強化学習メモ3.2.3

2015/12/16(水) 19:00~ 担当：高山 晃一

前節3.2.2：TD(λ)の解析と**改良**

TD(λ)は方策オフで解が不安定 ⇒ 擬勾配降下法で安定化

- 本文中ではλ = 0のときの解析と収束を紹介

コスト関数 $$J(\theta) = \|V_\theta - \Pi_{\mathcal{F},\nu} T V_\theta\|_\nu^2 = \mathbb{E}[\delta_{t+1}(\theta)\varphi_t]^\top \mathbb{E}[\varphi_t\varphi_t^\top]^{-1}\mathbb{E}[\delta_{t+1}(\theta)\varphi_t]$$

勾配 $$\nabla_\theta J(\theta) = -2\mathbb{E}[(\varphi_t - \gamma\varphi'_{t+1})\varphi_t^\top]\mathbb{E}[\varphi_t\varphi_t^\top]^{-1}\mathbb{E}[\delta_{t+1}(\theta)\varphi_t]$$

仮想的な重み $$w(\theta) = \mathbb{E}[\varphi_t\varphi_t^\top]^{-1}\mathbb{E}[\delta_{t+1}(\theta)\varphi_t]$$

・GTD2 (gradient temporal difference learning)

$$\nabla_\theta J(\theta) = -2\mathbb{E}[(\varphi_t - \gamma\varphi'_{t+1})\varphi_t^\top]w(\theta)$$

$$\theta_{t+1} = \theta_t + \alpha_t\left(\varphi_t - \gamma\varphi'_{t+1}\right)\varphi_t^\top w_t$$

・TDC (temporal difference learning with correlations)

$$\nabla_\theta J(\theta) = -2\Big(\mathbb{E}[\delta_{t+1}(\theta)\varphi_t] - \gamma\mathbb{E}[\varphi'_{t+1}\varphi_t^\top]w(\theta)\Big)$$

$$\theta_{t+1} = \theta_t + \alpha_t\left(\delta_{t+1}(\theta_t)\varphi_t - \gamma\varphi'_{t+1}\right)\varphi_t^\top w_t$$

$$w_{t+1} = w_t + \beta\left(\delta_{t+1}(\theta_t) - \varphi_t^\top w_t\right)\varphi_t$$

本節3.2.3：TD(λ)の**別解釈**と改良

擬勾配降下法の更新はAdaptive filteringのLMSに類似

・LMS (Least-means squares method)　入力φ, 出力y

コスト関数	$J(\theta) = \mathbb{E}[(y - \theta\varphi)^2]$
勾配	$\nabla_\theta J(\theta) = \mathbb{E}[\underline{(y - \theta\varphi)}\varphi]$
更新	$\theta_{t+1} = \theta_t + \alpha_t \underline{(y - \theta_t\varphi_t)}\varphi_t$

・TDC (temporal difference learning with correlations)

コスト関数	$J(\theta) = \|\|V_\theta - \Pi_{\mathcal{F},\nu} T V_\theta\|\|_\nu^2$ $= \mathbb{E}[\delta_{t+1}(\theta)\varphi_t]^\top \mathbb{E}[\varphi_t\varphi_t^\top]^{-1}\mathbb{E}[\delta_{t+1}(\theta)\varphi_t]$
勾配	$\nabla_\theta J(\theta) = -2\mathbb{E}[\underline{(\varphi_t - \gamma\varphi'_{t+1})\varphi_t^\top}]w(\theta)$
θの更新	$\theta_{t+1} = \theta_t + \alpha_t \underline{\left(\varphi_t - \gamma\varphi'_{t+1}\right)\varphi_t^\top} w_t$
仮想的な重み	$w(\theta) = \mathbb{E}[\varphi_t\varphi_t^\top]^{-1}\mathbb{E}[\delta_{t+1}(\theta)\varphi_t]$

⇒ 性能がステップ幅や行列A(P35)の固有値に敏感

本節3.2.3：TD(λ)の別解釈と**改良**

TD(λ)は収束が遅い&GTD系は調整がつらい ⇒ 最小二乗法

・LSTD (Least-squares temporal difference learning)

TD(0)の更新 $$\theta_{t+1} - \theta_t = \alpha_t \delta_{t+1}(\theta_t)\varphi_t$$

収束値 θ^*での振る舞い $$\mathbb{E}[\delta_{t+1}(\theta^*)\varphi_t] = 0$$

標本近似 $$\frac{1}{n}\sum_{t=0}^{n-1}\varphi_t\delta_{t+1}(\theta) = 0$$

・LSTD(λ)

TD(λ)の更新 $$\theta_{t+1} - \theta_t = \alpha_t \delta_{t+1}(\theta_t) z_{t+1}$$

標本近似 $$\frac{1}{n}\sum_{t=0}^{n-1} z_{t+1}\delta_{t+1}(\theta) = 0$$

適格度トレース $$\begin{aligned} z_{t+1} &= \nabla_\theta V_{\theta_t}(\varphi_t) + \gamma\lambda z_t \\ &= \varphi_t + \gamma\lambda z_t \end{aligned}$$

本節3.2.3：TD(λ)の別解釈と**改良**

LSTD系は計算量がつらい ⇒ 再帰的計算により回避

$O(nd^2 + d^3)$ $O(nd^2)$

TD(λ)は***O*(nd)**？

・RLSTD (Recursive LSTD)

$$\hat{A}_t = \frac{1}{t}\sum_{i=0}^{t-1} \varphi_i(\varphi_i - \gamma\varphi'_{i+1})^\top$$

$$A'_t = t\hat{A}_t$$

Sheman-Morrison: $$A'^{-1}_{t+1} = A'^{-1}_t - \frac{A'^{-1}_t \varphi_t(\varphi_t - \gamma\varphi'_{t+1})^\top A'^{-1}_t}{1 + (\varphi_t - \gamma\varphi'_{t+1})A'^{-1}_t \varphi_t}$$

$$C_t = A'^{-1}_t$$

$$C_{t+1} = C_t - \frac{C_t\varphi_t(\varphi_t - \gamma\varphi'_{t+1})^\top C_t}{1 + (\varphi_t - \gamma\varphi'_{t+1})^\top C_t\varphi_t}$$

$$\theta_{t+1} = \theta_t + \frac{C_t}{1 + (\varphi_t - \gamma\varphi'_{t+1})^\top C_t\varphi_t}\delta_{t+1}(\theta_t)\varphi_t$$

比較 (仮)

LSTD(λ)

<on>

TD(λ)より収束が早い

標本への標準的な仮定下で概収束

<off>

TD(λ)より収束が早い

標本仮定&極限解があれば概収束

解がill-definedな可能性がある

λ-LSPE

<on>

性能はLSTD(λ)に匹敵

標本仮定&ステップ幅条件下で概収束

<off>

性能はLSTD(λ)に匹敵？

収束は述べていない？

解は常にwell-defined

TD(λ)

<on>

収束のboundあり(but, slow)

<off>

発散しうる

GTD系: tuning is necessary

<on>

RM, ステップ幅, +α条件下で概収束

<off>

RM, ステップ幅, +α条件下で概収束？