# 強化学習　第6章
# 「関数近似を用いた強化学習」

価値関数：

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a|s)\{g(s,a) + \gamma \sum_{s' \in \mathcal{S}} p_T(s'|s,a) V^{\pi}(s')\}, \quad \forall s \in \mathcal{S}$$

は，状態行動空間が膨大である場合，テーブルの要素数が大きくなりすぎてしまい学習が困難となる．そこで，価値関数を関数近似器を用い近似することを考える．

## 1　価値関数の関数近似

$V^{\pi}$ をパラメータ $\boldsymbol{w} \in \mathcal{R}^d$ で規定される関数近似器 $\hat{V} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$ を用いて近似する．すなわち，

$$V^{\pi}(s) \approx \hat{V}(s; \boldsymbol{w}^*)$$

なるようなパラメータ $\boldsymbol{w}^*$ を学習することを考える．ここで，関数近似器 $\hat{V}(s; \boldsymbol{w})$ の集合を

$$\mathcal{V} \coloneqq \{\hat{V}_{\boldsymbol{w}}; \boldsymbol{w} \in \mathbb{R}^d\}, \quad (\hat{V}_{\boldsymbol{w}}(s) \coloneqq \hat{V}(s; \boldsymbol{w}))$$

と置くと，$\mathcal{V}$ は状態関数空間の部分空間である．($\mathcal{V} \subset \mathbb{R}^{\mathcal{S}}$ となる．$\mathcal{V} \equiv \mathbb{R}^{\mathcal{S}}$ でないことが重要．)

### 1.1　線形近似

線形関数近似器として，

$$\hat{V}_{\boldsymbol{w}}(s) \coloneqq {}^{\top}\boldsymbol{w}\boldsymbol{\phi}(s)$$

がある．ただし，$\boldsymbol{\phi} : \mathcal{S} \to \mathbb{R}^d$ は状態 $s$ の特徴ベクトル．例えば

$$\boldsymbol{w} \coloneqq \begin{pmatrix} w_1 \\ \vdots \\ w_{|\mathcal{S}|} \end{pmatrix}, \quad \boldsymbol{\phi}(s) \coloneqq \begin{pmatrix} \phi_1(s) \\ \vdots \\ \phi_{|\mathcal{S}|}(s) \end{pmatrix} \in \mathbb{R}^{|\mathcal{S}| \times d}$$

とし，$\phi_i(s) = \mathbb{I}_{\{i=s\}}$ とおけば，$\hat{V}_{\boldsymbol{w}}(s) = \boldsymbol{w}_s, \ \ \forall s \in \mathcal{S}$ はテーブル形式の関数と同値．(関数近似の意味なし)

## 1.2 テーブル形式方法の拡張

モデルフリー型強化学習で $V^\pi$ や $V^*$ を求めたことを関数近似へと拡張する．

### 1.2.1 バッチ学習の場合

これまで，価値関数を求める際，なにかしらのベルマン作用素 $B$(履歴データから近似されるものも含む) を状態関数 $v$ に作用させることで，

$$v'(s) \coloneqq Bv(s), \ \ \forall s \in \mathcal{S}$$

と更新してきた．一般に，$\mathcal{V} \subset \mathbb{R}^{\mathcal{S}}$ より，関数近似器 $\hat{V}_{\boldsymbol{w}}$ にベルマン作用素を適用することで部分空間 $\mathcal{V}$ の外に出てしまう場合がある．

$$\exists \hat{V}_{\boldsymbol{w}} \in \mathcal{V}, \quad B\hat{V}_{\boldsymbol{w}} \notin \mathcal{V}$$

このような場合，

$$\hat{V}'_{\boldsymbol{w}}(s) \coloneqq B\hat{V}_{\boldsymbol{w}}(s), \ \ \forall s \in \mathcal{S}$$

と従来の更新則を実行できない．そこで $B\hat{V}_{\boldsymbol{w}}$ を $\hat{V}_{\boldsymbol{w}}$ で近似するようにパラメータ $\boldsymbol{w}$ を更新するアプローチを考える．すなわち，

1. 各状態 $s \in \mathcal{S}$ に対し目的関数を算出する．(fixing target)

$$V^{\text{target}}(s) \coloneqq B\hat{V}_{\boldsymbol{w}}(s)$$

2. $\boldsymbol{w} \in \mathbb{R}^d$ を次のように更新する．

$$\boldsymbol{w} \coloneqq \arg\min_{\boldsymbol{w} \in \mathbb{R}^d} \sum_{s \in \mathcal{S}} (V^{\text{target}}(s) - \hat{V}_{\boldsymbol{w}}(s))^2$$

例えば，関数近似器 $\hat{V}_{\boldsymbol{w}}$ が線形モデルの場合，

$$\begin{aligned}\hat{V}_{\boldsymbol{w}}(s) &= {}^{\top}\boldsymbol{w}\boldsymbol{\phi}(s) = \begin{pmatrix} w_1 & \dots & w_{|\mathcal{S}|} \end{pmatrix} \begin{pmatrix} \phi_1(s) \\ \vdots \\ \phi_{|\mathcal{S}|}(s) \end{pmatrix}, \quad \forall s \in \mathcal{S} \\ \boldsymbol{v}^{\text{target}} &= \begin{pmatrix} V^{\text{target}}(1) \\ \vdots \\ V^{\text{target}}(|\mathcal{S}|) \end{pmatrix}\end{aligned}$$

とおけば，二乗誤差を微分して $\mathbf{0}$ とおけば

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{w}} \sum_{s \in \mathcal{S}} (V^{\text{target}}(s) - {}^{\top}\boldsymbol{w}\boldsymbol{\phi}(s))^2 &\propto \sum_{s \in \mathcal{S}} (V^{\text{target}}(s) - {}^{\top}\boldsymbol{w}\boldsymbol{\phi}(s))^{\top}\boldsymbol{\phi}(s) \\ &= \sum_{s \in \mathcal{S}} V^{\text{target}}(s)^{\top}\boldsymbol{\phi}(s) - {}^{\top}\boldsymbol{w} \sum_{s \in \mathcal{S}} \boldsymbol{\phi}(s)^{\top}\boldsymbol{\phi}(s) = \mathbf{0}\end{aligned}$$

ここで，

$$\begin{aligned}\sum_{s \in \mathcal{S}} V^{\text{target}}(s)^{\top}\boldsymbol{\phi}(s) &= \sum_{s \in \mathcal{S}} V^{\text{target}}(s) \begin{pmatrix} \phi_1(s) & \dots & \phi_{|\mathcal{S}|}(s) \end{pmatrix} \\ &= \begin{pmatrix} V^{\text{target}}(1) & \dots & V^{\text{target}}(|\mathcal{S}|) \end{pmatrix} \begin{pmatrix} \phi_1(1) & \dots & \phi_{|\mathcal{S}|}(1) \\ \vdots & \ddots & \vdots \\ \phi_1(|\mathcal{S}|) & \dots & \phi_{|\mathcal{S}|}(|\mathcal{S}|) \end{pmatrix} \\ \sum_{s \in \mathcal{S}} \boldsymbol{\phi}(s)^{\top}\boldsymbol{\phi}(s) &= \sum_{s \in \mathcal{S}} \begin{pmatrix} \phi_1(s) \\ \vdots \\ \phi_{|\mathcal{S}|}(s) \end{pmatrix} \begin{pmatrix} \phi_1(s) & \dots & \phi_{|\mathcal{S}|}(s) \end{pmatrix} \\ &= \begin{pmatrix} \phi_1(1) & \dots & \phi_1(|\mathcal{S}|) \\ \vdots & \ddots & \vdots \\ \phi_{|\mathcal{S}|}(1) & \dots & \phi_{|\mathcal{S}|}(|\mathcal{S}|) \end{pmatrix} \begin{pmatrix} \phi_1(1) & \dots & \phi_{|\mathcal{S}|}(1) \\ \vdots & \ddots & \vdots \\ \phi_1(|\mathcal{S}|) & \dots & \phi_{|\mathcal{S}|}(|\mathcal{S}|) \end{pmatrix}\end{aligned}$$

であるから

$$\Phi \coloneqq \begin{pmatrix} \phi_1(1) & \dots & \phi_{|\mathcal{S}|}(1) \\ \vdots & \ddots & \vdots \\ \phi_1(|\mathcal{S}|) & \dots & \phi_{|\mathcal{S}|}(|\mathcal{S}|) \end{pmatrix}$$

とおけば (計画行列)，

$$\begin{aligned}\mathbf{0} &= {}^{\top}\boldsymbol{v}^{\text{target}}\Phi - {}^{\top}\boldsymbol{w}^{\top}\Phi\Phi \\ \therefore \boldsymbol{w} &= ({}^{\top}\Phi\Phi)^{-1\top}\Phi\boldsymbol{v}^{\text{target}}\end{aligned}$$

と解析的に $\boldsymbol{w}$ を求めることができる．(最小二乗法)
しかし，この方法では結局全ての $s \in \mathcal{S}$ に対し計算を行うため，関数近似の意味がない．そこで，状態の単位ではなく経験 $(s, a, r, s')$ の単位で標本を準備して $\boldsymbol{w}$ を学習する適合価値反復法がある．履歴データを

$$\mathcal{D} \coloneqq \{(s_{(1)}, a_{(1)}.r_{(1)}, s'_{(1)}), \ldots, (s_{(N)}, a_{(N)}.r_{(N)}, s'_{(N)})\}$$

と置き $(1, ..., N$ は時間ステップとは関係なし)，適合価値反復法の例を二つ示す．

- 価値関数 $V^\pi$ の推定
  1. 各経験の目的変数を算出

$$V_{(n)}^{\text{target}} \coloneqq \hat{B}(\hat{V}; \{s_{(n)}, r_{(n)}, s'_{(n)}\})(s_{(n)}) = r_{(n)} + \gamma \hat{V}_{\boldsymbol{w}}(s'_{(n)})$$

  2. パラメータ $\boldsymbol{w}$ を更新

$$\boldsymbol{w} \coloneqq \arg\min_{\boldsymbol{w} \in \mathbb{R}^d} \frac{1}{N} \sum_{n=1}^{N} (V_{(n)}^{\text{target}} - \hat{V}_{\boldsymbol{w}}(s_{(n)}))^2$$

- 最適行動価値関数 $Q^*$ の推定 (適合 Q 反復)
  1. 各経験の目的変数を算出

$$Q_{(n)}^{\text{target}} \coloneqq \hat{\Upsilon}_*(\hat{Q}_{\boldsymbol{w}}; \{s_{(n)}, a_{(n)}, r_{(n)}, s'_{(n)}\})(s_{(n)}, a_{(n)}) = r_{(n)} + \gamma \max_{a' \in \mathcal{A}} \hat{Q}_{\boldsymbol{w}}(s_{(n+1)}, a')$$

