为什么off-policy中可以用一个策略采样来更新另外一个策略：

解释：因为采样策略π(θ’)和要更新的策略π(θ)获得的reward的期望相同（这就说明两个策略的分布近似），所以可以用π(θ’)来更新π(θ)。

在莫烦PPO解释中有这么一段话：

“**总的来说 PPO 是一套 Actor-Critic 结构, Actor 想最大化 J\_PPO, Critic 想最小化 L\_BL. Critic 的 loss 好说, 就是减小 TD error. 而 Actor 的就是在 old Policy 上根据 Advantage (TD error) 修改 new Policy, advantage 大的时候, 修改幅度大, 让 new Policy 更可能发生. 而且他们附加了一个 KL Penalty (惩罚项, 不懂的同学搜一下 KL divergence), 简单来说, 如果 new Policy 和 old Policy 差太多, 那 KL divergence 也越大, 我们不希望 new Policy 比 old Policy 差太多, 如果会差太多, 就相当于用了一个大的 Learning rate, 这样是不好的, 难收敛.**”

这里解释一下：

1. TD error是用π(θ’)经过采样得到的data来更新π(θ)后的第n+1步的V与第n步的V的差，因为要让π(θ’)和π(θ)差距较小，所以要让TD error较小，因为如果两者策略差距很大，则会导致不能用采样策略来更新原始策略。
2. 莫烦的说法里old Policy和new Policy并不是同一个策略的新旧更新态，而是指old Policy是我们要更新的策略，new Policy实际上是采样策略。PPO的目的就是梯度更新策略的时候使得采样策略和更新策略的分布近似，更新的即不是太快也不是太慢。