強化学習20151202

田中一樹

TD(λ)

- 1ステップのTD法(TD(0))

$$R_{t:0} = R_t + \gamma \hat{V}_t(X_{t+1})$$

- kステップのTD法

$$\mathrm{R}_{t:k} = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^k R_{t+k+1} + \gamma^{k+1} \hat{V}_t(X_{t+k+1})$$

X_tにおけるkステップ先を見た報酬は

$$\mathbb{E}[R_{t:k}] = \mathbb{E}[R_{t:0}]$$

しかし、これだと先の報酬を見てから現在の価値を更新するのでモンテカルロ法と同じく事後学習になる

λ収益: $R_t^\lambda = (1-\lambda) \sum_{k=0}^{\infty} \lambda^k R_{t:k}$

正規化係数(1-λ)があるので

$$\mathbb{E}[R_t] = \mathbb{E}[R_t^\lambda]$$

λ=0: TD(0)
λ=1: モンテカルロ法（無限回試行後の）

λ収益アルゴリズム

未来の報酬を用いて現在の価値を更新するアルゴリズム（更新則）

$$\hat{V}_{t+1}(X_t) = \hat{V}_t(X_t) + \alpha\left[R_t^{\lambda} - \hat{V}_t(X_t)\right]$$

- 下でTD(λ)がTD(0)とモンテカルロ法の中間ということがわかる

$$R_t^{\lambda} - \hat{V}_t(X_t) = (\gamma\lambda)^0\left[R_{t+1} + \gamma\hat{V}_t(X_{t+1}) - \hat{V}_t(X_t)\right]$$

XtでのTD(0)

$$+ (\gamma\lambda)^1\left[R_{t+2} + \gamma\hat{V}_t(X_{t+2}) - \hat{V}_t(X_{t+1})\right]$$

Xt+1でのTD(0)

$$+ \cdots$$

$$\approx \sum_{k=t}^{\infty} (\gamma\lambda)^{k-t}\delta_k$$

報酬の系列がわかってから更新
モンテカルロ法(事後学習)

これでは前方観測的 → 毎時刻ごとに学習を行うには過去の情報だけを用いる後方観測的な更新式の方が都合が良い

そこで適格度トレース! $z_t(x)$

$$\delta_{t+1} = R_{t+1} + \gamma \hat{V}_t(X_{t+1}) - \hat{V}_t(X_t),$$

適格度トレース

$$z_{t+1}(x) = \mathbb{I}_{\{x=X_t\}} + \gamma\lambda z_t(x),$$

$$\hat{V}_{t+1}(x) = \hat{V}_t(x) + \alpha_t \delta_{t+1} z_{t+1}(x),$$

$$z_0(x) = 0,$$

$$x \in \mathcal{X}.$$

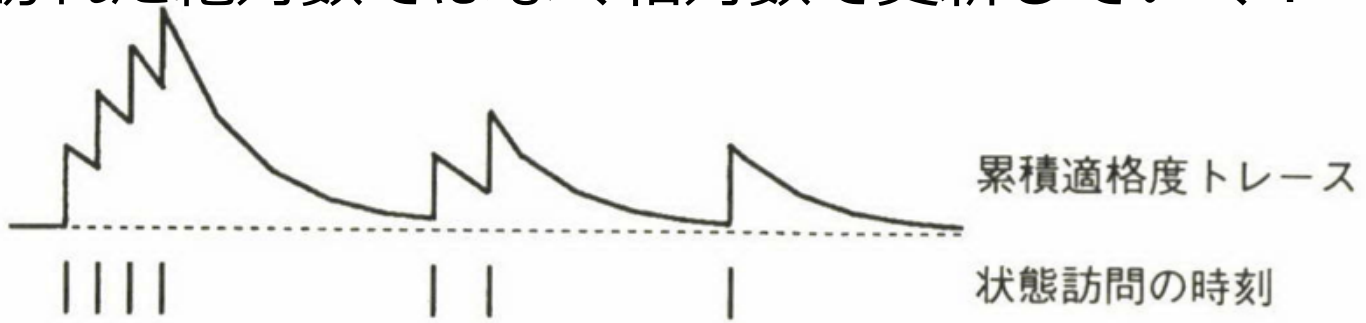
こういう風に更新式を作れば、
前方的な更新式と等価である
ということが証明できる
(Sutton, p189-)

累積トレース

X_tを一回訪問すれば1増加、それ以外は減衰

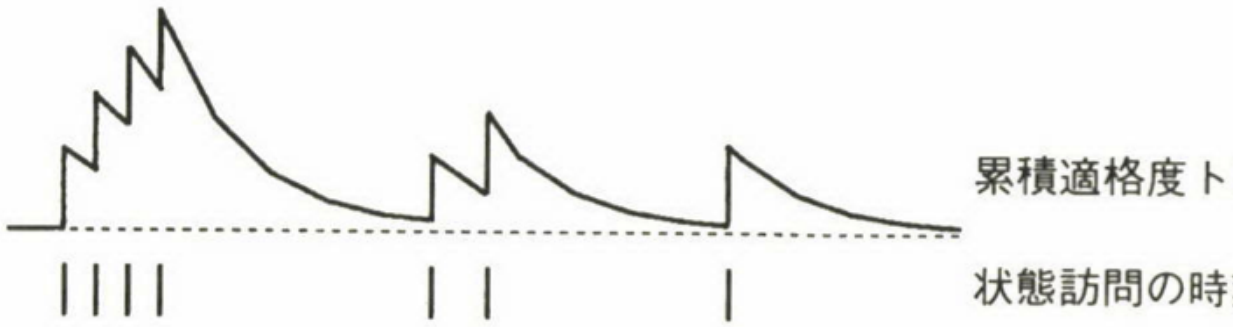
➡ 訪れたX_tのみだけでなく、一回訪れた価値も同時に更新

➡ 訪れた絶対数ではなく相対数で更新していく？



証明イメージ(γλが出てくるイメージ)

$$
\begin{aligned}
R_t - \hat{V}_t(X_t) = & - \hat{V}_t(X_t) \\
& + (1-\lambda)\lambda^0[R_{t+1} + \gamma \hat{V}_t(X_{t+1})] \\
& + (1-\lambda)\lambda^1[R_{t+1} + \gamma R_{t+2} + \gamma^2 \hat{V}_t(X_{t+2})] \\
& \cdots \\
= & (\gamma\lambda)^0[R_{t+1}\gamma \hat{V}_t(X_{t+1}) - \hat{V}_t(X_t) \\
& + (\gamma\lambda)^1[R_{t+2}\gamma \hat{V}_t(X_{t+2}) - \hat{V}_t(X_{t+1}) \\
& + \cdots \\
\approx & \sum_{k=t}^{\infty} (\gamma\lambda)^{k-t}\delta_k
\end{aligned}
$$



K個先まで見た収益$R_{t:k}$

$$R_{t:0} = R_{t+1} + \gamma \hat{V}_t(X_{t+1})$$

$$R_{t:1} = R_{t+1} + \gamma R_{t+2} \gamma^2 + \hat{V}_t(X_{t+2})$$

$$\cdots$$

$$R_{t:k} = R_{t+1} + \gamma R_{t+2} + \cdots \gamma^k R_{t+k+1} \gamma^{k+1} \hat{V}_t(X_{t+k+1})$$

$$\cdots$$

$$R_t = (1-\lambda) \sum_{k=0}^{\infty} \lambda^k R_{t:k}$$

重みλ^kで
正規化係数1-λ

K個先までの報酬を見て
現在の価値観数を更新する