



# Hindsight Credit Assignment - hoho

## 论文试图解决什么问题？

如何在状态 $x$ 下通过选择适当的动作 $a$ ，从而影响未来的回报？（how does choosing an action  $a$  in a state  $x$  affect future return?）

当前的问题：

1. 简单的通过平均回报来估计价值函数，譬如蒙特卡洛式的方法，不高效，因为会引入许多随机性，导致高方差
2. 部分观测问题（Partial observability）。TD方法，如Sarsa或Q-learning，会引入偏差，通过bootstrap估计价值函数，可能会难以收敛。
3. 依赖于时间。TD( $\lambda$ )可以平衡上述两点高方差和偏差问题，但是它十分依赖于时间作为相关性的衡量：the more recent the action, the more credit or blame it receives from a future reward.
4. 人们更希望使用同一个过程轨迹更新所有相关的动作，而不仅仅只有发生的动作才进行更新。

## 这是否是一个新的问题？

不是新问题

## 这篇文章要验证一个什么科学假设？

给定未来的输出（reward or state），如何衡量当时在状态 $x$ 下选择动作 $a$ 与这未来的输出的相关性（given the future outcome (reward or state), how relevant was the choice of  $a$  in  $x$  to achieve it?）

## 有哪些相关研究？如何归类？谁是这一课题在领域内值得关注的研究员？

1. 基于目标条件（goal conditioning），学习事后回溯的模型（backtracking model）的方法。
2. 使用注意力机制，基于时间回溯（backward）将credit有效分配的方法
3. 有很多降低方差的技术也应用到RL

## 论文中提到的解决方案之关键是什么？

提出了一个Hindsight Credit Assignment(HCA)方法，核心思想是：当用Monte Carlo方法采样只采样到很少你感兴趣的样本，然后做估计，结果往往不准确。这时可以使用测度变换（change measure）：使用另外一个分布，这个分布却可以采样很多你感兴趣的样本，然后用重要性采样方法去修正模型，即：

$$\mathbb{E}_{x \sim p}[X] = \mathbb{E}_{x \sim q}\left[\frac{p}{q}X\right]$$

1. 基于未来的状态的方法

定义： $h_k(a|x, \pi, y) = \mathbb{P}_{\tau \sim \tau(x, \pi)}(A_0 = a | X_k = y)$ ，其中第k步状态为y，用于量化当前动作a与未来状态y的相关性：若不相关，则 $h_k$ 与策略 $\pi$ 相等；若正相关，则 $h_k > \pi$ ；若负相关，则 $h_k < \pi$

可以通过下式理解：根据贝叶斯定理，有

$$\begin{aligned} \frac{h_k(a|x, \pi, y)}{\pi(a|x)} &= \frac{\mathbb{P}(X_k = y | X_0 = x, A_0 = a, \pi)}{\mathbb{P}(X_k = y | X_0 = x, \pi)} \\ &= \frac{\mathbb{P}_{\tau \sim \tau(x, a, \pi)}(X_k = y)}{\mathbb{P}_{\tau \sim \tau(x, \pi)}(X_k = y)} \end{aligned}$$

可见两个分布的比值就是不同采样条件下采样到未来的状态y的概率的比值。

- 若a与y不相关，说明无论一开始采用什么动作，到达状态y的概率应该是一样的。
- 若a与y正相关，说明一开始使用动作a进行采样，到达状态y的概率，这样的事件会比一开始不考虑指定动作的高，即 $h_k > \pi$
- 若a与y负相关，则说明一开始使用动作a采样最终转到状态y的概率会比一开始不考虑具体动作的概率低，即 $h_k < \pi$

据此，作者新定义了新的Q函数： $Q^\pi(x, a) = r(x, a) +$

$$\mathbb{E}_{\tau \sim \tau(x, \pi)} \left[ \sum_{k \geq 1} \gamma^k \frac{h_k(a|x, \pi, y)}{\pi(a|x)} R_k \right]$$

对应的优势函数： $A^\pi(x, a) = r(x, a) - r^\pi(x) + \mathbb{E}_{\tau \sim \tau(x, \pi)} \left[ \sum_{k \geq 1} \left( \frac{h_k(a|x, X_k)}{\pi(a|x)} - 1 \right) \gamma^k R_k \right]$

