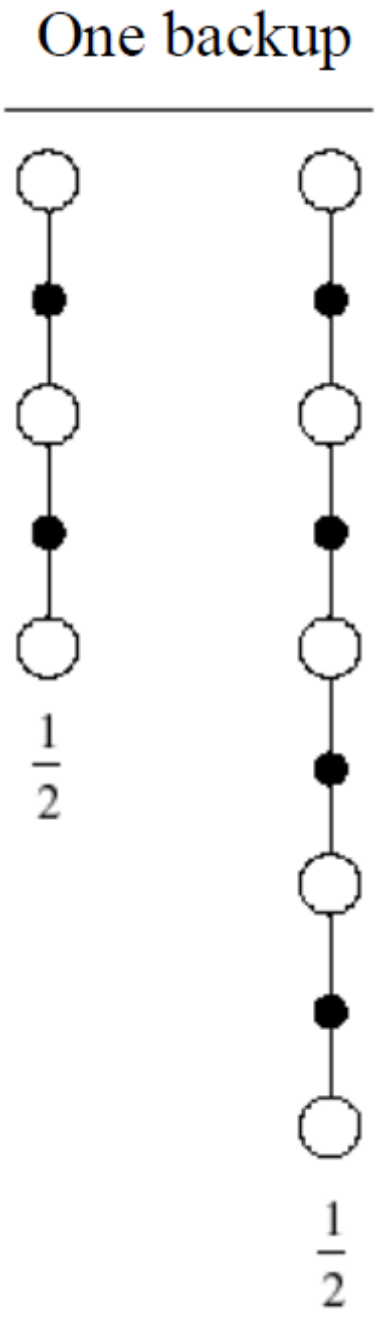
强化学习基础篇（二十六）$TD(\lambda)$预测

1、平均n-Step回报

从在上一篇中我们考虑了n-Step回报，在每个n的选择都有着相应的回报（Reward）。我们如果把不同的n-step回报都做一个如下的平均，例如对2-step和4-step回报可以这样：

