**Introduction**

奖励函数的设计是解决强化学习问题的关键挑战，对于复杂应用，我们很难手工设计出一个好的奖励函数，（which决定了强化学习的成败），同时，由于环境存在高风险，我们在训练智能体的过程中无法与环境进行交互和测试策略。

临床决策是一个典型的例子，我们希望从医学专家那里学习策略，但在训练过程中对患者实施策略是完全不切实际和不道德的。此外，我们不仅对演示的策略感兴趣，而且对演示者的偏好和目标（即指导行为的奖励函数）感兴趣。在这种情况下，学习必须完全基于专家演示者记录的轨迹进行，该轨迹仅显示访问的状态和在每种情况下采取的动作。

虽然模仿学习的策略可以匹配演示者的行为，但随着理解动机的增加，可以通过逆向强化学习来实现。(Abbeel & Ng，2004)将潜在奖励假设为特征的线性函数，然后使用前向强化学习相对于奖励的最优策略。(Klein et al., 2011、 Klein et al., 2013 )将该算法扩展为离线策略学习。(Lee et al., 2019)中的DSFN是目前离线最大边际方法中的state-of-art，然而DSFN存在以下问题：（1）使用DQN作为前向强化学习方法，而replay buffer是固定的，因此初始策略需要与学习策略相近，这就需要对初始策略进行预训练；（2）使用神经网络进行特征提取，最终得到的关于特征的奖励函数是不可解释的，这一点尤为重要，因为这与逆强化学习的初衷--得到演示者的目的和偏好相违背。

针对以上问题，我们在(Abbeel & Ng，2004)算法基础上，使用BCQ进行前向策略评估与改进，优点：一是使用的Double-Q学习方法可以对Q估计更加准确；二是在学习Q值的近似的同时，将动作限制在演示的动作集合中；三是只使用演示数据集，不再与环境进行交互，实现了完全离线学习。

我们使用raw state作为特征，来观察学出的奖励是否合理

**Background**

**Preliminaries**

我们使用标准的马尔科夫决策过程（MDP）设置，which定义为一个五元组（S,A，T，r，γ），with状态集合S，动作集合A，转移概率矩阵T（which we don’t know），奖励函数r，折扣因子γ。策略π（a|s）表示在状态s(∈S)处执行动作a(∈A)的概率。价值函数定义为，其中s0∈S0,表示初始状态。状态价值函数定义为衡量在策略π下状态s和动作a的优劣，其中表示策略π下的期望。RL目标是求出一个策略Π\*，使得Q和V的值最大，即。我们的设定是R是未知的，且训练过程中不可与环境交互，我们所能得到的只有根据专家策略采样得到专家数据集。

**Imitation Learning**

模仿学习【2】的目标是学习一个策略，使得和专家策略尽量接近。 （s属于S，）。

其中L代表某种损失函数。BC【5】是一个高效但十分简单的方法，它忽略了状态之间的转移关系这一宝贵信息，仅仅使用监督学习进行分类预测。由于BC未考虑到长序列规划（no-longterm-planning），这导致BC受到compounding errors的影响而漂移到远离专家数据的地方【6】。一旦飘移到这些out-of-distribution的状态，智能体无法做出正确的决策。

**APPrenticeship Learning**

为了使用状态转移的信息，APP算法被提出，这其中包括了逆强化学习算法和对抗模仿学习【7】，这种学习方式大多是在线学习【注：本文中所提到的在线学习都是指 需要与环境交互，offline或batch都是指完全不与环境交互】，它们产生的奖励函数也是不可解释的。IRL逆强化学习不仅要求出最优策略，而且要recover一个合适的奖励来解释行为。NG【8】提出的算法通过最大化学习策略与专家策略的奖励间距来优化，需要进行交互进行策略评价和优化，这一部分的细节在之后章节。

**Batch IRL**

但是，很多时候我们并不能获得与环境的交互。

假设我们，并且假设奖励函数是关于状态动作特征的线性函数，，其中是关于状态s和动作a特征映射。关于特定状态s和动作a的特征期望定义为策略π下，累积折扣“特征访问”的期望：