为了降低对时间的依赖性，进一步修正为

$A^\pi(x, a) = r(x, a) - r^\pi(x) + \mathbb{E}_{\tau \sim \tau(x, \pi)} \left[ \sum_{k \geq 1} \left( \frac{h_\beta(a|x, X_k)}{\pi(a|x)} - 1 \right) \gamma^k R_k \right]$ ，其中 $\beta \in [0, 1)$ ，表示在每个时间步的“存活概率”

## 2. 基于未来的回报的方法

定义 $h_z(a|x, \pi, z) = \mathbb{P}_{\tau \sim \tau(x, \pi)}(A_0 = a | Z(\tau) = z)$ ，其中 $z$ 是整个过程的回报

价值函数改为 $V^\pi(x) = \mathbb{E}_{\tau \sim \tau(x, a, \pi)} \left[ Z(\tau) \frac{\pi(a|x)}{h_z(a|x, Z(\tau))} \right]$

优势函数改为 $A^\pi(x, a) = \mathbb{E}_{\tau \sim \tau(x, a, \pi)} \left[ \left( 1 - \frac{\pi(a|x)}{h_z(a|x, Z(\tau))} \right) Z(\tau) \right]$ (注意比值跟“基于未来状态的方法”的是倒过来的)

$c(a|x, Z) = 1 - \frac{\pi(a|x)}{h_z(a|x, Z)}$ 反映了动作 $a$ 对回报 $Z$ 的贡献程度：若为0，则 $a$ 无贡献（ $a$ 与其他动作同效用）；若小于0，其他动作比 $a$ 更有效；若大于0，对获得回报 $Z$ ， $a$ 比其他动作更有效；

另外，相应的策略梯度方法也做相应改进。

由此得出：

### 1. 基于状态的HCA策略梯度算法

---

**Algorithm 1** State-conditional HCA

---

**Given:** Initial  $\pi, h_\beta, V, \hat{r}$ ; horizon  $T$ 

```
1: for  $k = 1, \dots$  do
2:   Sample  $\tau = X_0, A_0, R_0, \dots, R_T$  from  $\pi$ 
3:   for  $i = 0, \dots, T - 1$  do ▷ Train hindsight distribution
4:     for  $j = i, \dots, T$  do
5:       Train  $h_\beta(A_i|X_i, X_j)$  via cross-entropy
6:     end for
7:   end for
8:   for  $i = 0, \dots, T - 1$  do ▷ Train baseline and reward predictor
9:      $Z = 0$ 
10:    for  $j = i, \dots, T - 1$  do
11:       $Z \leftarrow Z + \gamma^{j-i} R_j$ 
12:    end for
13:     $Z \leftarrow Z + \gamma^{T-i} V(X_T)$ 
14:    Update  $V(X_i)$  towards  $Z$ 
15:    Update  $\hat{r}$  towards  $R_i$ 
16:  end for
17:  for  $i = 0, \dots, T - 1$  do ▷ Train policy of all actions with the hindsight-conditioned return
18:    for all actions  $a$  do
19:       $Z_h = \pi(a|X_i, a) \hat{r}(X_i, a)$ 
20:      for  $j = i + 1, \dots, T - 1$  do
21:         $Z_h \leftarrow Z_h + \gamma^{j-i} \frac{h_\beta(a|X_i, X_j)}{\pi(a|X_i)} R_j$ 
22:      end for
23:       $Z_{h,a} \leftarrow Z_h + \gamma^{T-i} \frac{h_\beta(a|X_i, X_T)}{\pi(a|X_i)} V(X_T)$ 
24:    end for
25:    Follow the gradient  $\sum_a \nabla \pi(a|X_i) Z_{h,a}$ 
26:  end for
27: end for
```

---

(hoho\_todo: 上图via cross entropy具体怎么做？ground truth是啥？)

然后可以根据下式子估计所有动作的回报：

$$Q^x(X_s, a) \approx \hat{r}(X_s, a) + \sum_{t=s+1}^{T-1} \gamma^{t-s} \frac{h_\beta(a|X_s, X_t)}{\pi(a|X_s)} R_t + \gamma^{T-s} \frac{h_\beta(a|X_s, X_T)}{\pi(a|X_s)} V(X_T).$$