  2. パラメータ $\boldsymbol{w}$ を更新

$$\boldsymbol{w} \coloneqq \arg\min_{\boldsymbol{w} \in \mathbb{R}^d} \frac{1}{N} \sum_{n=1}^{N} (Q_{(n)}^{\text{target}} - \hat{Q}_{\boldsymbol{w}}(s_{(n)}, a_{(n)}))^2$$

以上のような価値関数の関数近似の際には，収束性に気を配る必要がある．例 6.2 において，関数近似器を

$$\hat{V}_w(s) = \begin{cases} w, & \text{if } s = 1 \\ 2w, & \text{if } s = 2 \end{cases}$$

と定義し，正しい $w$ の値を求めることを考える．$w$ が $w = 0$ へと収束すれば良い．ベルマン最適 (期待) 作用素を $\hat{V}_w$ に適用することで，目的変数は

$$V^{\text{target}} \coloneqq B(\hat{V}_w)(s) = \begin{cases} 0 + \gamma \times 2w, & (\text{if } s = 1) \\ 0 + \gamma \times 2w, & (\text{if } s = 2) \end{cases}$$

なので，第 $k$ 回目の $w$ の更新は，

$$\begin{aligned} w_{k+1} &= \arg\min_{w\in\mathbb{R}} \sum_{s\in\mathcal{S}} (V^{\text{target}}(s) - \hat{V}_w(s))^2 \\ &= \arg\min_{w\in\mathbb{R}} \{(V^{\text{target}}(1) - \hat{V}_w(1))^2 + (V^{\text{target}}(2) - \hat{V}_w(2))^2\} \\ &= \arg\min_{w\in\mathbb{R}} \{(2\gamma w_k - w)^2 + (2\gamma w_k - 2w)^2\} \\ &= \arg\min_{w\in\mathbb{R}} \{5(w - \frac{6}{5}\gamma w_k)^2 + \frac{4}{5}\gamma^2 w_k^2\} \end{aligned}$$

となるので，パラメータ $w$ の更新則は

$$w_{k+1} = \frac{6}{5}\gamma w_k$$

となる．ゆえ，$\gamma > \frac{5}{6}$ の場合，$w$ は単調増加で発散してしまう．
このように，関数近似器集合が真の関数を含んでいたとしても，更新則によってはパラメータが発散してしまう場合がある．そのため，関数近似器 $\mathcal{V}$ の表現力を評価する上で,「最悪ケースの関数近似誤差」を用いる場合がある．関数近似誤差とは，$\boldsymbol{w}$ を

$$\boldsymbol{w} \coloneqq \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \sum_{s\in\mathcal{S}} (V^{\text{target}}(s) - \hat{V}_{\boldsymbol{w}}(s))^2$$

と更新する場合，

$$\inf_{v'\in\mathcal{V}} \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} (V^{\text{target}}(s) - v'(s))^2$$

と定義できるので,「最悪ケース」とは $v'$ による誤差が最大になってしまうケースであり，そのときの近似誤差は

$$\epsilon^*(\mathcal{V}; B) \coloneqq \sup_{v\in\mathcal{V}} \inf_{v'\in\mathcal{V}} \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} (Bv(s) - v'(s))^2$$

となる．つまり，最悪ケースの近似誤差とは,「頑張って誤差が小さくなるよう $v'$ を設定した時生じた誤差の最大値」である．指標 $\epsilon^*$ が高いほど，$\mathcal{V}$ の表現力が低い (任意の $V^{\text{target}}$ に対する誤差を小さくする能力がない) ことを意味する．例 6.2 では，近似誤差は

$$\frac{4\gamma^2}{5} w_k^2$$

より，最悪近似誤差は $w_k \to \infty$ で無限大に発散してしまうため，不良な関数近似器を用いていたことがわかる．

### 1.2.2　オンライン学習の場合

行動方策 $\pi$ による価値 $V^\pi$ を求めるため，関数近似器 $\hat{V}_{\boldsymbol{w}}$ を用いて近似 TD 法を考える．関数近似をしない TD 法では特定の状態 $s_t$ に対してしか $\hat{V}$ を更新しなかったが，近似 TD 法では $\boldsymbol{w}_t$ を更新するため，経験した状態 $s_t$ に応じてパラメータを更新すると別の状態に対しても推定価値が変わる事になる．TD 誤差を

$$\delta_t \coloneqq r_t + \gamma \hat{V}_{\boldsymbol{w}_t}(s_{t+1}) - \hat{V}_{\boldsymbol{w}_t}(s_t)$$

とし，$\hat{V}_{\boldsymbol{w}}$ の $\boldsymbol{w}$ に関する偏微分ベクトル (定数ベクトル) を

$$\nabla_{\boldsymbol{w}} \hat{V}_{\boldsymbol{w}_t}(s) \coloneqq \begin{pmatrix} \left.\frac{\partial \hat{V}_{\boldsymbol{w}}(s)}{\partial w_1}\right|_{\boldsymbol{w}=\boldsymbol{w}_t} \\ \vdots \\ \left.\frac{\partial \hat{V}_{\boldsymbol{w}}(s)}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{w}=\boldsymbol{w}_t} \end{pmatrix}$$

と定義する．今，$\hat{V}_{\boldsymbol{w}_t+\Delta\boldsymbol{w}}(s)$ を $\boldsymbol{w}_t$ 周りでテイラー展開すると，

$$\begin{aligned} \hat{V}_{\boldsymbol{w}_t+\Delta\boldsymbol{w}}(s) &= \hat{V}_{\boldsymbol{w}_t}(s) + {}^\top(\nabla_{\boldsymbol{w}} \hat{V}_{\boldsymbol{w}_t}(s))(\boldsymbol{w}_t + \Delta\boldsymbol{w} - \boldsymbol{w}_t) + o(\|\Delta(\boldsymbol{w})\|^2) \\ &= \hat{V}_{\boldsymbol{w}_t}(s) + {}^\top(\nabla_{\boldsymbol{w}} \hat{V}_{\boldsymbol{w}_t}(s))\Delta\boldsymbol{w} + o(\|\Delta(\boldsymbol{w})\|^2) \end{aligned}$$

となる．すなわち，$\boldsymbol{w}_t$ を $\Delta\boldsymbol{w}$ だけ変化させた時の $\hat{V}_{\boldsymbol{w}}$ の変化量は ${}^\top(\nabla_{\boldsymbol{w}} \hat{V}_{\boldsymbol{w}_t}(s))\Delta\boldsymbol{w}$ となる．さて，近似 TD 法において，

$$\boldsymbol{w}_{t+1} \coloneqq \boldsymbol{w}_t + \alpha\delta_t \nabla_{\boldsymbol{w}} \hat{V}_{\boldsymbol{w}}(s_t) \tag{1}$$

と $\boldsymbol{w}$ を更新することにすると，この時の $s_t$ における推定価値の変化量は，

$$\begin{aligned} \hat{V}_{\boldsymbol{w}_{t+1}}(s_t) &= \hat{V}_{\boldsymbol{w}_t}(s_t) + {}^\top(\nabla_{\boldsymbol{w}} \hat{V}_{\boldsymbol{w}_t}(s_t))\alpha\delta_t \nabla_{\boldsymbol{w}} \hat{V}_{\boldsymbol{w}}(s_t) \\ &= \hat{V}_{\boldsymbol{w}_t}(s_t) + \alpha\delta_t \|\nabla_{\boldsymbol{w}} \hat{V}_{\boldsymbol{w}}(s_t)\|^2 \end{aligned}$$

となり，関数近似をしない TD 法における価値関数の更新とほとんど等価になる．((1) 式の右辺第二項を $\|\nabla_{\boldsymbol{w}} \hat{V}_{\boldsymbol{w}_t}(s_t)\|^2$ で割った更新則もある．こうすれば関数近似をしない TD 法と等価．) また，$\tilde{s} \neq s_t$ であるような $\tilde{s}$ における価値関数は，

$$\hat{V}_{\boldsymbol{w}_{t+1}}(\tilde{s}) = \hat{V}(\tilde{s}) + \alpha\delta^\top(\nabla_{\boldsymbol{w}} \hat{V}_{\boldsymbol{w}_t}(s_t))\nabla_{\boldsymbol{w}} \hat{V}_{\boldsymbol{w}_t}(\tilde{s})$$

と，内積 ${}^\top(\nabla_{\boldsymbol{w}} \hat{V}_{\boldsymbol{w}_t}(s_t))\nabla_{\boldsymbol{w}} \hat{V}_{\boldsymbol{w}_t}(\tilde{s})$ に依存して更新される．よって，状態 $(s_t, \tilde{s})$ 間の類似度と内積 ${}^\top(\nabla_{\boldsymbol{w}} \hat{V}_{\boldsymbol{w}_t}(s_t))\nabla_{\boldsymbol{w}} \hat{V}_{\boldsymbol{w}_t}(\tilde{s})$ が正の相関をもつように関数近似器 $\hat{V}_{\boldsymbol{w}}$ を設定できれば，現時点での状態 $s_t$ の推定価値の更新と同時に $s$ とよ

く似た状態 $\tilde{s}$ の推定価値も更新することができる．

TD($\lambda$) 法においては，

$$\hat{V}(s) \coloneqq \hat{V}(s) + \alpha_t \delta_t \sum_{\tau=0}^{t} \mathbb{I}_{s_\tau = s} (\lambda\gamma)^{t-\tau}, \quad \forall s \in \mathcal{S}$$

と価値関数を更新していた．今，近似 TD($\lambda$) 法において

$$\boldsymbol{w}_{t+1} \coloneqq \boldsymbol{w}_t + \alpha \delta_t \sum_{\tau=0}^{t} (\gamma\lambda)^{t-\tau} \nabla_{\boldsymbol{w}} \hat{V}_{\boldsymbol{w}_\tau}(s_\tau)$$

と $\boldsymbol{w}$ を更新することにすると，状態 $s_t$ における推定価値の変化量は

$$\hat{V}_{\boldsymbol{w}_{t+1}}(s_t) = \hat{V}_{\boldsymbol{w}_t}(s_t) + \alpha_t \delta_t \sum_{\tau=0}^{t} (\gamma\lambda)^{t-\tau \top} (\nabla_{\boldsymbol{w}} \hat{V}_{\boldsymbol{w}_t}(s_t)) \nabla_{\boldsymbol{w}} \hat{V}_{\boldsymbol{w}_\tau}(s_\tau)$$

と，関数近似をしない TD($\lambda$) 法における価値関数の更新とほとんど等価になる．

行動価値関数の学習に関数近似器 $\hat{Q}_{\boldsymbol{w}}$ を用いる場合でも，TD 誤差を

$$\delta_t^{(q)} \coloneqq r + \gamma \max_{a' \in \mathcal{A}} \hat{Q}_{\boldsymbol{w}_t}(s_{t+1}, a') - \hat{Q}_{\boldsymbol{w}_t}(s_t, a_t)$$