$$\frac{1}{2}G^{(2)}+\frac{1}{2}G^{(4)}$$

这样我们得到的回报信息就结合了2个不同的时间步的结果。

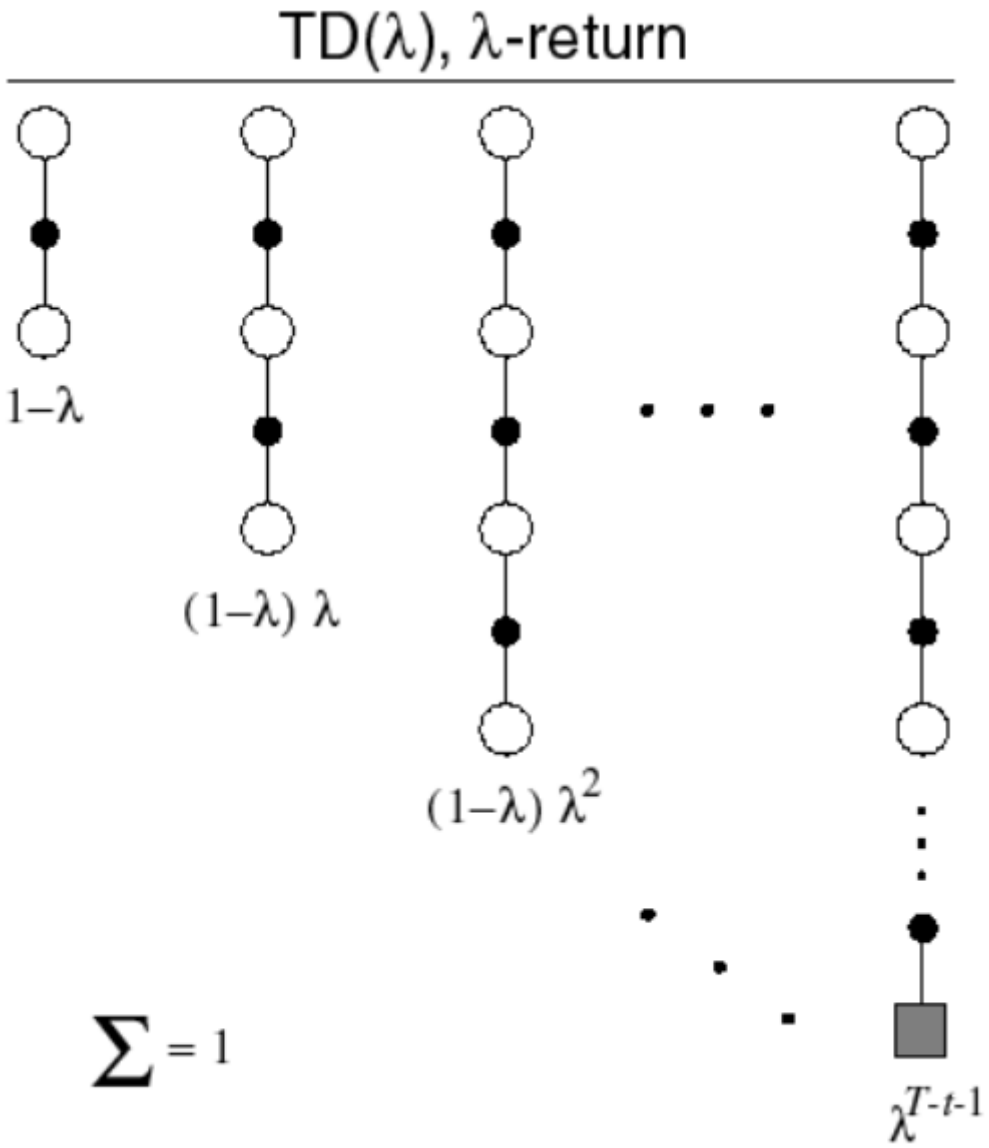


2、$\lambda-return$

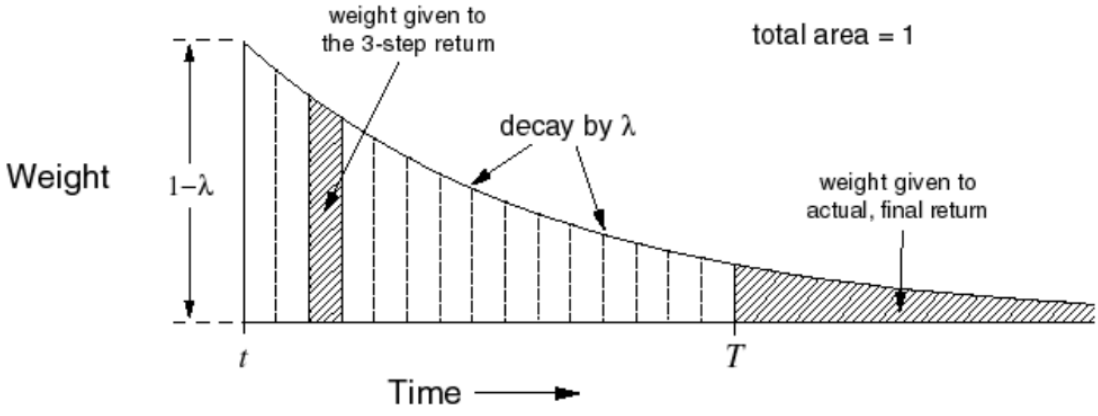
如果我们考虑结合所有的n-step的回报$G_t^{(n)}$，就可以得到$\lambda-return$，即G_t^{λ}。其定义为：

$$G_t^{\lambda} = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{(n)}$$

这里是把$TD(\lambda)$算法视为平均n步更新的一种特例。这里的平均值包括了所有可能的n步更新，每一个都按照比例λ^{n-1}进行加权，其中$\lambda \in [0,1]$，最后乘上正则项$1-\lambda$保证权值和为1。可视化如下：



$\lambda - return$中每一个n步回报的权重如下所示：



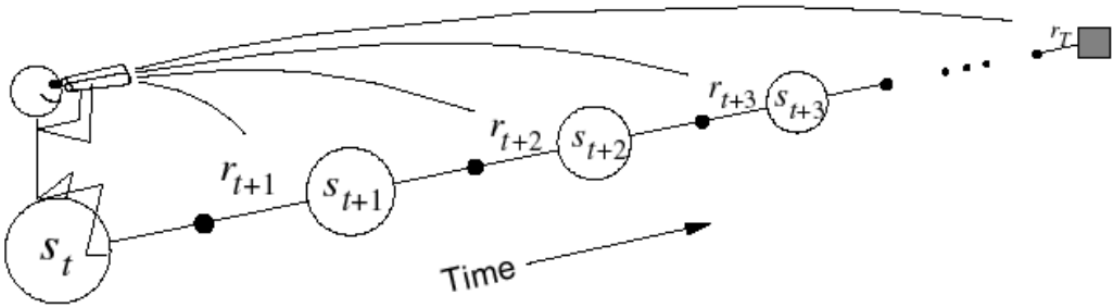
其中单步回报获得了最大的权值$(1-\lambda)$，两步回报为$(1-\lambda)\lambda$，三步回报为$(1-\lambda)\lambda^2$，以此类推进行衰减。

