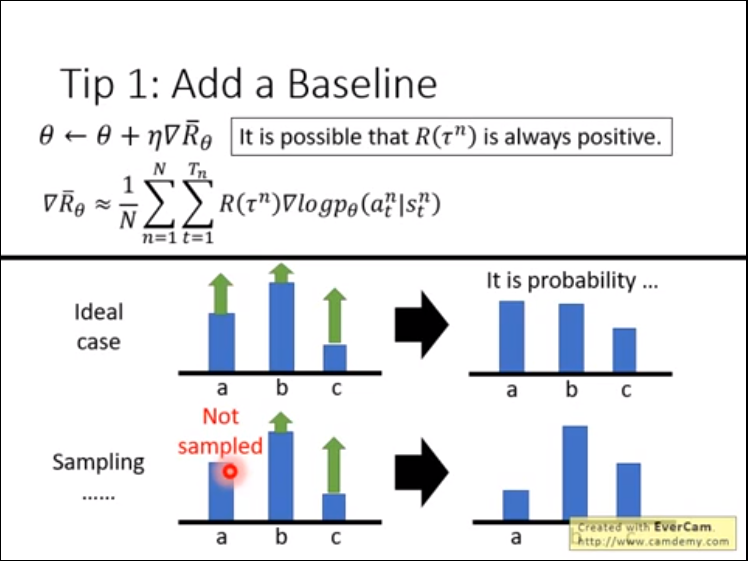
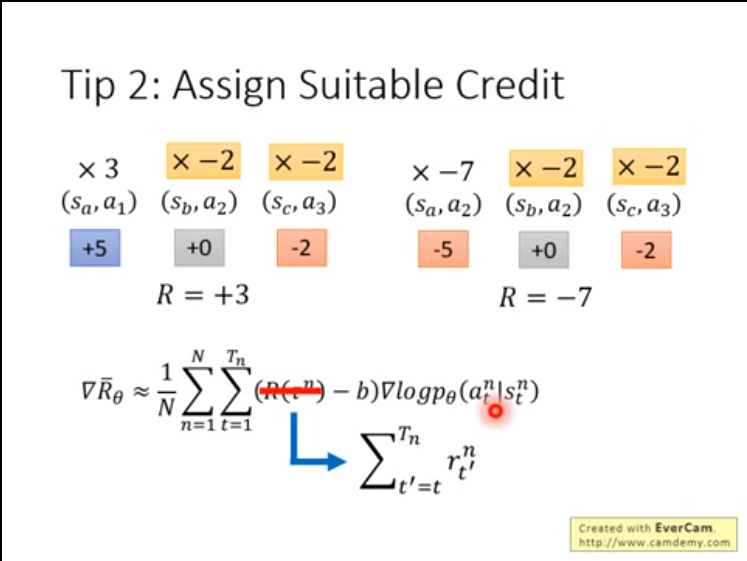
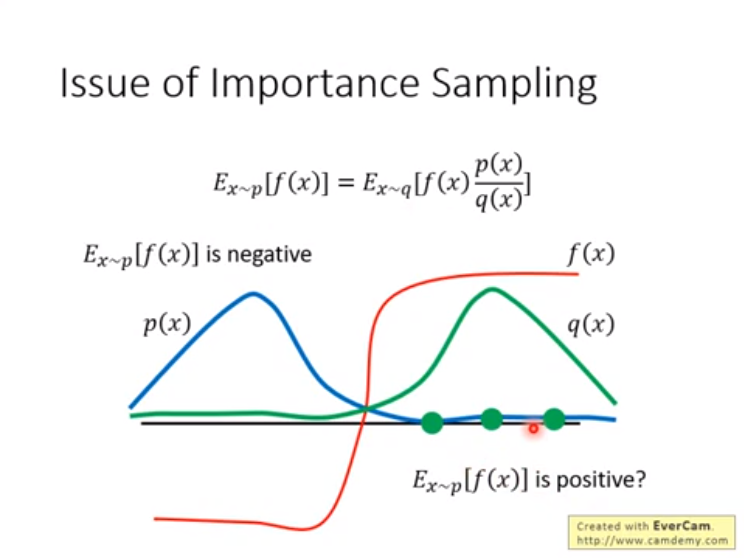
1.策略梯度方法回报R全为正，但sample结果差可能会导致的问题：a没有被采样到，它的回报很高但agent采取它的概率会变低