とおけば

$$\boldsymbol{w}_{t+1} \coloneqq \boldsymbol{w}_t + \alpha \delta_t^{(q)} \nabla_{\boldsymbol{w}} \hat{Q}_{\boldsymbol{w}_t}(s_t, a_t)$$

のようにパラメータを更新できる．

### 1.3 損失関数に基づく近似価値関数学習法

テーブル形式の方法を単純に拡張した方法では，推定値 $\hat{V}_{\boldsymbol{w}}$ が収束しない場合があった．そこで，損失関数を導入し，損失関数の最小化を考える．「ベルマン作用素を繰り返し適用して $V^\pi(s) = B_\pi V^\pi(s)$ が成り立つような $V^\pi$ を近似する，というこれまでのアイデアとは異なり，損失関数を最小化するような $\boldsymbol{w}$ を勾配法や最小二乗法によって解く．

### 1.3.1 ベルマン残差

ベルマン方程式：

$$V^{\pi}(s) = B_{\pi}V^{\pi}(s)$$

より，任意の $s$ に対し $\hat{V}_{\boldsymbol{w}}(s)$ と $B\hat{V}_{\boldsymbol{w}}(s)$ の差が小さくなるように $\boldsymbol{w}$ を設定したい．そこで，ベルマン残差：

$$L_{\mathrm{BR}}(\boldsymbol{w}) \coloneqq \sum_{s\in\mathcal{S}} \mu(s)(\hat{V}_{\boldsymbol{w}}(s) - B\hat{V}_{\boldsymbol{w}}(s))^2$$

の最小化を考える．今，関数近似器が線形かつ $B$ に真のベルマン作用素 $B_{\pi}$ を用いる：

$$\hat{V}_{\boldsymbol{w}}(s) = {}^{\top}\boldsymbol{w}\boldsymbol{\phi}(s), \quad B(\hat{V}_{\boldsymbol{w}})(s) = \mathbb{E}_{A_t|S_t\sim\pi, S_{t+1}|S_t,A_t\sim p_T}[g(S_t,A_t) + \gamma\hat{V}_{\boldsymbol{w}}(S_{t+1})|S_t = s]$$

とする．これを $L_{\mathrm{BR}}$ の式に代入することで，

$$\begin{aligned}
L_{\mathrm{BR}}(\boldsymbol{w}) &= \sum_{s\in\mathcal{S}} \mu(s)(\hat{V}_{\boldsymbol{w}}(s) - \mathbb{E}_{A_t|S_t\sim\pi, S_{t+1}|S_t,A_t\sim p_T}[g(S_t,A_t) + \gamma\hat{V}_{\boldsymbol{w}}(S_{t+1})|S_t = s])^2 \\
&= \sum_{s\in\mathcal{S}} \mu(s)({}^{\top}\boldsymbol{w}\boldsymbol{\phi}(s) - \mathbb{E}_{A_t|S_t\sim\pi, S_{t+1}|S_t,A_t\sim p_T}[g(S_t,A_t) + \gamma\,{}^{\top}\boldsymbol{w}\boldsymbol{\phi}(S_{t+1})|S_t = s])^2 \\
&= \sum_{s\in\mathcal{S}} \mu(s)({}^{\top}\boldsymbol{w}\boldsymbol{\phi}(s) \\
&\qquad\qquad -\mathbb{E}_{A_t|S_t\sim\pi}[g(S_t,A_t)|S_t = s] \\
&\qquad\qquad -\gamma\,{}^{\top}\boldsymbol{w}\mathbb{E}_{A_t|S_t\sim\pi, S_{t+1}|S_t,A_t\sim p_T}[\boldsymbol{\phi}(S_{t+1})|S_t = s])^2 \\
&= \sum_{s\in\mathcal{S}} \mu(s)\{{}^{\top}\boldsymbol{w}(\boldsymbol{\phi}(s) - \gamma\bar{\boldsymbol{\phi}}^{\pi}_{+1}(s)) - \bar{g}^{\pi}(s)\}^2
\end{aligned}$$

ただし，

$$\begin{aligned}
\bar{g}^{\pi}(s) &= \mathbb{E}_{A_t|S_t\sim\pi}[g(S_t,A_t)|S_t = s] = \sum_{a\in\mathcal{A}} \pi(a|s)g(s,a) \\
\bar{\boldsymbol{\phi}}^{\pi}_{+1}(s) &= \mathbb{E}_{A_t|S_t\sim\pi, S_{t+1}|S_t,A_t\sim p_T}[\boldsymbol{\phi}(S_{t+1})|S_t = s] = \sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}} \pi(a|s)p_T(s'|s,a)\boldsymbol{\phi}(s')
\end{aligned}$$

とおいた．従って，$L_{\mathrm{BR}}$ を $\boldsymbol{w}$ について微分して $\mathbf{0}$ とおくことで $L_{\mathrm{BR}}$ の最小値を与えるパラメータの最適値 $\boldsymbol{w}^*_{\mathrm{BR}}$ が求まる．

$$\begin{aligned}
\frac{\partial L_{\mathrm{BR}}(\boldsymbol{w})}{\partial \boldsymbol{w}} &= 2\sum_{s\in\mathcal{S}}\mu(s)\{^\top\boldsymbol{w}(\boldsymbol{\phi}(s)-\gamma\bar{\boldsymbol{\phi}}^\pi_{+1}(s))-\bar{g}^\pi(s)\}(\boldsymbol{\phi}(s)-\gamma\bar{\boldsymbol{\phi}}^\pi_{+1}(s)) \qquad (1)\\
&= 2^\top\boldsymbol{w}\sum_{s\in\mathcal{S}}\mu(s)(\boldsymbol{\phi}(s)-\gamma\bar{\boldsymbol{\phi}}^\pi_{+1}(s))^\top(\boldsymbol{\phi}(s)-\gamma\bar{\boldsymbol{\phi}}^\pi_{+1}(s))\\
&\quad -2\sum_{s\in\mathcal{S}}\mu(s)\bar{g}^\pi(s)(\boldsymbol{\phi}(s)-\gamma\bar{\boldsymbol{\phi}}(s))\\
\therefore \boldsymbol{w}^*_{\mathrm{BR}} &= \{\sum_{s\in\mathcal{S}}\mu(s)(\boldsymbol{\phi}(s)-\gamma\bar{\boldsymbol{\phi}}^\pi_{+1}(s))^\top(\boldsymbol{\phi}(s)-\gamma\bar{\boldsymbol{\phi}}^\pi_{+1}(s))\}^{-1}\\
&\quad \times\sum_{s\in\mathcal{S}}\mu(s)\bar{g}^\pi(s)(\boldsymbol{\phi}(s)-\gamma\bar{\boldsymbol{\phi}}^\pi_{+1}(s)) \qquad (2)
\end{aligned}$$

$\boldsymbol{w}^*_{\mathrm{BR}}$ は状態行動空間が膨大である場合には計算することができない，そこで，履歴データ $h_T=\{s_0,a_0,r_0,s_{t+1},...,s_{T-1},a_{T-1},r_{T-1},s_T\}$ から $\boldsymbol{w}^*_{\mathrm{BR}}$ の推定量 $\hat{\boldsymbol{w}}^*_{\mathrm{BR}}$ を求める．$\mu(s)$ に状態の定常分布 $p^\pi_\infty(s)$ を取ることにすれば，(2) 式は

$$\begin{aligned}
\boldsymbol{w}^*_{\mathrm{BR}} &= \{\mathbb{E}_{S_t\sim p^\pi_\infty}[(\boldsymbol{\phi}(S_t)-\gamma\bar{\boldsymbol{\phi}}^\pi_{+1}(S_t))^\top(\boldsymbol{\phi}(S_t)-\gamma\bar{\boldsymbol{\phi}}^\pi_{+1}(S_t))]\}^{-1}\\
&\quad \times\mathbb{E}_{S_t\sim p^\pi_\infty}[\bar{g}^\pi(S_t)(\boldsymbol{\phi}(S_t)-\gamma\bar{\boldsymbol{\phi}}^\pi_{+1}(S_t))]
\end{aligned}$$

と書けるので，これを標本近似して，

$$\hat{\boldsymbol{w}}^*_{\mathrm{BR}}=A^{-1}\boldsymbol{b}$$

とすることができる．ただし，

$$\begin{aligned}
A &\coloneqq \frac{1}{T}\sum_{t=0}^{T-1}(\boldsymbol{\phi}(s_t)-\boldsymbol{\phi}(s_{t+1}))^\top(\boldsymbol{\phi}(s_t)-\boldsymbol{\phi}(\tilde{s}_{t+1}))\\
\boldsymbol{b} &\coloneqq \frac{1}{T}\sum_{t=0}^{T-1}r_t(\boldsymbol{\phi}(s_t)-\gamma\boldsymbol{\phi}(s_{t+1}))
\end{aligned}$$

とおいた．オンラインの場合は (1) 式より勾配降下法を行えばよく，

$$\boldsymbol{w}\coloneqq\boldsymbol{w}-\alpha_t\{^\top\boldsymbol{w}(\boldsymbol{\phi}(s_t)-\gamma\boldsymbol{\phi}(s_{t+1}))-r_t\}(\boldsymbol{\phi}(s_t)-\gamma\boldsymbol{\phi}(\tilde{s}_{t+1}))$$

となる．しかし，以上の推定式では $s_t$ の次状態として $s_t,\tilde{s}_{t+1}$ を独立にサンプリングしないと不偏推定量たりえない．どういうことかというと，例えば

$$\bar{\boldsymbol{\phi}}^\pi_{+1}(s)=\sum_{a\in\mathcal{A}}\sum_{s\in\mathcal{S}}\pi(a|s)p_T(s'|s,a)\boldsymbol{\phi}(s')$$

を推定する際は，$\boldsymbol{\phi}(s')$ の実現値の平均を取ることで

$$\frac{1}{\sum_{t=0}^{T-1} \mathbb{I}_{s_t=s}} \sum_{t=0}^{T-1} \boldsymbol{\phi}(s_{t+1}) \mathbb{I}_{s_t=s}$$

とすればこれは標本平均なのだから不偏推定量であるが，

$$\bar{\boldsymbol{\phi}}^{\pi}_{+1}(s)^{\top} \bar{\boldsymbol{\phi}}^{\pi}_{+1}(s) = \sum_{a\in\mathcal{A}} \sum_{s\in\mathcal{S}} \pi(a|s) p_T(s'|s,a) \boldsymbol{\phi}(s') \times {}^{\top}\{ \sum_{a\in\mathcal{A}} \sum_{s\in\mathcal{S}} \pi(a|s) p_T(s'|s,a) \boldsymbol{\phi}(s') \}$$