## 2. 基于回报的HCA

算法流程

---

**Algorithm 2** Return-conditional HCA

---

**Given:** Initial  $\pi, h_z, V$

```
1: for  $k = 1, \dots$  do
2:   Sample  $\tau = X_0, A_0, R_0, \dots$  from  $\pi$ 
3:   for  $i = 0, 1, \dots$  do
4:     Compose the return  $Z(\tau_{i:\infty})$  starting from  $X_i$ 
5:     Train  $h_z(A_i|X_i, Z_i)$  via cross-entropy
6:      $Z_h \leftarrow \left(1 - \frac{\pi(A_i|X_i)}{h_z(A_i|X_i, Z(\tau_{i:\infty}))}\right) Z(\tau_{i:\infty})$ 
7:     Follow the gradient  $\nabla \log \pi(A_i|X_i) Z_h$ 
8:   end for
9: end for
```

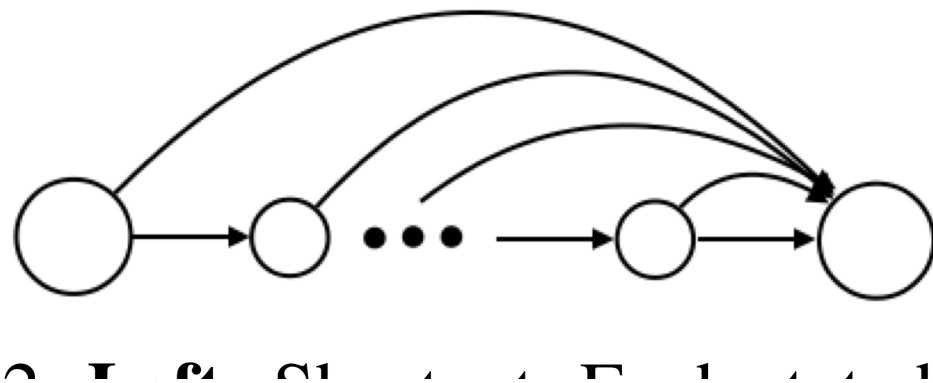
---

## 论文中的实验是如何设计的？

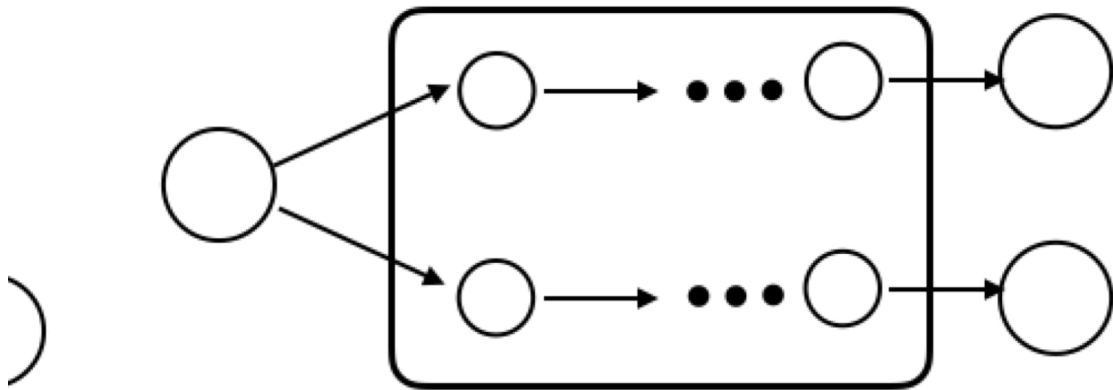
与基本的策略梯度算法作对比

实验设置了以下三种情形的过程：

1. short cut（反映了问题3和问题4）：每个状态都有两个动作：一个直接到最后的  
状态，一个到下一个状态

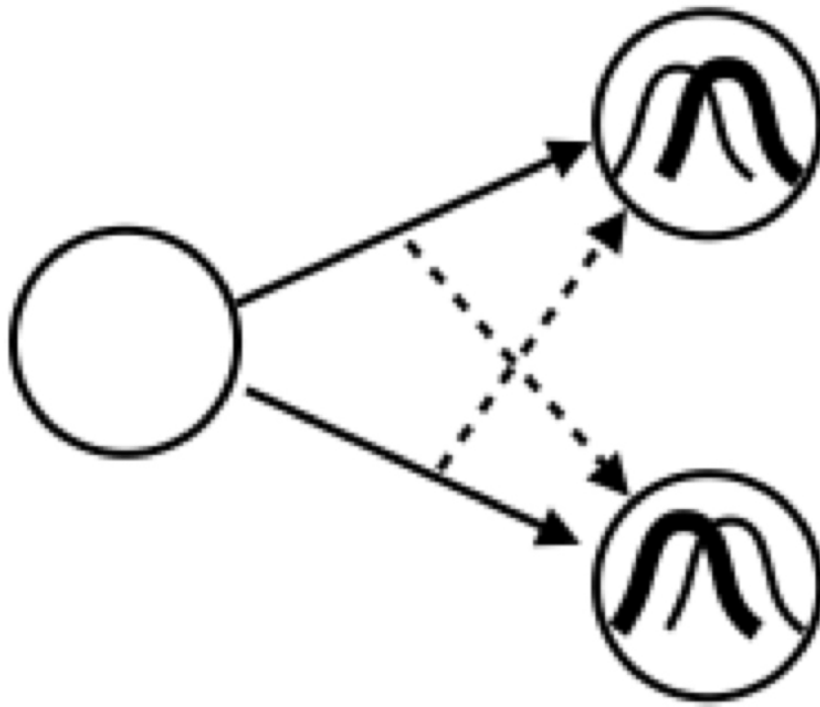