2.策略梯度方法每次计算采取总的回报会出现问题：虽然整个回合的回报很高但不能说明回合中每一步的回报都很高，比如说(sc,a3)就是低回报的actor

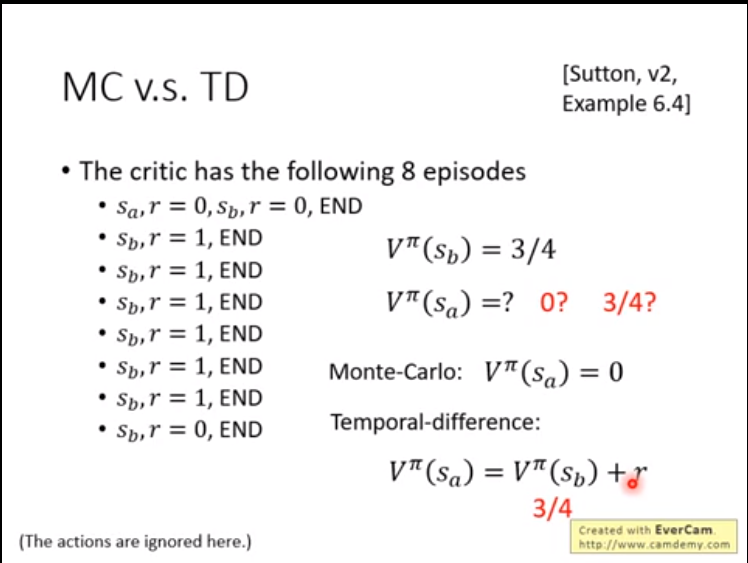


3.重要性采样sample次数不够会导致的问题：等式左边值为负，而右边因为sample次数不够，所以采样到的点都是右边的绿点，计算出的等式右边结果为正，会导致等式有误差

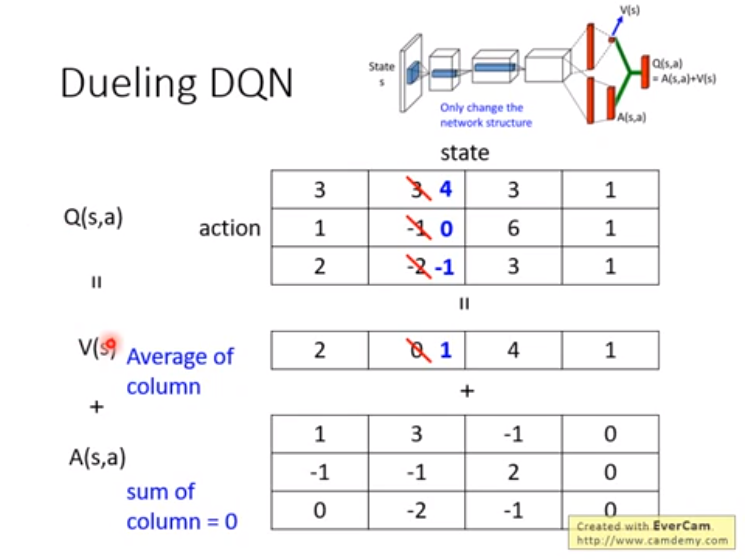


4.TRPO是把θ和θ’的KL散度放在约束里，PPO是把KL散度放在公式里，实际操作来看PPO要比TRPO容易得多

5.MC和TD方法对值函数估计结果的不同：V(Sb)都一样，V(Sa)会不一样，采用MC方法的话就一个样本，所以V(Sa)=0；采用TD方法的话V(Sa)=r+V(Sb)=3/4。这样就会造成MC和TD方法结果的不同，但是不能说谁对谁错，因为两种方法的假设条件不一样。