の推定量を

$$\frac{1}{\sum_{t=0}^{T-1} \mathbb{I}_{s_t=s}} \sum_{t=0}^{T-1} \boldsymbol{\phi}(s_{t+1})^{\top} \boldsymbol{\phi}(s_{t+1}) \mathbb{I}_{s_t=s}$$

としてしまうと，これが近似しているのは

$$\sum_{a\in\mathcal{A}} \sum_{s\in\mathcal{S}} \pi(a|s) p_T(s'|s,a) \{ \boldsymbol{\phi}(s')^{\top} \boldsymbol{\phi}(s') \}$$

であって，$\bar{\boldsymbol{\phi}}^{\pi}_{+1}(s)^{\top} \bar{\boldsymbol{\phi}}^{\pi}_{+1}(s)$ に対する不偏推定量ではない．そのため，

$$\frac{1}{\sum_{t=0}^{T-1} \mathbb{I}_{s_t=s}} \sum_{t=0}^{T-1} \boldsymbol{\phi}(s_{t+1})^{\top} \boldsymbol{\phi}(\tilde{s}_{t+1}) \mathbb{I}_{s_t=s}$$

と，$s_{t+1}$ を観測した後で $\tilde{s}_{t+1}$ を前状態 $s_t$ に戻って $s_{t+1}$ とは独立にサンプリングしなければならない．これが二重サンプリング問題である．一般の強化学習問題では，二重サンプリングは実施が難しいため，残差勾配法では $\tilde{s}_{t+1}$ を $s_{t+1}$ と独立にサンプリングすることを諦め，$s_{t+1}$ で置き換えてしまう．

$$\begin{aligned} \boldsymbol{w} &\coloneqq \boldsymbol{w} - \alpha_t \{ {}^{\top}\boldsymbol{w}(\boldsymbol{\phi}(s_t) - \gamma\boldsymbol{\phi}(s_{t+1})) - r_t \}(\boldsymbol{\phi}(s_t) - \gamma\boldsymbol{\phi}(s_{t+1})) \\ &= \boldsymbol{w} + \alpha_t (r_t + \gamma \hat{V}_{\boldsymbol{w}}(s_{t+1}) - \hat{V}_{\boldsymbol{w}}(s_t))(\boldsymbol{\phi}(s_t) - \gamma\boldsymbol{\phi}(s_{t+1})) \end{aligned}$$

これは，次の期待二乗 TD 誤差：

$$\begin{aligned} L_{\mathrm{TD}}(\boldsymbol{w}) &\coloneqq \sum_{s\in\mathcal{S}} \mu(s) \mathbb{E}_{A_t|S_t\sim\pi, S_{t+1}|S_t,A_t\sim p_T} [\{ g(S_t,A_t) + \gamma \hat{V}_{\boldsymbol{w}}(S_{t+1}) - \hat{V}_{\boldsymbol{w}}(S_t) \}^2 | S_t] \\ &= \sum_{s\in\mathcal{S}} \mu(s) [ \sum_{a\in\mathcal{A}} \{ \pi(a|s) g(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_T(s'|s,a) \hat{V}_{\boldsymbol{w}}(s') \} - \hat{V}_{\boldsymbol{w}}(s) ]^2 \\ &= \sum_{s\in\mathcal{S}} \mu(s) [ \sum_{a\in\mathcal{A}} \{ \pi(a|s) g(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_T(s'|s,a) {}^{\top}\boldsymbol{w}\boldsymbol{\phi}(s') \} - {}^{\top}\boldsymbol{w}\boldsymbol{\phi}(s) ]^2 \end{aligned}$$

の重み関数を $\mu = p^{\pi}_{\infty}$ とした時の不偏な確率的勾配法に対応する (微分して $\mathbf{0}$ とおけば明らか．)．残差勾配法は対応する損失関数 $L_{\mathrm{TD}}$ をもつため，安定した学習を期待できるが，実は．$L_{\mathrm{TD}}$ の最小化では $V^{\pi}$ には収束しない．後に詳述するが，$L_{\mathrm{TD}}$ とは，報酬 $r_t = g(s_t, a_t)$ を $\hat{V}(s_t) = r_t + \gamma\hat{V}(s_{t+1})$ を用いて線形モデル ${}^{\top}(\boldsymbol{\phi}(s_t) - \gamma\boldsymbol{\phi}(s_{t+1}))\boldsymbol{w}$ で予測する場合の

$$r_t = {}^{\top}(\boldsymbol{\phi}(s_t) - \gamma\boldsymbol{\phi}(s_{t+1}))\boldsymbol{w} + \sigma_t$$

の誤差項 $\sigma_t$ の期待二乗値に対応する．

### 1.3.2 射影ベルマン残差

$B\hat{V}_{\boldsymbol{w}}$ が必ずしも関数近似器空間 $\mathcal{V}$ に含まれているとは限らないことから，射影ベルマン残差：

$$L_{\mathrm{PBR}}(\boldsymbol{w}) \coloneqq \sum_{s\in\mathcal{S}} \mu(s)(\hat{V}_{\boldsymbol{w}}(s) - \Gamma(B\hat{V}_{\boldsymbol{w}})(s))^2$$

を用いる．ただし，$\Gamma$ は任意の状態関数 $v$ を関数近似器空間 $\mathcal{V}$ へと直交射影する作用素：

$$\Gamma(v) \coloneqq \arg\min_{\hat{V}_{\boldsymbol{w}}\in\mathcal{V}} \sum_{s\in\mathcal{S}} \mu(s)(v(s) - \hat{V}_{\boldsymbol{w}}(s))^2$$

である．$L_{PBR}$ の最小化を考える．今，関数近似器は線形かつ $B$ に真のベルマン作用素 $B_{\pi}$ を用いる：

$$\begin{aligned} \hat{V}_{\boldsymbol{w}}(s) &= {}^{\top}\boldsymbol{w}\boldsymbol{\phi}(s) = w_1\phi_1(s) + ... + w_{|\mathcal{V}|}\phi_{|\mathcal{V}|}(s) \\ B(\hat{V}_{\boldsymbol{w}})(s) &= \mathbb{E}_{A_t|S_t\sim\pi, S_{t+1}|S_t, A_t\sim p_T}[g(S_t, A_t) + \gamma\hat{V}_{\boldsymbol{w}}(S_{t+1})|S_t = s] \end{aligned}$$

まずは，次のような行列表記を用意する．

- 状態遷移確率 $P^{\pi}$ ... $(i, j)$ 要素が状態 $i$ から状態 $j$ へと遷移する確率であるような行列．

$$(P^{\pi})_{i,j} \coloneqq p^{\pi}_{MC}(j|i) = \sum_{a\in\mathcal{A}} \pi(a|i)p_T(j|i, a)$$

- 報酬ベクトル $\boldsymbol{r}^{\pi}$ ... 第 $i$ 要素が状態 $i$ における報酬の期待値であるような行列．

$$(\boldsymbol{r}^{\pi})_i \coloneqq \mathbb{E}_{A_t|S_t\sim\pi}[g(A_t, S_t)|S_t = i] = \sum_{a\in\mathcal{A}} \pi(a|i)g(i, a)$$

- 損失関数の重み行列 $U$ ... $i$ 番目の対角要素が $L_{\mathrm{PBR}}$ の状態 $i$ の重みである行列.

$$(U)_{i,j} \coloneqq \begin{cases} \mu(i), & (i=j) \\ 0, & \text{otherwise} \end{cases}$$

- 価値ベクトル $\boldsymbol{v}^\pi$ と推定価値ベクトル $\hat{\boldsymbol{v}}_{\boldsymbol{w}}$ ... 第 $i$ 要素がそれぞれ状態 $i$ の価値及び推定価値であるようなベクトル．前者は $|\mathcal{S}|$ 次元，後者は $|\mathcal{V}|$ 次元.

$$\begin{aligned} (\boldsymbol{v}^\pi) &\coloneqq V^\pi(i) \\ \hat{\boldsymbol{v}}_{\boldsymbol{w}} &\coloneqq \begin{pmatrix} \hat{V}_{\boldsymbol{w}}(1) \\ \vdots \\ \hat{V}_{\boldsymbol{w}}(|\mathcal{S}|) \end{pmatrix} = \Phi\boldsymbol{w} = \begin{pmatrix} \phi_1(1) & \dots & \phi_{|\mathcal{V}|}(1) \\ \vdots & \ddots & \vdots \\ \phi_1(|\mathcal{S}|) & \dots & \phi_{|\mathcal{V}|}(|\mathcal{S}|) \end{pmatrix} \begin{pmatrix} \boldsymbol{w}_1 \\ \vdots \\ \boldsymbol{w}_{|\mathcal{V}|} \end{pmatrix} = \begin{pmatrix} {}^\top\boldsymbol{\phi}(1) \\ \vdots \\ {}^\top\boldsymbol{\phi}(|\mathcal{S}|) \end{pmatrix} \boldsymbol{w} \end{aligned}$$

今，ある状態関数 $v$ に対し，

$$\boldsymbol{v} = \begin{pmatrix} v(1) \\ \vdots \\ v(|\mathcal{S}|) \end{pmatrix}$$

と置く．$\boldsymbol{v}$ は $|\mathcal{S}|$ 次元のベクトルである．対して，$|\mathcal{V}| < |\mathcal{S}|$ ならば，$\hat{\boldsymbol{v}}_{\boldsymbol{w}}$ は $|\mathcal{V}|$ 次元のベクトルである．$\Gamma(v)$ において，二乗誤差を微分して $\mathbf{0}$ と置く．

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{w}} \sum_{s\in\mathcal{S}} \mu(s)(v(s) - {}^\top\boldsymbol{w}\boldsymbol{\phi}(s))^2 &\propto \sum_{s\in\mathcal{S}} \mu(s)(v(s) - {}^\top\boldsymbol{w}\boldsymbol{\phi}(s)){}^\top\boldsymbol{\phi}(s) \\ &= \sum_{s\in\mathcal{S}} \mu(s)v(s){}^\top\boldsymbol{\phi}(s) - {}^\top\boldsymbol{w}\sum_{s\in\mathcal{S}} \mu(s)\boldsymbol{\phi}(s){}^\top\boldsymbol{\phi}(s) = \mathbf{0} \\ \therefore \mathbf{0} &= {}^\top\boldsymbol{v}{}^\top U\Phi - {}^\top\boldsymbol{w}{}^\top\Phi U\Phi \\ \therefore \hat{\boldsymbol{w}} &= ({}^\top\Phi U\Phi)^{-1}{}^\top\Phi U\boldsymbol{v} \end{aligned}$$

