注意，这是一个同轨学习方法，因为产生样本的策略是 $\pi(a \mid s, \theta_t)$，而策略改进正是为了改进这一策略。

第三个改进是通过重要性采样使得 A2C 具备离轨学习能力。重要性采样是指通过从一个行为分布采样得到的数据，对服从另一个目标分布的随机变量进行估计。例如，已知一个目标分布 T，一个行为分布 B，则服从目标分布的随机变量 X 的数学期望：

$$E_{X\sim T}[X] = \sum_x p_T(x)x = \sum_x p_B(x)\frac{p_T(x)}{p_B(x)}x = E_{X\sim B}\left[\frac{p_T(x)}{p_B(x)}X\right] \approx \frac{1}{n}\sum_{i=1}^{n}\frac{p_T(x_i)}{p_B(x_i)}x_i$$

可见，通过对从行为分布采样得到的数据进行加权，即可实现对服从目标分布的随机变量数学期望的估计。这里，目标分布与行为分布在给定值处概率质量/概率密度之比，称为该值的重要性权重。而应用重要性采样估计数学期望时，应利用重要性权重计算加权平均。

因此，在离轨学习中，如果数据来自于行为策略 β 而非目标策略 π，则策略梯度的计算应遵循离轨策略梯度定理，即：

$$\nabla_\theta J(\theta) = E_{S\sim\eta, A\sim\beta}\left[\frac{\pi(A|S,\theta)}{\beta(A|S)}\nabla_\theta \ln \pi(A \mid S,\theta) q_\pi(S,A)\right]$$

简言之，应该将同轨学习中策略对数的梯度乘以重要性权重 $\pi(A|S,\theta)/\beta(A|S)$，即目标策略与行为策略在给定状态下采取特定行动的概率之比。值得注意的是，离轨策略梯度仍然具有基线不变性，因此可以使用优势函数代替行动价值。

综合上述分析，使用优势函数的离轨 A2C 算法，其伪代码如下：

注意其中加入的重要性权重。