目标函数梯度计算公式中不存在对动作的积分，所以即使作为 off-policy 算法，也不需要使用重要性采样**。