ただし，$\hat{w}$ は $\Gamma$ によるパラメータの推定値である．$v$ は $\Gamma$ によって $\mathcal{V}$ 内の関数近似器 $\hat{V}_{\hat{\boldsymbol{w}}}$ へと射影されることになる．したがって，ある状態関数 $v$ に対し，任意の $s$ について $\Gamma$ を作用させることは，$\boldsymbol{v}$ に対し以下のように $\boldsymbol{\Gamma}$ を作用させて部分空間に射影することと言える．

$$\boldsymbol{\Gamma}\boldsymbol{v} \coloneqq \begin{pmatrix} \Gamma(v)(1) \\ \vdots \\ \Gamma(v)(|\mathcal{S}|) \end{pmatrix} = \begin{pmatrix} \hat{V}_{\hat{\boldsymbol{w}}}(1) \\ \vdots \\ \hat{V}_{\hat{\boldsymbol{w}}}(|\mathcal{S}|) \end{pmatrix} = \Phi\hat{\boldsymbol{w}} = \Phi({}^\top\Phi U\Phi)^{-1}{}^\top\Phi U\boldsymbol{v}$$

ここで，行列

$$\mathbf{\Gamma} = \Phi({}^\top\Phi U\Phi)^{-1}{}^\top\Phi U$$

を射影行列と呼ぶ．射影行列 $\mathbf{\Gamma}$ を用いることで，射影ベルマン残差 $L_{\rm PBR}$ は，

$$\begin{aligned}
L_{\rm PBR} &= {}^\top\{\Phi\boldsymbol{w} - \mathbf{\Gamma}(\boldsymbol{r}^\pi + \gamma P^\pi\Phi\boldsymbol{w})\}U\{\Phi\boldsymbol{w} - \mathbf{\Gamma}(\boldsymbol{r}^\pi + \gamma P^\pi\Phi\boldsymbol{w})\} \\
&= {}^\top(\Phi\boldsymbol{w} - \boldsymbol{r}^\pi - \gamma P^\pi\Phi\boldsymbol{w})^\top\mathbf{\Gamma}U\mathbf{\Gamma}(\Phi\boldsymbol{w} - \boldsymbol{r}^\pi - \gamma P^\pi\Phi\boldsymbol{w}) \\
&\quad (\because \mathbf{\Gamma}\Phi\boldsymbol{w} = \Phi\boldsymbol{w}) \\
&= {}^\top\{\boldsymbol{r}^\pi - (\Phi - \gamma P^\pi\Phi)\boldsymbol{w}\}^\top\mathbf{\Gamma}U\mathbf{\Gamma}\{\boldsymbol{r}^\pi - (\Phi - \gamma P^\pi\Phi)\boldsymbol{w}\} \\
&= {}^\top(\boldsymbol{r}^\pi - L^\pi\boldsymbol{w})U\Phi({}^\top\Phi U\Phi)^{-1}{}^\top\Phi U(\boldsymbol{r}^\pi - L^\pi\boldsymbol{w}) \\
&\quad (\because \mathbf{\Gamma} = \Phi({}^\top\Phi U\Phi)^{-1}{}^\top\Phi U)
\end{aligned}$$

ただし，$L^\pi = \Phi - \gamma P^\pi\Phi$ と置いた．よって $L_{\rm PBR}(\boldsymbol{w})$ の一階微分を $\mathbf{0}$ と置くことで，

$$\frac{\partial}{\partial\boldsymbol{w}}L_{\rm PBR}(\boldsymbol{w}) = -2^\top L^\pi U\Phi({}^\top\Phi U\Phi)^{-1}{}^\top\Phi U(\boldsymbol{r}^\pi - L^\pi\boldsymbol{w}) = \mathbf{0}$$

より，$L_{\rm PBR}(\boldsymbol{w})$ を最小化する最適パラメータ $\boldsymbol{w}^*_{\rm PBR}$ は，

$$\begin{aligned}
\boldsymbol{w}^*_{\rm PBR} &= (L^\pi)^{-1}\boldsymbol{r}^\pi \\
&= (L^\pi)^{-1}U^{-1}{}^\top\Phi^{-1}{}^\top\Phi^{-1}U\boldsymbol{r}^\pi \\
&= ({}^\top\Phi UL^\pi)^{-1}{}^\top\Phi U\boldsymbol{r}^\pi
\end{aligned}$$

ここで，

$$\begin{aligned}
{}^\top\Phi UL^\pi &= {}^\top\Phi U(\Phi - \gamma P^\pi\Phi) \\
&= {}^\top\Phi U\Phi - \gamma^\top\Phi UP^\pi\Phi \\
&= \sum_{s\in\mathcal{S}}\mu(s)\boldsymbol{\phi}(s)^\top\boldsymbol{\phi}(s) - \gamma\sum_{s\in\mathcal{S}}\mu(s)^\top\boldsymbol{\phi}(s)\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\pi(a|s)p_T(s'|s,a)\phi(s') \\
&= \sum_{s\in\mathcal{S}}\mu(s)\boldsymbol{\phi}(s)^\top\boldsymbol{\phi}(s) - \gamma\sum_{s\in\mathcal{S}}\mu(s)\phi(s)^\top\bar{\phi}^\pi_{+1}(s) \\
&= \sum_{s\in\mathcal{S}}\mu(s)\boldsymbol{\phi}(s)^\top(\boldsymbol{\phi}(s) - \gamma\bar{\boldsymbol{\phi}}^\pi_{+1}(s)) \\
{}^\top\Phi U\boldsymbol{r}^\pi &= \sum_{s\in\mathcal{S}}\mu(s)\bar{g}^\pi(s)\boldsymbol{\phi}(s)
\end{aligned}$$

なので，

$$\boldsymbol{w}^*_{\rm PBR} = \{\sum_{s\in\mathcal{S}}\mu(s)\boldsymbol{\phi}(s)^\top(\boldsymbol{\phi}(s) - \gamma\bar{\boldsymbol{\phi}}^\pi_{+1}(s))\}^{-1}\sum_{s\in\mathcal{S}}\mu(s)\bar{g}^\pi(s)\boldsymbol{\phi}(s)$$

$\mu(s) = p^{\pi}_{\infty}$ とすれば，上式は履歴データ $\{s_0, r_0, ..., s_{T-1}, r_{T-1}, s_T\}$ から推定することができ，

$$\hat{\boldsymbol{w}}^{*}_{\mathrm{PBR}} = \{\frac{1}{T}\sum_{t=0}^{T-1}\boldsymbol{\phi}(s_t)^{\top}(\boldsymbol{\phi}(s_t) - \gamma\boldsymbol{\phi}(s_{t+1}))\}^{-1}\{\frac{1}{T}\sum_{t=0}^{T-1} r_t\boldsymbol{\phi}(s_t)\}$$

である．

## 2　方策の関数近似

方策パラメータ $\boldsymbol{\theta} \in \mathbb{R}^d$ で確率的方策 $\pi_{\boldsymbol{\theta}} : \mathcal{A} \times \mathcal{S} \to [0, 1]$ で直接規定して，$\boldsymbol{\theta}$ を学習させることを考える．$\pi$ は例えば

$$\pi_{\mathrm{normal}}(a|s;\boldsymbol{\theta}) \coloneqq \frac{1}{\sqrt{2\pi(\sigma(s;\boldsymbol{\theta}))^2}}\exp\{-\frac{(a-\mu(s;\boldsymbol{\theta}))^2}{2(\sigma(s;\boldsymbol{\theta}))^2}\}$$

のような正規分布に基づくものがある．目的関数には，時間ステップ $t = 0$ からのリターン $C_0$ の期待値

$$\begin{aligned} f_0(\boldsymbol{\theta}) &\coloneqq \mathbb{E}_{S_0\sim p_{S_0}, A_t|S_t\sim\pi_{\boldsymbol{\theta}}, S_{t+1}|S_t,A_t\sim p_T}[C_0] \\ &= \mathbb{E}_{S_0\sim p_{S_0}}[V^{\pi_{\boldsymbol{\theta}}}(S_0)] \\ &= \sum_{s\in\mathcal{S}} p_{S_0}(s)V^{\pi_{\boldsymbol{\theta}}}(s) \end{aligned}$$

や，マルコフ連鎖がエルゴート性を満たす場合のみ全時間点での期待リターンの平均

$$\begin{aligned} f_\infty(\boldsymbol{\theta}) &\coloneqq \lim_{T\to\infty}\mathbb{E}_{S_0\sim p_{S_0}, A_t|S_t\sim\pi_{\boldsymbol{\theta}}, S_{t+1}|S_t,A_t\sim p_T}[\frac{1}{T}\sum_{t=0}^{T-1} C_t] \\ &\propto \lim_{T\to\infty}\mathbb{E}_{S_0\sim p_{S_0}, A_t|S_t\sim\pi_{\boldsymbol{\theta}}, S_{t+1}|S_t,A_t\sim p_T}[\frac{1}{T}\sum_{t=0}^{T-1} g(S_t, A_t)] \\ &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p^{\pi_{\boldsymbol{\theta}}}_{\infty}(s)\pi_{\boldsymbol{\theta}}(a|s)g(s,a) \end{aligned}$$

などを用いる．方策勾配とは，目的関数 $f$ の $\boldsymbol{\theta}$ の一階微分ベクトル

$$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial f(\boldsymbol{\theta})}{\partial\theta_1} \\ \vdots \\ \frac{\partial f(\boldsymbol{\theta})}{\partial\theta_d} \end{pmatrix}$$

のことであり，これを求めることができれば

$$\boldsymbol{\theta} \coloneqq \boldsymbol{\theta} + \alpha_t \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$$

と，目的関数が大きくなる方向にパラメータを更新するアプローチを取ることができる．しかし，$f_0(\boldsymbol{\theta})$ および $f_\infty(\boldsymbol{\theta})$ の微分はこのままでは求めることができない．($V^{\pi_{\boldsymbol{\theta}}}$ はそのままでは微分できない)
まずは，$\nabla_{\boldsymbol{\theta}} V^{\pi_{\boldsymbol{\theta}}}$, ひいては $\nabla_{\boldsymbol{\theta}} f_0$ を求めるために，いくつかの概念を導入する．

エピソード ... 確率 $\gamma$ で行動選択し，$1-\gamma$ で終了する時，確率過程が終了するまでの系列 $\{s_0, a_0, r_0, \ldots, s_{t-1}, a_{t-1}, r_{t-1}, s_t\}$

期待割引累積訪問数 ... 1 エピソードで状態 $s$ に訪れる回数の期待値

$$\begin{aligned} d^\pi(s) &= \mathbb{E}_{S_0 \sim p_{S_0}, A_t | S_t \sim \pi, S_{t+1} | A_t, S_t \sim p_T}[\sum_{t=0}^{\infty} \gamma^t \mathbb{I}_{S_t = s}] \\ &= \sum_{t=0}^{\infty} \gamma^t Pr[S_t = s | M(\pi)], \quad s \in \mathcal{S} \end{aligned}$$