所以前向的$TD(\lambda)$算法使用G_t^λ替换target，得到价值函数的更新为：

$$V(S_t) \leftarrow V(S_t) + \alpha\left(G_t^\lambda - V(S_t)\right)$$

3、$TD(\lambda)$的前向视图

之前我们介绍的所有算法，理论上都是前向的(Forward-view)。对于访问的每一个状态，我们向前（未来的方向）探索所有可能的收益并决定如何将他们提供的信息进行有效结合利用。如下所示：



我们可以想象处于一个状态流之中，从每一个状态向前看并决定如何更新这个状态。每次更新完一个状态，移动到下一个状态并且不再更新以前经过的路径的状态。在另一个方面，未来的状态会从之前的位置被重复地观测并被处理。

引入了λ之后，会发现要更新一个状态的状态价值$V(S_t)$，必须要走完整个episode获得每一个状态的即时奖励以及最终状态获得的即时奖励。这和MC算法的要求一样，因此$TD(\lambda)$算法有着和MC方法一样的劣势。

- TD更新为：

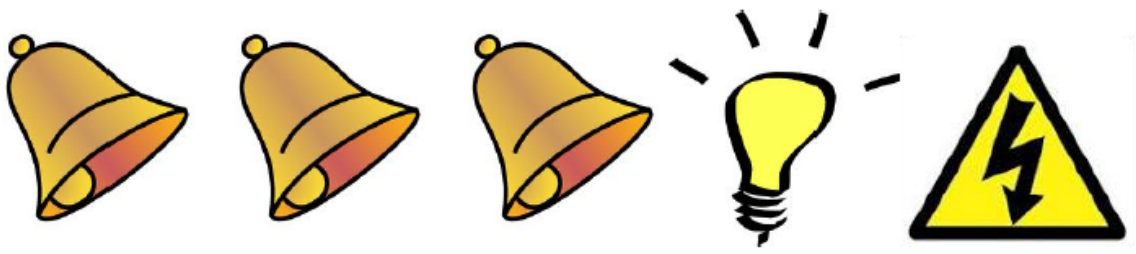
$$V(S_t) \leftarrow V(S_t) + \alpha\left(G_t^\lambda - V(S_t)\right)$$
$$G_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{(n)}$$
$$G_t^{(n)} = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{n-1}R_{t+n} + \gamma^n V(S_{t+n})$$

- MC更新为：

$$V(S_t) \leftarrow V(S_t) + \alpha\left(G_t - V(S_t)\right)$$

4、资格迹（Eligibility Traces）

考虑如下一个问题，老鼠在连续接受了3次响铃和1次亮灯信号后遭到了电击，那么在分析遭电击的原因时，到底是响铃的因素较重要还是亮灯的因素更重要呢？



问题的归因可以考虑两种情况：

- 频率启发（Frequency heuristic）：将原因归因于出现频率最高的状态，所以老鼠被点击的主要原因会是铃铛。

- 就近启发 （Recency heuristic）：将原因归因于较近的几次状态，这种考虑之下，老鼠被点击的主要原因会是灯。

结合频率启发（Frequency heuristic）与就近启发 （Recency heuristic）两种思想，可以引出资格迹（Eligibility Traces）。

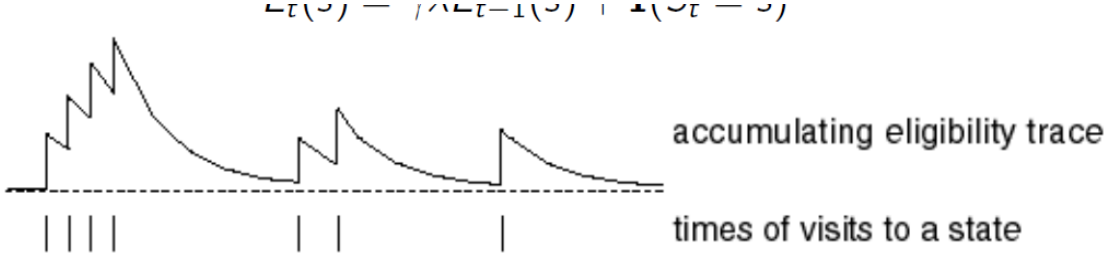
定义如下：

$$E_0(s) = 0$$
$$E_t(s) = \gamma\lambda E_{t-1}(s) + \mathbf{1}\left(S_t = s\right)$$

