1. 在策略，因为探索和训练的为同一个策略。
2. 解答如下：
   - 损失函数为 $LOSS=-\frac{1}{m}\sum_{i=1}^{m}\log\left(P\left(\tau,\theta\right)\right)\left(R\left(\tau\right)-b\right)$
   - 使用熵函数，保证动作不会过于集中。
3. 随着 epoch 获得 reward 的变化如下图：