エルゴート性を満たさないマルコフ連鎖に対しエピソードを定義することで，エピソードごとにパラメータ更新をすることを考えることができる．今，エピソード $\gamma$ を実行するときの期待値演算子を $\mathbb{E}^\gamma$ と置くことにすると，

$$\begin{aligned} d^\pi(s) &= \mathbb{E}^\gamma_{S_t \sim Pr}[\sum_{t=0}^{\infty} \mathbb{I}_{S_t = s}] \\ V^\pi(s) &= \mathbb{E}_{A_t | S_t \sim \pi, S_{t+1} | S_t, A_t \sim p_T}[\sum_{t=0}^{\infty} \gamma^t g(S_t, A_t) | S_0 = s] \\ &= \mathbb{E}^\gamma_{S_t, A_t | S_0 \sim Pr}[\sum_{t=0}^{\infty} g(S_t, A_t) | S_0 = s] \\ Q^\pi(s, a) &= \mathbb{E}_{S_{t+1} | S_t, A_t \sim p_T, A_{t+1} | S_{t+1} \sim \pi}[\sum_{t=0}^{\infty} \gamma g(S_t, A_t) | S_0 = s, A_0 = a] \\ &= \mathbb{E}^\gamma_{S_t, A_t | S_0, A_0 \sim Pr}[\sum_{t=0}^{\infty} g(S_t, A_t) | S_0 = s, A_0 = a] \end{aligned}$$

(割引率 $\gamma$ と行動選択確率 $\gamma$ は同値．最初はエピソードで切る場合と無限時間の場合で価値関数が同値になってしまうのが変な気がしたが，よく見ると上式はエピソードで切ることは考えていなくて,「割引率を $\gamma$ とする価値関数は，エピソード $\gamma$ を実行するときの期待値演算子で表現出来る」ことが言い

たい．)
以上より，目的関数 $f_0(\boldsymbol{\theta})$ は，

$$\begin{aligned} f_0(\boldsymbol{\theta}) &= \sum_{s\in\mathcal{S}} p_{S_0}(s)V^{\pi_{\boldsymbol{\theta}}}(s) \\ &= \sum_{s\in\mathcal{S}} p_{S_0}(s)\mathbb{E}^{\gamma}_{S_t,A_t|S_0\sim Pr}[\sum_{t=0}^{\infty} g(S_t,A_t)|S_0=s] \\ &= \mathbb{E}^{\gamma}_{S_t,A_t\sim Pr}[\sum_{t=0}^{\infty} g(S_t,A_t)] \\ &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t=0}^{\infty}\gamma^t Pr[S_t=s]\pi_{\boldsymbol{\theta}}(a|s)g(s,a) \\ &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^{\pi_{\boldsymbol{\theta}}}(s)\pi_{\boldsymbol{\theta}}(a|s)g(s,a) \end{aligned}$$

次に，目的関数 $f_\infty$ を価値関数で表現するため，割引現在価値を用いない価値関数として，差分価値関数を導入する．

$$\begin{aligned} Q^{\pi_{\boldsymbol{\theta}}}_{\infty}(s,a) &= \sum_{t=0}^{\infty}\mathbb{E}_{S_{t+1}|A_t,S_t\sim p_T,A_{t+1}|S_{t+1}\sim\pi_{\boldsymbol{\theta}}}[R_t - f_\infty(\boldsymbol{\theta})|S_0=s,A_0=a] \\ V^{\pi_{\boldsymbol{\theta}}}_{\infty}(s) &= \sum_{a\in\mathcal{A}}\pi_{\boldsymbol{\theta}}(a|s)Q^{\pi_{\boldsymbol{\theta}}}_{\infty}(s,a) \end{aligned}$$

割引現在価値 $\gamma$ を使わないまま無限和を取ると発散してしまうため，差分価値関数では各時点で $f_\infty$ を引いている，$t\to\infty$ で

$$\mathbb{E}_{S_t|A_{t-1},S_{t-1}\sim p_T,A_t|S_t\sim\pi_{\boldsymbol{\theta}}}[g(S_t,A_t)|A_{t-1},S_{t-1}] = \mathbb{E}_{S_t\sim p^{\pi}_{\infty},A_t\sim\pi_{\boldsymbol{\theta}}}[g(S_t,A_t)] = f_\infty(\boldsymbol{\theta})$$

となるため，差分価値関数は有限である．差分価値関数には，次のような再起式が成り立つ．

$$\begin{aligned} Q^{\pi_{\boldsymbol{\theta}}}_{\infty}(s,a) &= g(s,a) - f_\infty(\boldsymbol{\theta}) + \sum_{s'\in\mathcal{S}} p_T(s'|s,a)\sum_{a'\in\mathcal{A}}\pi_{\boldsymbol{\theta}}(a'|s')Q^{\pi_{\boldsymbol{\theta}}}_{\infty}(s',a') \\ V^{\pi_{\boldsymbol{\theta}}}_{\infty}(s) &= \sum_{a\in\mathcal{A}}\{g(s,a) - f_\infty(\boldsymbol{\theta}) + \sum_{s'\in\mathcal{S}} p_T(s'|s,a)V^{\pi_{\boldsymbol{\theta}}}_{\infty}(s)\} \end{aligned}$$

目的関数 $f$ の $\boldsymbol{\theta}$ による微分のことを方策勾配 $\nabla_{\boldsymbol{\theta}} f$ という．方策勾配は，次の性質を持つ．

方策勾配定理

a. 平均報酬 $f_\infty$ の方策勾配は，任意の状態関数 $b : \mathcal{S} \to \mathbb{R}$ を用いて

$$\nabla_{\boldsymbol{\theta}} f_\infty(\boldsymbol{\theta}) = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} p_\infty^{\pi_{\boldsymbol{\theta}}}(s)\pi_{\boldsymbol{\theta}}(a|s)\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s)(Q_\infty^{\pi_{\boldsymbol{\theta}}}(s,a) - b(s))$$

b. 価値関数の割引現在価値 $\gamma$ による重みつき和 $f_0$ の方策勾配は，

$$\nabla_{\boldsymbol{\theta}} f_0(\boldsymbol{\theta}) = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} d^{\pi_{\boldsymbol{\theta}}}(s)\pi_{\boldsymbol{\theta}}(a|s)\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s)(Q^{\pi_{\boldsymbol{\theta}}}(s,a) - b(s))$$

(証明 a.)
差分価値関数 $Q_\infty^{\pi_{\boldsymbol{\theta}}}$ の再帰式：

$$Q_\infty^{\pi_{\boldsymbol{\theta}}}(s,a) = g(s,a) - f_\infty(\boldsymbol{\theta}) + \sum_{s'\in\mathcal{S}} p_T(s'|s,a) \sum_{a'\in\mathcal{A}} \pi_{\boldsymbol{\theta}}(a'|s')Q_\infty^{\pi_{\boldsymbol{\theta}}}(s',a')$$

の両辺を $\boldsymbol{\theta}$ で微分すると，

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} Q_\infty^{\pi_{\boldsymbol{\theta}}}(s,a) &= -\nabla_{\boldsymbol{\theta}} f_\infty(\boldsymbol{\theta}) \\ &\quad + \sum_{s'\in\mathcal{S}} \sum_{a'\in\mathcal{A}} p_T(s'|s,a)\{\nabla_{\boldsymbol{\theta}}\pi_{\boldsymbol{\theta}}(a'|s')Q_\infty^{\pi_{\boldsymbol{\theta}}}(s',a') + \pi_{\boldsymbol{\theta}}(a'|s')\nabla_{\boldsymbol{\theta}} Q_\infty^{\pi_{\boldsymbol{\theta}}}(s',a')\}\end{aligned}$$

ここで，

$$\nabla_{\boldsymbol{\theta}}\pi_{\boldsymbol{\theta}}(a'|s') = \pi_{\boldsymbol{\theta}}(a'|s')\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a'|s')$$

を用いて

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} f_\infty(\boldsymbol{\theta}) &= -\nabla_{\boldsymbol{\theta}} Q_\infty^{\pi_{\boldsymbol{\theta}}}(s,a) \\ &\quad + \sum_{s'\in\mathcal{S}} \sum_{a'\in\mathcal{A}} p_T(s'|s,a)\pi_{\boldsymbol{\theta}}(a'|s')\{\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a'|s')Q_\infty^{\pi_{\boldsymbol{\theta}}}(s',a') + \nabla_{\boldsymbol{\theta}} Q_\infty^{\pi_{\boldsymbol{\theta}}}(s',a')\}\end{aligned}$$

また，$f_\infty(\boldsymbol{\theta})$ は $s, a$ に非依存なので，

$$f_\infty(\boldsymbol{\theta}) = \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} p_\infty^{\pi_{\boldsymbol{\theta}}}(s)\pi_{\boldsymbol{\theta}}(a|s)\nabla_{\boldsymbol{\theta}} f_\infty(\boldsymbol{\theta})$$

が成り立つのだから，

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} f_\infty(\boldsymbol{\theta}) &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_\infty^{\pi_{\boldsymbol{\theta}}}(s)\pi_{\boldsymbol{\theta}}(a|s)[\ -\nabla_{\boldsymbol{\theta}} Q_\infty^{\pi_{\boldsymbol{\theta}}}(s,a) \\
&\quad + \sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} p_T(s'|s,a)\pi_{\boldsymbol{\theta}}(a'|s')\{\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a'|s')Q_\infty^{\pi_{\boldsymbol{\theta}}}(s',a') + \nabla_{\boldsymbol{\theta}} Q_\infty^{\pi_{\boldsymbol{\theta}}}(s',a')\}\ ] \\
&= -\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_\infty^{\pi_{\boldsymbol{\theta}}}(s)\pi_{\boldsymbol{\theta}}(a|s)\nabla_{\boldsymbol{\theta}} Q_\infty^{\pi_{\boldsymbol{\theta}}}(s,a) \\
&\quad + \sum_{s'\in\mathcal{S}}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_\infty^{\pi_{\boldsymbol{\theta}}}(s)\pi_{\boldsymbol{\theta}}(a|s)p_T(s'|s,a) \\
&\quad \times \sum_{a'\in\mathcal{A}} \pi_{\boldsymbol{\theta}}(a'|s')\{\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a'|s')Q_\infty^{\pi_{\boldsymbol{\theta}}}(s',a') + \nabla_{\boldsymbol{\theta}} Q_\infty^{\pi_{\boldsymbol{\theta}}}(s',a')\} \\
&= -\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_\infty^{\pi_{\boldsymbol{\theta}}}(s)\pi_{\boldsymbol{\theta}}(a|s)\nabla_{\boldsymbol{\theta}} Q_\infty^{\pi_{\boldsymbol{\theta}}}(s,a) \\
&\quad + \sum_{s'\in\mathcal{S}}\sum_{a'\in\mathcal{A}} p_\infty^{\pi_{\boldsymbol{\theta}}}(s')\pi_{\boldsymbol{\theta}}(a'|s')\{\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a'|s')Q_\infty^{\pi_{\boldsymbol{\theta}}}(s',a') + \nabla_{\boldsymbol{\theta}} Q_\infty^{\pi_{\boldsymbol{\theta}}}(s',a')\} \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_\infty^{\pi_{\boldsymbol{\theta}}}(s)\pi_{\boldsymbol{\theta}}(a|s)\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a|s)Q_\infty^{\pi_{\boldsymbol{\theta}}}(s,a)
\end{aligned}$$