其中$\mathbf{1}\left(S_t = s\right)$是一个条件判断，可以改写$E_t(s)$为：

$$E_t(s) = \begin{cases} \gamma\lambda E_{t-1} & \text{if } S_t \neq s \\ \gamma\lambda E_{t-1} + 1 & \text{if } S_t = s \end{cases}$$

直观上如下图：



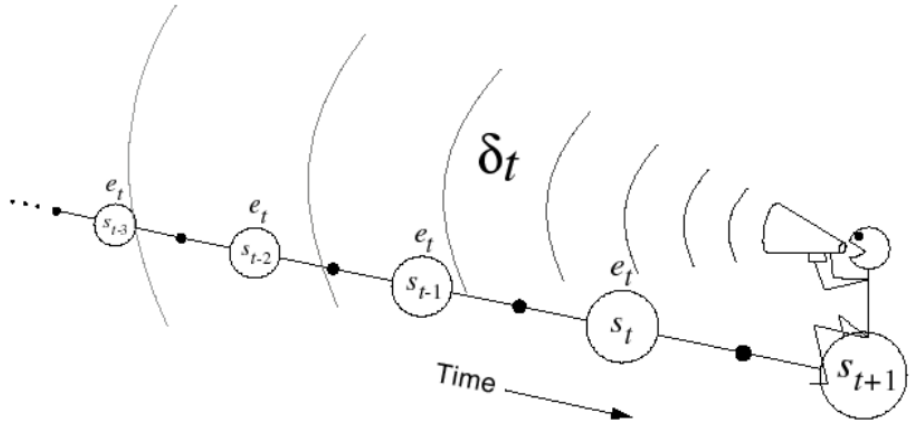
该图横坐标是时间，横坐标下有竖线的位置代表当前进入了状态s，纵坐标是资格迹E，可以看出当某一状态连续出现， E值会在一定衰减的基础上有一个单位数值的提高，此时将增加该状态对于最终收获贡献的比重，因而在更新该状态价值的时候可以较多地考虑最终收获的影响。同时如果该状态距离最终状态较远，则其对最终收获的贡献越小，在更新该状态时也不需要太多的考虑最终收获。

资格迹的提出是基于一个**信度分配（Credit Assignment）**问题的，打个比方，最后我们去跟别人下围棋，最后输了，那到底该中间我们下的哪一步负责？或者说，每一步对于最后我们输掉比赛这个结果，分别承担多少责任？这就是一个信度分配问题。对于小鼠问题，小鼠先听到三次铃声，然后看见灯亮，接着就被电击了，小鼠很生气，它仔细想，究竟是铃声导致的它被电击，还是灯亮导致的呢？如果按照事件的发生频率来看，是铃声导致的，如果按照最近发生原则来看，那就是灯亮导致的，但是，更合理的想法是，这二者共同导致小鼠被电击了，于是小鼠为这两个事件分别分配了权重，如果某个事件s发生，那么s对应的资格迹的值就加1，如果在某一段时间s未发生，则按照某个衰减因子进行衰减，这也就是上面的资格迹的计算公式了。

资格迹E值并不需要等到完整的episode结束才能计算出来，它可以每经过一个时刻就得到更新。

5、$TD(\lambda)$的后向视图

后向视角使用了我们刚刚定义的资格迹，每个状态s都保存了一个资格迹。我们可以将资格迹理解为一个权重，状态s被访问的时间离现在越久远，其对于值函数的影响就越小，状态s被访问的次数越少，其对于值函数的影响也越小。



$TD(\lambda)$的后向视角解释：有个人坐在状态流上，手里拿着话筒，面朝着已经经历过的状态获得当前回报并利用下一个状态的值函数得到TD偏差之后，此人会向已经经历过的状态喊话告诉这些已经经历过的状态处的值函数需要利用当前时刻的TD偏差进行更新。此时过往的每个状态值函数更新的大小应该跟距离当前状态的步数有关。

