

ベルマン方程式とその証明

状態価値関数の定義式を用いて、下記のベルマン方程式が成立することを示す

$$V_{\pi}(s) = \sum_a \pi(s, a) \sum_{s'} P(s_{t+1} = s' | s_t = s, a_t = a) [r_{t+1} + \gamma V_{\pi}(s')]$$

ここで状態価値関数(状態 s から始め方策 π に従うときに得られる割引累積報酬の期待値)の定義は

$$V_{\pi}(s) = E_{\pi}[R_t | s_t = s] = E_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right] \quad (7.3)$$

関係する式

報酬 $r_{t+1} = r(s_t, a_t)$

方策 $\pi(s, a) = P(a_t = a | s_t = s)$ (時刻 t によらず) 状態 s において行動 a をとる確率 (7.3)

$$\text{割引累積報酬 } R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \quad (7.4)$$

$P(s_{t+1} = s' | s_t = s, a_t = a)$: 状態 s で行動 a をとったときに次状態が s' となる状態遷移確率

証明:

状態価値関数の定義から

$$\begin{aligned} V_{\pi}(s) &= E_{\pi}[R_t | s_t = s] = E_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right] \\ &= E_{\pi}\left[r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} | s_t = s\right] \quad \text{累積割引報酬の定義から} \\ &= E_{\pi}[r_{t+1} | s_t = s] + \gamma E_{\pi}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} | s_t = s\right] \quad \text{期待値の計算式を二つに分割} \quad (A) \end{aligned}$$

(A)の第1項について: (確認: t は時刻であり、 s_t は時刻 t における状態、 a_t はその時の行動を表す)

報酬 $r_{t+1} = r(s_t, a_t)$ は状態 s_t において行動 a_t をとり状態遷移確率 $P(s_{t+1} | s_t, a_t)$ によって次状態(s_{t+1})になったときに得られる報酬(時刻 t における状態とその行動に対する報酬であり、 $t+1$ の状態に対する報酬ではないことに注意)。このことを明確にするため、報酬関数 $r(s', s, a)$ を定義する。

$r(s', s, a)$ は「状態 s で行動 a をとり、次状態が s' になったときの報酬」

ここで S を「状態 s で行動 a により遷移する状態すべての集合」とすれば、(次状態が s' になるかどうかは状態遷移確率で決まるので $r(s', s, a)$ にその確率を掛け算したものの総和)

$\sum_{s' \in S} P(s_{t+1} = s' | s_t = s, a_t = a) r(s', s, a)$ は「(時刻 t における)状態 s で行動 a をとったときの報酬の期待値」を表す。

さて、状態 $s_t = s$ で取れる行動はいろいろあるので、選択可能な行動の集合を A で表し $a \in A$ とすると、「状態 s で行動 a をとる」確率が方策 $\pi(s, a)$ であるので、「時刻 t で状態 $s_t = s$ にいるという条件のもとで、なんらかの行動を取って得られる報酬の期待値」は

$$\sum_{a \in A} \pi(s, a) \sum_{s' \in S} P(s_{t+1} = s' | s_t = s, a_t = a) r(s', s, a)$$

となる。

ここで改めて $\pi(s, a) = P(a_t = a | s_t = s)$ であり、 $\sum_{a \in A}$ を $\sum_a$ と書き、 $\sum_{s' \in S}$ を $\sum_{s'}$ として

$r(s', s, a) = r_{t+1}(= r(s, a))$ で書き直すと、「状態 $s_t = s$ から $t+1$ の状態にいたときの報酬 r_{t+1} 」の期待値は次のように表されることが示せた:

(正直言えば、 s と a の関数 $r(s, a)$ を定数のような r_{t+1} で表すのは気持ちが悪い)

$$\begin{aligned} E_\pi[r_{t+1} | s_t = s] &= \sum_{a \in A} \pi(s, a) \sum_{s' \in S} r(s', s, a) \\ &= \sum_a P(a_t = a | s_t = s) \sum_{s'} P(s_{t+1} = s' | s_t = s, a_t = a) r_{t+1} \end{aligned}$$

(A)の第2項について

$$\gamma E_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} | s_t = s \right]$$

これは状態価値関数の定義と比較すると分かるようによく似ている(r_{t+k+1} ではなく r_{t+k+2} であることが違う) $V_\pi(s) = E_\pi[R_t | s_t = s] = E_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s \right]$ (7.3再掲)

これを意識して式を変形していく。

ここで状態 $s_t = s$ から何らかの行動によって次状態 $s_{t+1} = s'$ になることを考える。すると、状態 $s_t = s$ から状態 $s_{t+1} = s'$ に遷移する確率を $P(s_{t+1} = s' | s_t = s)$ と表し、 S を「状態 s で行動 a により遷移する状態すべての集合」とすると、

$$E_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} | s_t = s \right] = \sum_{s' \in S} E_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} | s_{t+1} = s' \right] P(s_{t+1} = s' | s_t = s)$$

と書ける。ここで状態価値関数の定義から $E_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} | s_{t+1} = s' \right] = V_\pi(s')$ (もうちょっと書く

と、 $t+1$ を T と書き直すと、 $E_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} | s_{t+1} = s' \right] = E_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{T+k+1} | s_T = s' \right]$) であるた

め)。そして、掛け算の順番を変え、 $\sum_{s' \in S}$ を $\sum_{s'}$ と書き直すと次のように書ける:

$$E_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} | s_t = s \right] = \sum_{s'} P(s_{t+1} = s' | s_t = s) V_\pi(s')$$

また状態 $s_t = s$ から次状態 $s_{t+1} = s'$ になるには行動を選択しなければならず、しかもそのような行動は一つとは限らないので、選択可能な行動の集合を A で表すと

$$P(s_{t+1} = s' | s_t = s) = \sum_{a \in A} P(a_t = a | s_t = s) P(s_{t+1} = s' | s_t = s, a_t = a) \quad (\text{周辺化})$$

が成り立つ。よって、 $\pi(s, a) = P(a_t = a | s_t = s)$ を用い、 $\sum_{a \in A}$ を $\sum_a$ と書き換えれば

$$\begin{aligned} \gamma E_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+2} | s_t = s \right] &= \sum_{s'} \sum_a P(a_t = a | s_t = s) P(s_{t+1} = s' | s_t = s, a_t = a) V_\pi(s') \\ &= \gamma \sum_a \pi(s, a) \sum_{s'} P(s_{t+1} = s' | s_t = s, a_t = a) V_\pi(s') \end{aligned}$$

($\pi(s, a)$ は s' を含まないので外にだした)

以上から、二つの式を合わせて

$$\begin{aligned}
V_\pi(s) &= E_\pi[r_{t+1}|s_t = s] + \gamma E_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+2}|s_t = s\right] \\
&= \sum_a \pi(s, a) \sum_{s'} P(s_{t+1} = s'|s_t = s, a_t = a) r_{t+1} + \gamma \sum_a \pi(s, a) \sum_{s'} P(s_{t+1} = s'|s_t = s, a_t = a) V_\pi(s') \\
&= \sum_a \pi(s, a) \sum_{s'} P(s_{t+1} = s'|s_t = s, a_t = a) [r_{t+1} + \gamma V_\pi(s')]
\end{aligned}$$

以上により証明された。