ここで，任意の状態関数 $b(s)$ に対し，

$$\begin{aligned}
\sum_{a\in\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s)\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a|s)b(s) &= \sum_{a\in\mathcal{A}} \nabla_{\boldsymbol{\theta}}\pi_{\boldsymbol{\theta}}(a|s)b(s) \\
&= b(s)\nabla_{\boldsymbol{\theta}}\{\sum_{a\in\mathcal{A}}\pi_{\boldsymbol{\theta}}(a|s)\} = 0
\end{aligned}$$

なのだから，

$$\nabla_{\boldsymbol{\theta}} f_\infty(\boldsymbol{\theta}) = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p_\infty^{\pi_{\boldsymbol{\theta}}}(s)\pi_{\boldsymbol{\theta}}(a|s)\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}\pi(a|s)(Q_\infty^{\pi_{\boldsymbol{\theta}}}(s,a) - b(s))$$

が成り立つ．

(証明 b.)

$$\nabla_{\boldsymbol{\theta}} f_0(\boldsymbol{\theta}) = \sum_{s\in\mathcal{S}} p_{S_0}(s)\nabla_{\boldsymbol{\theta}} V^{\pi_{\boldsymbol{\theta}}}(s)$$

より，まずは $\nabla_{\boldsymbol{\theta}} V^{\pi_{\boldsymbol{\theta}}}(s)$ を求める．

$$V^{\pi_{\boldsymbol{\theta}}}(s) = \sum_{a\in\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s)Q^{\pi_{\boldsymbol{\theta}}}(s,a)$$

の両辺を $\boldsymbol{\theta}$ で微分することで,

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} V^{\pi_{\boldsymbol{\theta}}}(s) &= \sum_{a\in\mathcal{A}} \{\nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a) + \pi_{\boldsymbol{\theta}}(a|s) \nabla_{\boldsymbol{\theta}} Q^{\pi_{\boldsymbol{\theta}}}(s,a)\} \\ &= \sum_{a\in\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s) \{\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a) + \nabla_{\boldsymbol{\theta}} Q^{\pi_{\boldsymbol{\theta}}}(s,a)\}\end{aligned}$$

ここで,

$$Q^{\pi_{\boldsymbol{\theta}}}(s,a) = g(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_T(s'|s,a) V^{\pi_{\boldsymbol{\theta}}}(s')$$

より

$$\nabla_{\boldsymbol{\theta}} Q^{\pi_{\boldsymbol{\theta}}}(s,a) = \gamma \sum_{s'\in\mathcal{S}} p_T(s'|s,a) \nabla_{\boldsymbol{\theta}} V^{\pi_{\boldsymbol{\theta}}}(s')$$

であるから,

$$\nabla_{\boldsymbol{\theta}} V^{\pi_{\boldsymbol{\theta}}}(s) = \sum_{a\in\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s) \{\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_T(s'|s,a) \nabla_{\boldsymbol{\theta}} V^{\pi_{\boldsymbol{\theta}}}(s')\}$$

この式より，ベクトル $\nabla_{\boldsymbol{\theta}} V^{\pi_{\boldsymbol{\theta}}}(s)$ の第 $i$ 要素は,

$$\begin{aligned}\frac{\partial}{\partial\theta_i} V^{\pi_{\boldsymbol{\theta}}}(s) &= \sum_{a\in\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s) \{\frac{\partial}{\partial\theta_i} \log \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_T(s'|s,a) \frac{\partial}{\partial\theta_i} V^{\pi_{\boldsymbol{\theta}}}(s')\} \\ &= \sum_{a\in\mathcal{A}} \pi(a|s) \{g_i(s,a) + \gamma \sum_{s'\in\mathcal{S}} p_T(s'|s,a) \frac{\partial}{\partial\theta_i} V^{\pi_{\boldsymbol{\theta}}}(s')\}\end{aligned}$$

ただし,

$$g_i(s,a) = \frac{\partial}{\partial\theta_i} \log \pi_{\boldsymbol{\theta}}(a|s) Q^{\pi_{\boldsymbol{\theta}}}(s,a)$$

より，$\nabla_{\boldsymbol{\theta}} V^{\pi_{\boldsymbol{\theta}}}(s)$ とは，第 $i$ 要素が報酬関数を $g_i(s,a)$ に持つ価値関数であることがわかる．ゆえに，第 $i$ 要素が $g_i(s,a)$ であるようなベクトル $\boldsymbol{g}(s,a)$ を用いて,

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} V^{\pi_{\boldsymbol{\theta}}}(s) &= \mathbb{E}_{A_t|S_t\sim\pi, S_{t+1}|S_t,A_t\sim p_T}[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{g}(S_t,A_t)|S_0 = s] \\ &= \mathbb{E}^{\gamma}_{S_t,A_t|S_0\sim Pr}[\sum_{t=0}^{\infty} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A_t|S_t) Q^{\pi_{\boldsymbol{\theta}}}(S_t,A_t)|S_0 = s]\end{aligned}$$

以上より，

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} f_0(\boldsymbol{\theta}) &= \sum_{s\in\mathcal{S}} p_{S_0}(s)\nabla_{\boldsymbol{\theta}} V^{\pi_{\boldsymbol{\theta}}}(s) \\
&= \sum_{s\in\mathcal{S}} p_{S_0}(s)\mathbb{E}^{\gamma}_{S_t,A_t|S_0\sim Pr}[\sum_{t=0}^{\infty}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(A_t|S_t)Q^{\pi_{\boldsymbol{\theta}}}(S_t,A_t)|S_0=s] \\
&= \mathbb{E}^{\gamma}_{S_t,A_t\sim Pr}[\sum_{t=0}^{\infty}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(A_t|S_t)Q^{\pi_{\boldsymbol{\theta}}}(S_t,A_t)] \\
&= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^{\pi_{\boldsymbol{\theta}}}(s)\pi_{\boldsymbol{\theta}}(a|s)\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a|s)Q^{\pi_{\boldsymbol{\theta}}}(s,a)
\end{aligned}$$

方策勾配定理は，空間平均を時間平均に置き換えることで次のように書き換えることができる．

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} f_\infty(\boldsymbol{\theta}) &= \lim_{T\to\infty}\mathbb{E}^{\pi}_{S_0\sim p_{S_0},A_t|S_t\sim\pi,S_{t+1}|A_t,S_t\sim p_T}[\frac{1}{T}\sum_{t=0}^{T}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(A_t|S_t)(Q^{\pi_{\boldsymbol{\theta}}}_{\infty}(S_t,A_t)-b(S_t))] \\
\nabla_{\boldsymbol{\theta}} f_0(\boldsymbol{\theta}) &= \mathbb{E}^{\gamma}_{S_t,A_t\sim Pr}[\sum_{t=0}^{\infty}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(A_t|S_t)(Q^{\pi_{\boldsymbol{\theta}}}(S_t,A_t)-b(S_t))]
\end{aligned}$$

これを用いることで，各時点で $\{s_t, a_t, r_t, s_{t+1}\}$ が得られた時，

$$\nabla_{\boldsymbol{\theta}}\log_{\boldsymbol{\theta}}\pi_{\boldsymbol{\theta}}(a_t|s_t)(\hat{Q}_t - b(s_t))$$

を計算することで勾配を推定できる．

## 2.1 モンテカルロ方策勾配法 (REINFORCE 法)

エピソードの履歴データ $\{s_0, a_0, r_0, ..., s_{T-1}, a_{T-1}, r_{T-1}, s_T\}$ を用いて，$Q^\pi$ の推定量として

$$c_t \coloneqq \sum_{k=t}^{T-1} r_k, \qquad \forall t = \{0, ..., T-1\}$$

を計算し，方策勾配を次のように更新する．

$$\boldsymbol{\theta} \coloneqq \boldsymbol{\theta} + \alpha_n \frac{1}{T}\sum_{t=0}^{T-1}(c_t - b(s_t))\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a_t|s_t)$$

## 2.2 アクタークリティック方策勾配法

行動価値関数 $Q_\infty^\pi$ を $\hat{Q}_{\boldsymbol{w}}$ で近似することを考える．これにより，価値関数と方策の両方を関数近似するアクタークリティックが得られる．

- クリティック . . . 近似 TD($\lambda$) 法によって方策の価値 $Q_\infty^\pi$ を更新する．
  - 推定平均報酬 $\hat{f}$ の更新

$$\hat{f}_t \coloneqq \hat{f}_{t-1} + \alpha_t^{\mathrm{critic}}$$

  - TD 誤差 $\delta$ の更新

$$\delta_t \coloneqq r_t - \hat{f}_t + \hat{Q}_{\boldsymbol{w}_t}(s_{t+1}, a_{t+1}) - \hat{Q}_{\boldsymbol{w}_t}(s_t, a_t)$$

  - 関数近似器パラメータ $\boldsymbol{w}$ の更新

$$\boldsymbol{w}_{t+1} \coloneqq \boldsymbol{w}_t + \alpha_t^{\mathrm{critic}} \delta_t \sum_{\tau=0}^{t} \lambda^{t-\tau} \nabla_{\boldsymbol{w}} \hat{Q}_{\boldsymbol{w}_t}(s_t, a_t)$$

- アクター . . . 方策勾配法によって方策 $\pi_{\boldsymbol{\theta}}$ を更新する．
  - 方策パラメータ $\boldsymbol{\theta}$ の更新

$$\boldsymbol{\theta}_{t+1} \coloneqq \boldsymbol{\theta}_t + \alpha_t^{\mathrm{actor}} \hat{Q}_{\boldsymbol{w}_t}(s_t, a_t) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_t | s_t)$$

アクタークリティック方策勾配法において用いる関数近似器 $\hat{Q}_{\boldsymbol{w}}$ について考える．$\hat{Q}_{\boldsymbol{w}}$ のパラメータ $\boldsymbol{w}$ を，ベクトル $\boldsymbol{w}_1, \boldsymbol{w}_2$ を用いて

$$\boldsymbol{w} = \begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{pmatrix}$$

と表し，$\hat{Q}_{\boldsymbol{w}}$ を

$$\hat{Q}_{\boldsymbol{w}}(s, a) \coloneqq {}^\top \boldsymbol{w}_1 \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s) + b_{\boldsymbol{w}_2}(s), \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}$$