假设当前状态为s，TD偏差为δ_t，那么s_{t-1}处的值函数更新应该乘以一个衰减因子$\gamma\lambda$，状态 s_{t-2}处的值函数更新应该乘以 $(\gamma\lambda)^2$，以此类推。状态价值的更新为：

$$
\begin{aligned}
&\delta_t = R_{t+1} + \gamma V\left(S_{t+1}\right) - V\left(S_t\right) \\
&V(s) \leftarrow V(s) + \alpha \delta_t E_t(s) \\
&\text{其中 } E_t(s) = \begin{cases} \gamma\lambda E_{t-1} & \text{if } S_t \neq s \\ \gamma\lambda E_{t-1} + 1 & \text{if } S_t = s \end{cases}, \quad E_0(s) = 0
\end{aligned}
$$

6、$TD(\lambda)$的前向视图与后向视图的关系

$TD(\lambda)$与TD（0）

当$\lambda = 0$的时候，$\gamma\lambda = 0$。只有当前状态会得到更新，资格迹只会记录脉冲信号。其等价于$TD(0)$算法：

$$
\begin{aligned}
&E_t(s) = \mathbf{1}\left(S_t = s\right) \\
&V(s) \leftarrow V(s) + \alpha \delta_t E_t(s) \quad \left(\text{其中，} \delta_t = R_{t+1} + \gamma V\left(S_{t+1}\right) - V\left(S_t\right)\right) \\
&V\left(S_t\right) \leftarrow V\left(S_t\right) + \alpha \delta_t \\
&V\left(S_t\right) \leftarrow V\left(S_t\right) + \alpha(R_{t+1} + \gamma V\left(S_{t+1}\right) - V\left(S_t\right))
\end{aligned}
$$

$TD(1)$与MC

当$\lambda = 1$的时候，信度分配只会在episode结束的时候才会被定义，$TD(\lambda)$与MC将会等价。

理论上，对于状态s的总更新量前向视角和后向视角是等价的，如下等式的左边为后向视角的总更新量，等式右边为前向视角的总更新量。

$$
\sum_{t=1}^{T} \alpha \delta_t E_t(s) = \sum_{t=1}^{T} \alpha \left(G_t^{\lambda} - V\left(S_t\right)\right) \mathbf{1}\left(S_t = s\right)
$$

我们假设处在某个episode中，状态s在k时刻被访问了一次，那么$TD(1)$的资格迹会随时间进行衰减（在k时刻之前，资格迹E为0，自k时刻开始衰减）：

$$
\begin{aligned}
E_t(s) &= \gamma E_{t-1}(s) + \mathbf{1}\left(S_t = s\right) \\
&= \begin{cases} 0 & \text{if } t < k \\ \gamma^{t-k} & \text{if } t \geq k \end{cases}
\end{aligned}
$$

$TD(1)$的在线更新过程中，他的累积误差可以表述为：

$$\sum_{t=1}^{T-1}\alpha\delta_t E_t(s)=\alpha\sum_{t=k}^{T-1}\gamma^{t-k}\delta_t=\alpha\left(G_k-V\left(S_k\right)\right)$$

在episode介绍的时候总的累积误差是：

$$\delta_k+\gamma\delta_{k+1}+\gamma^2\delta_{k+2}+\ldots+\gamma^{T-1-k}\delta_{T-1}$$

当$\lambda=1$的时候，总的累积误差与MC误差的关系可以进行如下关联：

$$\begin{aligned}
&\delta_t+\gamma\delta_{t+1}+\gamma^2\delta_{t+2}+\ldots+\gamma^{T-1-t}\delta_{T-1}\\
=&R_{t+1}+\gamma V\left(S_{t+1}\right)-V\left(S_t\right)\\
+&\gamma R_{t+2}+\gamma^2 V\left(S_{t+2}\right)-\gamma V\left(S_{t+1}\right)\\
+&\gamma^2 R_{t+3}+\gamma^3 V\left(S_{t+3}\right)-\gamma^2 V\left(S_{t+2}\right)\\
&\vdots\\
+&\gamma^{T-1-t}R_T+\gamma^{T-t}V\left(S_T\right)-\gamma^{T-1-t}V\left(S_{T-1}\right)\\
=&R_{t+1}+\gamma R_{t+2}+\gamma^2 R_{t+3}\ldots+\gamma^{T-1-t}R_T-V\left(S_t\right)\\
=&G_t-V\left(S_t\right)
\end{aligned}$$