总体的特征期望定义为对初始状态集合的期望：

**Method**

模型框架如图1，我们使用神经网络来近似特征期望，以及动作价值函数Q；输入为专家数据集，输出为特征期望参数、最优策略以及奖励函数参数。

可分为三个部分：

1. 根据当前策略，使用DSFN估计特征期望；
2. 求解二次优化QP或者使用投影法优化奖励函数参数；
3. 用BCQ进行策略评估以及策略改进得到

下面是这三个部分的细节：

1. DSFN估计特征期望

这部分参考于(Lee et al., 2019)

我们想要学习一个神经网络来近似特征期望，使得，该网络可以用贝尔曼方程中的时域差分误差TD-errors来训练。给定π，φ，贝尔曼目标值如下：



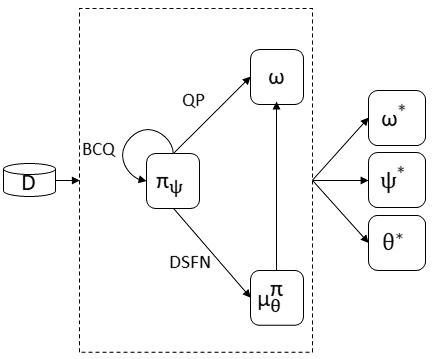
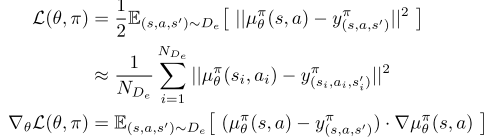


图1.模型框架图

使用均方误差来训练DSFN，损失和梯度计算如下：

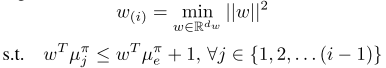


训练的过程和DQN十分相似，不同的是DSFN只进行策略评价而不进行策略改进。因为无法与环境交互来评估DSFN的表现，因此只能使用交叉验证，当损失小于设定的固定阈值δ时训练结束。

1. 计算奖励函数参数

这部分参考自(Abbeel & Ng，2004)。

通过二次规划QP求解最大间隔问题：



投影法求解：



这部分计算在和上的正交投影。



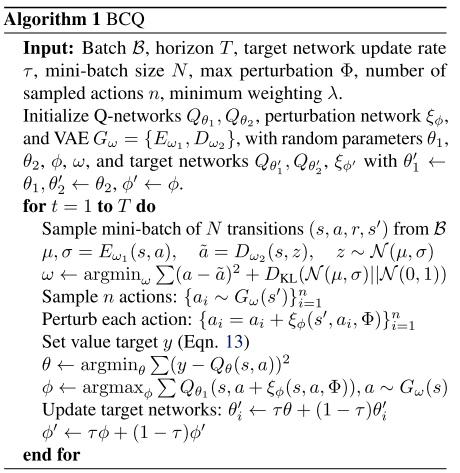
表示专家特征期望和学习到的特征期望的距离，通常整个算法的结束条件为减小到某一阈值。

1. 用BCQ改进策略

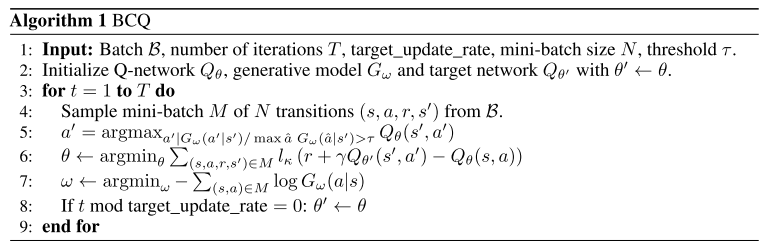
这部分参考自(Fujimoto et al, 2019)