と定義する．ただし，$b_{\boldsymbol{w}_2} : \mathcal{S} \to \mathbb{R}$ は $\boldsymbol{w}_2$ で規定される任意の状態関数．

方策勾配の適合関数

- $M(\pi_{\boldsymbol{\theta}})$ がエルゴート性を満たす場合,

$$\frac{\partial}{\partial \boldsymbol{w}_1}\mathbb{E}_{S_0\sim p_{S_0},A_t|S_t\sim\pi,S_{t+1}|A_t,S_t\sim p_T} [\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}(Q^{\pi}_{\infty}(S_t,A_t)-\hat{Q}_{\boldsymbol{w}}(S_t,A_t))^2]\Big|_{\boldsymbol{w}=\boldsymbol{w}^*_{\infty}} = \mathbf{0}$$

を満たすパラメータ $\boldsymbol{w}^*_{\infty}$ を持つ関数近似器 $\hat{Q}_{\boldsymbol{w}^*_{\infty}}$ は次のように方策勾配に適合している.

$$\nabla_{\boldsymbol{\theta}} f_{\infty}(\boldsymbol{\theta}) = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p^{\pi_{\boldsymbol{\theta}}}_{\infty}(s)\pi_{\boldsymbol{\theta}}(a|s)\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a|s)\hat{Q}_{\boldsymbol{w}^*_{\infty}}(s,a)$$

- $\hat{Q}_{\boldsymbol{w}_0}$ が

$$\frac{\partial}{\partial \boldsymbol{w}_1}\mathbb{E}^{\gamma}_{S_t,A_t\sim Pr} = [\sum_{t=0}^{\infty}(Q^{\pi}(S_t,A_t)-\hat{Q}_{\boldsymbol{w}}(S_t,A_t))^2]\Big|_{\boldsymbol{w}=\boldsymbol{w}^*_0} = \mathbf{0}$$

を満たすパラメータ $\boldsymbol{w}^*_{\infty}$ を持つとする．この時，関数近似器 $\hat{Q}_{\boldsymbol{w}_0}$ は次のように方策勾配に適合している.

$$\nabla_{\boldsymbol{\theta}} f_0(\boldsymbol{\theta}) = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^{\pi_{\boldsymbol{\theta}}}(s)\pi_{\boldsymbol{\theta}}(a|s)\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a|s)\hat{Q}_{\boldsymbol{w}^*_0}(s,a)$$

(証明)
$M(\pi_{\boldsymbol{\theta}})$ はエルゴート性を満たすので，時間平均は空間平均と一致するので

$$\begin{aligned}
&\mathbb{E}_{S_0\sim p_{S_0},A_t|S_t\sim\pi,S_{t+1}|S_t,A_t\sim p_T}[\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}(Q^{\pi}_{\infty}(S_t,A_t)-\hat{Q}_{\boldsymbol{w}}(S_t,A_t))^2] \\
= &\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p^{\pi_{\boldsymbol{\theta}}}_{\infty}(s)\pi_{\boldsymbol{\theta}}(a|s)(Q^{\pi}_{\infty}(s,a)-\hat{Q}_{\boldsymbol{w}}(s,a))^2
\end{aligned}$$

が成り立つ．今，右辺を $\boldsymbol{w}_1$ で偏微分して $\mathbf{0}$ となるような $\boldsymbol{w} = \boldsymbol{w}^*_\infty$ が存在するので，

$$\begin{aligned} & \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p^{\pi_{\boldsymbol{\theta}}}_\infty(s)\pi_{\boldsymbol{\theta}}(a|s)(Q^\pi_\infty(s,a) - \hat{Q}_{\boldsymbol{w}^*_\infty}(s,a))\nabla_{\boldsymbol{w}_1}\hat{Q}_{\boldsymbol{w}}(s,a) \\ = & \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p^{\pi_{\boldsymbol{\theta}}}_\infty(s)\pi_{\boldsymbol{\theta}}(a|s)\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a|s)(Q^\pi_\infty(s,a) - \hat{Q}_{\boldsymbol{w}^*_\infty}(s,a)) = \mathbf{0} \end{aligned}$$

が成り立つ．以上より，

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} f_\infty(\boldsymbol{\theta}) &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p^{\pi_{\boldsymbol{\theta}}}_\infty(s)\pi_{\boldsymbol{\theta}}(a|s)\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a|s)Q^\pi_\infty(s,a) \\ &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p^{\pi_{\boldsymbol{\theta}}}_\infty(s)\pi_{\boldsymbol{\theta}}(a|s)\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a|s)\hat{Q}_{\boldsymbol{w}^*_\infty}(s,a) \end{aligned}$$

を得る．
エルゴート性を満たさない場合についても，

$$\begin{aligned} & \mathbb{E}^\gamma_{S_t,A_t\sim Pr}[\sum_{t=0}^\infty (Q^\pi(S_t,A_t) - \hat{Q}_{\boldsymbol{w}}(S_t,A_t))^2] \\ = & \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t=0}^\infty \gamma^t Pr[S_t = s, A_t = a](Q^\pi(s,a) - \hat{Q}_{\boldsymbol{w}}(s,a))^2 \\ = & \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\sum_{t=0}^\infty \gamma^t Pr[S_t = s]\pi(a|s)(Q^\pi(s,a) - \hat{Q}_{\boldsymbol{w}}(s,a))^2 \\ = & \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} d^{\pi_{\boldsymbol{\theta}}}(s)\pi(a|s)(Q^\pi(s,a) - \hat{Q}_{\boldsymbol{w}}(s,a))^2 \end{aligned}$$

とすれば，同様に証明できる．

(お気持ち)
アクタークリティックの際，関数 $\hat{Q}_{\boldsymbol{w}}$ は線形でも十分に行動価値関数 $Q^{\pi_{\boldsymbol{\theta}}}_\infty$ を表現しうる．また，ベースライン関数 $b_{\boldsymbol{w}_2}(s)$ は任意なので用いなくてもよく，この場合は $\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a|s)$ の張る空間で行動価値関数を近似すれば十分であ

ることがわかる．ただし，任意の $\boldsymbol{w}_1$ に対し，

$$\begin{aligned}
\hat{V}_{\boldsymbol{w}}(s) &= \sum_{a\in\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s)\hat{Q}_{\boldsymbol{w}}(s,a) \\
&= \sum_{a\in\mathcal{A}} \pi_{\boldsymbol{\theta}}(a|s)^\top \boldsymbol{w}_1 \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s) \\
&= {}^\top\boldsymbol{w}_1 \sum \nabla_{\boldsymbol{\theta}} \pi_{\boldsymbol{\theta}}(a|s) \\
&= {}^\top\boldsymbol{w}_1 \nabla_{\boldsymbol{\theta}} \sum \pi_{\boldsymbol{\theta}}(a|s) \\
&= {}^\top\boldsymbol{w}_1 \nabla_{\boldsymbol{\theta}} 1 = 0
\end{aligned}$$

が成り立ち，$V_\infty^\pi$ に対する表現能力を一切持たなくなってしまい，方策勾配法を $\hat{V}_{\boldsymbol{w}}$ で置き換えることができない．そこで，特徴ベクトル $\boldsymbol{\phi}(s)$ を追加した

$$\boldsymbol{\varphi}(s,a) \coloneqq \begin{pmatrix} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s) \\ \boldsymbol{\phi}(s) \end{pmatrix}$$

を用いて

$$\hat{Q}_{\boldsymbol{w}}(s,a) = {}^\top\boldsymbol{w}\boldsymbol{\phi}(s)$$

とすることが考えられる．

状態価値関数 $V_\infty^{\pi_{\boldsymbol{\theta}}}$ の TD 誤差を

$$\delta_t \coloneqq g(s_t,a_t) - f_\infty(\boldsymbol{\theta}) + V_\infty^{\pi_{\boldsymbol{\theta}}}(s_{t+1}) - V_\infty^{\pi_{\boldsymbol{\theta}}}(s_t)$$

と定義する．今，差分価値関数 $Q_\infty^{\pi_{\boldsymbol{\theta}}}$ についての再帰式：

$$\begin{aligned}
Q_\infty^{\pi_{\boldsymbol{\theta}}}(s,a) &= g(s,a) - f_\infty(\boldsymbol{\theta}) + \sum_{s'\in\mathcal{S}} p_T(s'|s,a) \sum_{a'\in\mathcal{A}} \pi_{\boldsymbol{\theta}}(a'|s') Q_\infty^{\pi_{\boldsymbol{\theta}}}(s',a') \\
&= g(s,a) - f_\infty(\boldsymbol{\theta}) + \sum_{s'\in\mathcal{S}} p_T(s'|s,a) V_\infty^{\pi_{\boldsymbol{\theta}}}(s')
\end{aligned}$$

より，TD 誤差の前半 3 項の期待値を取ると，

$$\begin{aligned}
& \mathbb{E}_{S_{t+1}|S_t,A_t\sim p_T, A_{t+1}|S_{t+1}\sim\pi}[g(S_t,A_t) - f_\infty(\boldsymbol{\theta}) + V_\infty^{\pi_{\boldsymbol{\theta}}}(S_{t+1})|S_t=s_t, A_t=a_t] \\
= \;& \mathbb{E}_{A_t|S_t\sim\pi, S_{t+1}|S_t,A_t\sim p_T}[g(S_t,A_t) - f_\infty(\boldsymbol{\theta}) + Q_\infty^{\pi_{\boldsymbol{\theta}}}(S_{t+1})|S_t=s_t, A_t=a_t] \\
= \;& g(s_t,a_t) - f_\infty(\boldsymbol{\theta}) + \sum_{s'\in\mathcal{S}} p_T(s'|s_t,a_t) \sum_{a'\in\mathcal{A}} \pi_{\boldsymbol{\theta}}(a'|s') Q_\infty^{\pi_{\boldsymbol{\theta}}}(s',a') \\
= \;& Q_\infty^{\pi_{\boldsymbol{\theta}}}(s_t,a_t)
\end{aligned}$$

となるので,

$$\mathbb{E}[\delta_t | S_t = s, A_t = a] = Q_\infty^{\pi_{\boldsymbol{\theta}}}(s_t, a_t) - V_\infty^{\pi_{\boldsymbol{\theta}}}$$

である．従って，$t$ 時点で $\{s_t, a_t, r_t, s_{t+1}\}$ が得られた時，$V_\infty^{\pi_{\boldsymbol{\theta}}}$ を $\hat{V}_{\boldsymbol{w}}$ で推定し，ベースライン関数を $\hat{V}_{\boldsymbol{w}}$ に設定して,

$$\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_t|s_t)(r_t - \hat{f}_t + \hat{V}_{\boldsymbol{w}}(s_{t+1}) - \hat{V}_{\boldsymbol{w}}(s_t))$$

を計算することで勾配を推定できる．