上式推导过程中，只是简单展开后删除中间项，这样的结果中G_t相当于MC方法的总的回报，$V(S_t)$为当前的状态值函数，$G_t-V\left(S_t\right)$相当于蒙特卡洛的更新量。所以当$\lambda=1$时，TD总的累积误差会缩小为MC误差。

所以简单总结下：

- $TD(1)$和每次访问的蒙特卡洛方法是大致是等价的，不过也有区别。
- $TD(1)$是在线对误差进行累积，每步都会更新。
- $TD(1)$如果也等到episode结束后离线更新，那么$TD(1)$和MC就完全等价。

$TD(\lambda)$的$\lambda-error$

对于一般的λ，不是极端的0与1之外的情况，我们也可以证明总误差等价于$G_t^\lambda-V\left(S_t\right)$。

$$\begin{aligned}
G_t^\lambda-V\left(S_t\right)=&-V\left(S_t\right)+(1-\lambda)\lambda^0\left(R_{t+1}+\gamma V\left(S_{t+1}\right)\right)\\
&+(1-\lambda)\lambda^1\left(R_{t+1}+\gamma R_{t+2}+\gamma^2 V\left(S_{t+2}\right)\right)\\
&+(1-\lambda)\lambda^2\left(R_{t+1}+\gamma R_{t+2}+\gamma^2 R_{t+3}+\gamma^3 V\left(S_{t+3}\right)\right)\\
&\cdots\\
=&-V\left(S_t\right)+(\gamma\lambda)^0\left(R_{t+1}+\gamma V\left(S_{t+1}\right)-\gamma\lambda V\left(S_{t+1}\right)\right)\\
&+(\gamma\lambda)^1\left(R_{t+2}+\gamma V\left(S_{t+2}\right)-\gamma\lambda V\left(S_{t+2}\right)\right)\\
&+(\gamma\lambda)^2\left(R_{t+3}+\gamma V\left(S_{t+3}\right)-\gamma\lambda V\left(S_{t+3}\right)\right)\\
&+\quad\cdots\\
=&\quad(\gamma\lambda)^0\left(R_{t+1}+\gamma V\left(S_{t+1}\right)-V\left(S_t\right)\right)\\
&+(\gamma\lambda)^1\left(R_{t+2}+\gamma V\left(S_{t+2}\right)-V\left(S_{t+1}\right)\right)\\
&+(\gamma\lambda)^2\left(R_{t+3}+\gamma V\left(S_{t+3}\right)-V\left(S_{t+2}\right)\right)\\
&+\ldots\\
=&\quad\delta_t+\gamma\lambda\delta_{t+1}+(\gamma\lambda)^2\delta_{t+2}+\ldots
\end{aligned}$$

前向视角和后向视角的$TD(\lambda)$

假设在某个episode中，状态s在k时刻被访问了一次，那么$TD(\lambda)$的资格迹会随时间进行衰减（在k时刻之前，资格迹E为0，自k时刻开始衰减）

$$
\begin{aligned}
E_t(s) &= \gamma\lambda E_{t-1}(s) + \mathbf{1}\left(S_t = s\right) \\
&= \begin{cases} 0 & \text{if } t < k \\ (\gamma\lambda)^{t-k} & \text{if } t \geq k \end{cases}
\end{aligned}
$$

后向视角的$TD(\lambda)$在这个过程中不断累积误差：

$$
\sum_{t=1}^{T} \alpha\delta_t E_t(s) = \alpha \sum_{t=k}^{T} (\gamma\lambda)^{t-k}\delta_t = \alpha\left(G_k^{\lambda} - V\left(S_k\right)\right)
$$

当整个片段完成时，后向视角方法对于值函数$V(s)$ 的增量等于$\lambda - return$；如果状态s被访问了多次，那么资格迹就会累积，从而相当于累积了更多的$V(s)$ 的增量。这直观地解释了前向视角和后向视角的等价性。

总结

离线更新	$\lambda = 0$	$\lambda \in (0,1)$	$\lambda = 1$
后向视角	TD(0)	TD(λ)	TD(1)
	⇕	⇕	⇕
前向视角	TD(0)	前向 TD(λ)	MC
在线更新	$\lambda = 0$	$\lambda \in (0,1)$	$\lambda = 1$
后向视角	TD(0)	TD(λ)	TD(1)
	⇕	⇕	⇕
前向视角	TD(0)	前向 TD(λ)	MC
	⇕	⇕	⇕
真实在线更新	TD(0)	真实在线 TD(λ)	真实在线 TD(1)