6.Dueling DQN的好处：可以通过更改V(s)将所有的Q(s,a)都更改，提高效率



1. 稀疏奖励的解决办法：

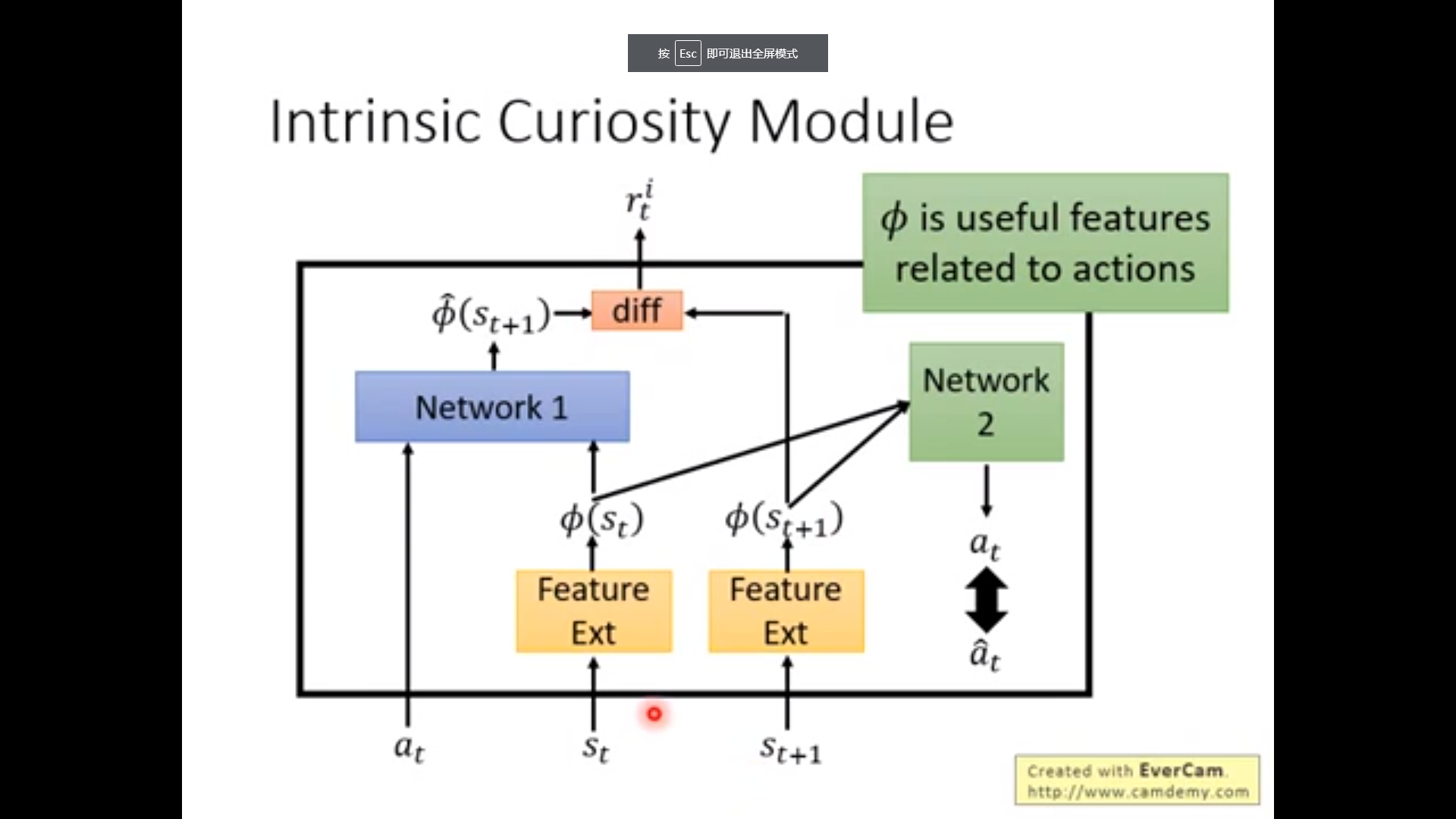
设计伪奖励

好奇心：计算 预估s’和实际s的区别作为另一个奖励，以鼓励探索

Curriculum learning：无法直接完成最终目标，设定一个个小目标一点一点逼近最终完成任务

Hierarchical reinforcement learning：有好几个agent，类似校长教授研究生，每个agent负责 不同level的任务

多目标强化学习



1. 模仿学习的几个方法：

Behavior Cloning：直接复制行为，但是对于采样未出现的状态不会处理；而且agent可能学到专家行为中不好的部分  
data aggregation：一个专家坐在车上，每次专家给出自己在当前状态会采取的行为但是agent继续采取自己的行为，这种情况可以解决采样未出现的情况，但是容易牺牲expert

Inverse reforcement learning：逆向强化学习，采用expert策略和agent策略进行采样，计算合适的回报函数(核心)使得expert策略的回报比agent策略大，然后通过正向强化学习和回报函数更新agent策略，再比较expert策略和agent策略设计新的回报函数

