Policy Gradient主要思想：在每一个状态下，根据现有的$P(a_t|s_t)$采样a_t，如此往复，获得一组状态-行为对：$s_1, a_1, s_2, a_2, \ldots, s_T$，此时获得最终的奖励函数$r_T$，这里我们假设$r_T$可取正负值，其中正值表示获得奖励，负值表示获得惩罚。最终我们根据r_T去修改每一步的$P(a_t|s_t)$：

$$P(a_t|s_t) = P(a_t|s_t) + \alpha r_{\mathrm{T}}$$

如果$P(a_t|s_t) \sim Q(s_t, a_t, \theta)$，则有：

$$\theta = \theta + \alpha r_t \nabla_\theta Q(s_t, a_t, \theta)$$