BCQ是一种离线强化学习算法，主要解决的问题是在离线设置下产生的外推误差。外推误差即，针对专家演示集中未出现的状态动作对，Q价值网络产生的错误估计。BCQ的解决办法是在最大化奖励的同时，将动作选择限制在演示集中出现的动作中。

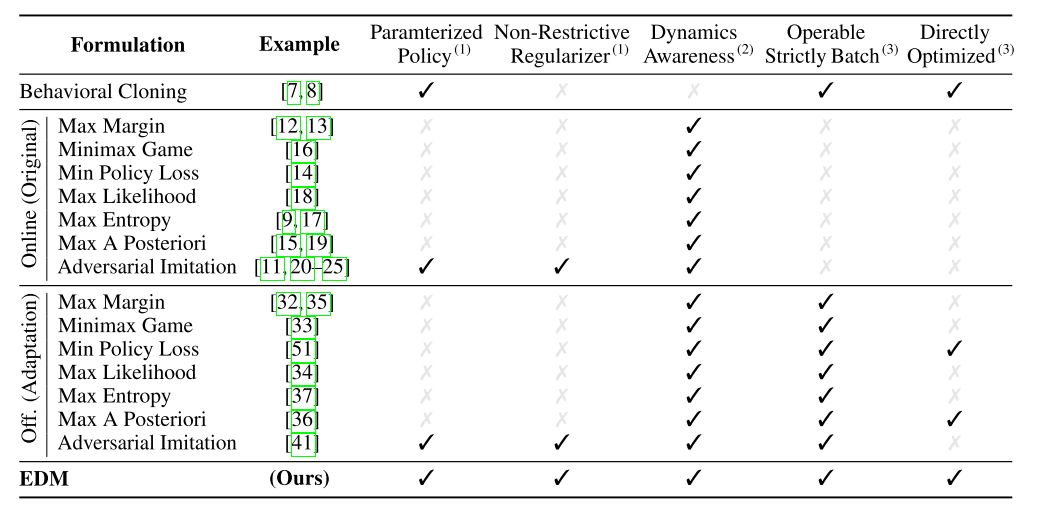
对于连续动作，其算法如下：



对于离散动作：



(Benchmarking Batch Deep Reinforcement Learning Algorithms, Fujimoto 2019)



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | 是否  ｏｆｆｌｉｎｅ | 奖励的可解释性 | 不需要ｗａｒｍｓｔａｒｔ |  |
| BC |  | YES | NO | ／ |  |
| APP |  | NO | YES | YES |  |
|  |  |  |  |  |  |
| CIRL |  | YES | YES | NO |  |
| DSFN |  | YES | NO | NO |  |
| OURS |  | YES | YES | YES |  |

注意：YES表示不需要warm-start

一般脱离了在线条件，对策略进行离线评估，就需要进行warm-start，而warm-start对数据量的要求较高，离线评估时，会遇到外推误差，即选取到了专家序列中没有的动作时，Q值的估计会出现较大的偏差。解决的办法是将动作限制在一定范围内

**实验**

为了验证我们方法的有效性，我们在gym环境上进行了实验。我们比较了我们的方法和DSFN，LSTD和BC

To evaluate the effectiveness of Batch-Apprenticeship learning in a high-dimensional setting, we focus on MuJoCo environments in OpenAI gym (Todorov et al., 2012; Brockman et al., 2016). We compare our method with DSFN，behavioral cloning method (BC).

**Training Details** For BCQ model, we first trained a TRIL network for warm start. We used a 70-30 training- validation split and following [Duan et al., 2016], included a Gaussian output layer that learns the means and standard deviations for the transition prediction — necessary to learn the uncertainty in our highly stochastic clinical domain exper- iment (Section 7). Further training details in terms of TRIL, DSFN model architecture and hyperparameters are provided in the appendix (Table 4). The IRL update was computed with the max-margin algorithm [Abbeel and Ng, 2004](Algorithm 1).

Exact implementation and experimental details are provided in the Supplementary Material.

Conclusion

在未来，要讨论非线性函数的情况。