2. 延迟效果（反映了问题2）：开始状态后可选两个动作，分别导致最后不同的状态



has two actions. one transitions direct

3. 混淆情形（反映了问题1）：每个动作都有大概率到一个确定的状态，但也有小概率到其他状态。

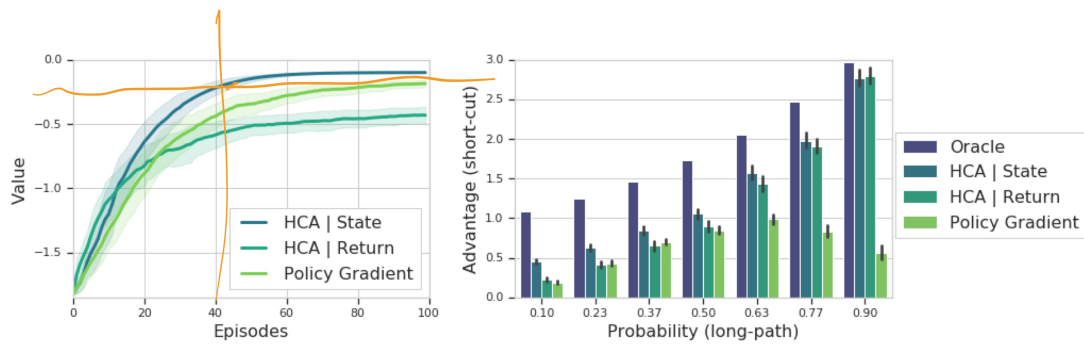


**用于定量评估的数据集是什么？代码有没有开源？**

无。无

**论文中的实验及结果有没有很好地支持需要验证的科学假设？**

- short cut 评估结果：



左图HCA-state收敛最快，右图（x轴表示long-path策略的概率，即不直接走到最终状态，而走一个经历很多状态才能到最终状态的过程），可见随着这样的情况概率上升，使用HCA优势值越来越明显。

- Delay effect评估结果：

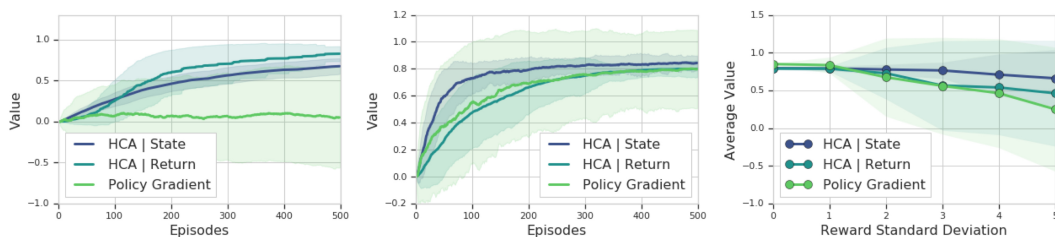


Figure 4: Delayed effect. Left: Bootstrapping. The learning curves for  $\gamma = 5$ ,  $\alpha = 0$ , and a 2 step

左图：使用bootstrapping，采样5个时间步，观察3步的回报，可见HCA方法持续有上升

中图：使用Monte Carlo采样整个时间步（开始到回合结束），观察3步的回报，可见HCA收敛的快且稳定。

## 这篇论文到底有什么贡献？

提出了一种衡量动作与回报相关度的方法，一种解决credit-assignment的方案

## 下一步呢？有什么工作可以继续深入？

后续可以关注下注意力机制加入到过程的每个时间步，现有的方法是有偏的。



