Conservative Policy Iteration まとめ

北村俊徳

May 9, 2023

目次

このスライドで説明する内容

方策更新で単調な性能の向上を保証するアルゴリズムの解説

1. Conservative Policy Iteration [Kakade and Langford, 2002]
2. Safe Policy Iteration [Pirotta et al., 2013]
3. Trust Region Policy Optimization [Schulman et al., 2015]
4. Proximal Policy Optimization [Schulman et al., 2017]
5. Constrained Policy Optimization [Achiam et al., 2017]

を解説するよ

表記 I

- $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma\mu \rangle$:
  状態空間, 行動空間, 遷移確率, 方策関数, 割引率, 初期分布
  (煩雑なので初期分布の表記は必要な箇所以外省略する)
- ベルマン方程式:

$$V^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \left(\mathcal{R}(s, a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}\left(s' \mid s, a\right) V^{\pi}\left(s'\right) \right) \tag{1}$$

- 目的関数:

$$J^{\pi} = \sum_{s \in \mathcal{S}} \mu(s) V^{\pi}(s) = \sum_{s \in \mathcal{S}} d^{\pi}(s) \sum_{a \in \mathcal{A}} \pi(a \mid s) \mathcal{R}(s, a) \tag{2}$$

ここで $d^{\pi} = \sum_{t=0}^{\infty} \gamma^t \Pr\left(s_t = s \mid \pi, \mu\right)$ は初期分布 μ から始まって π に従って動いた時の正規化されてない割引定常分布[1].

表記 II

- 方策 π に従って動いたときの期待値: $\mathbb{E}_{(a_t,s_t)\sim\pi}$
- 方策 π のアドバンテージ:

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s) \tag{3}$$

- π のアドバンテージの π' についての期待値

$$A_{\pi}^{\pi'}(s) = \sum_{a\in\mathcal{A}} \pi'(a \mid s) A^{\pi}(s,a) \tag{4}$$

- $A_{\pi}^{\pi'}(s)$ の定常分布 d^{π} についての期待値

$$\mathbb{A}_{\pi,d^{\pi}}^{\pi'} = \sum_{s\in\mathcal{S}} d^{\pi}(s) A_{\pi}^{\pi'}(s) \tag{5}$$

[1] 正規化の有無で式展開が結構変わる．注意しよう．

目次

CPI: 概要

基本的なアイディア

方策を $\pi' = \alpha\bar{\pi} + (1-\alpha)\pi$ の形で更新すると、目的関数について

$$J^{\pi'} - J^{\pi} \geq \alpha \mathbb{A}^{\bar{\pi}}_{\pi, d^{\pi}} - \frac{2\alpha^2\gamma\epsilon}{(1-\gamma(1-\alpha))} \tag{6}$$

の不等式が成立する. ここで, $\epsilon := \frac{1}{(1-\gamma)}\left(\max_s A^{\bar{\pi}}_{\pi}(s)\right)$
この左辺が常に 0 以上 & 最大になるように α を選びたい.

式 (6) の導出 I

まず, $J^{\pi'} - J^{\pi}$ をアドバンテージで表そう.
任意の方策 π' と π, π' に従って動いたときの定常分布 $d^{\pi'}$ について,

$$J^{\pi'} - J^{\pi} = \mathbb{A}^{\pi'}_{\pi, d^{\pi'}} \tag{7}$$

が成立する.

式 (6) の導出 II

式 (7) の導出

$$
\begin{aligned}
V^{\pi'}(s) &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{(a_t,s_t)\sim\pi'} \left[\mathcal{R}\left(s_t, a_t\right)\right] \\
&= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{(a_t,s_t)\sim\pi'} \left[\mathcal{R}\left(s_t, a_t\right) + V^{\pi}\left(s_t\right) - V^{\pi}\left(s_t\right)\right] \\
&= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{(a_t,s_t,s_{t+1})\sim\pi'} \left[\mathcal{R}\left(s_t, a_t\right) + \gamma V^{\pi}\left(s_{t+1}\right) - V^{\pi}\left(s_t\right)\right] + V^{\pi}(s) \\
&= V^{\pi}(s) + \mathbb{A}^{\pi'}_{\pi, d^{\pi'}}
\end{aligned}
$$

2 行目から 3 行目の変換: 時刻 t と $t+1$ で $\gamma V^{\pi}\left(s_{t+1}\right)$ と $\gamma V^{\pi}\left(s_t\right)$ が消し合い, $-V^{\pi}(s)$ が出てくるため.

式 (6) の導出 III

次に, $\mathbb{A}^{\pi'}_{\pi,d^{\pi'}}$ に関する不等式を考えよう.
式 (7) の π' を $\pi' = \alpha\bar{\pi} + (1-\alpha)\pi$ のように 2 つの方策を混ぜたものと考えると, 強力な不等式が出てくる. この α は行動が $\bar{\pi}$ から出る確率とみなせる.
まず, アドバンテージを同じ方策について期待値を取ると 0 になる性質を利用して,

$$
\begin{aligned}
A^{\pi'}_{\pi}(s) &= \sum_a \pi'(s;a) A^{\pi}(s,a) \\
&= \sum_a \left((1-\alpha)\pi(a;s) + \alpha\bar{\pi}(a;s)\right) A^{\pi}(s,a) \\
&= \alpha \sum_a \bar{\pi}(a;s) A^{\pi}(s,a) \\
&= \alpha A^{\bar{\pi}}_{\pi}(s)
\end{aligned}
$$

式 (6) の導出 IV

ここで, c_t を時刻 t 以前に行動が $\bar{\pi}$ から選ばれた回数を示す確率変数とし, $\rho_t \equiv \Pr(c_t \geq 1) = 1 - (1-\alpha)^t$ する. このとき,

$$
\begin{aligned}
&\mathbb{E}_{s\sim P(s_t;\pi')}\left[\alpha A_\pi^{\bar{\pi}}(s)\right] \\
=&\alpha(1-\rho_t)\,\mathbb{E}_{s\sim P(s_t|c_t=0;\pi')}\left[A_\pi^{\bar{\pi}}(s)\right] \\
&+\alpha\rho_t\mathbb{E}_{s\sim P(s_t|c_t\geq 1;\pi')}\left[A_\pi^{\bar{\pi}}(s)\right] \\
\geq&\alpha\mathbb{E}_{s\sim P(s_t|c_t=0;\pi')}\left[A_\pi^{\bar{\pi}}(s)\right] - 2\alpha\rho_t \max_s(A_\pi^{\bar{\pi}}(s)) \\
=&\alpha\mathbb{E}_{s\sim P(s_t;\pi)}\left[A_\pi^{\bar{\pi}}(s)\right] - 2\alpha\rho_t \max_s(A_\pi^{\bar{\pi}}(s))
\end{aligned}
$$

式 (6) の導出 V

式 (6) の導出

$$
\begin{aligned}
&J^{\pi'} - J^{\pi} = \mathbb{A}^{\pi'}_{\pi, d^{\pi'}} \\
=&\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{s \sim P(s_t; \pi')} \left[A^{\pi'}_{\pi}(s) \right] \\
=&\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{s \sim P(s_t; \pi')} \left[\alpha A^{\bar{\pi}}_{\pi}(s) \right] \\
\geq&\alpha \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{s \sim P(s_t; \pi)} \left[A^{\bar{\pi}}_{\pi} \right] - 2\alpha \max_s (A^{\bar{\pi}}_{\pi}(s)) \sum_{t=0}^{\infty} \gamma^t \left(1 - (1-\alpha)^t \right) \\
=&\alpha \left(\mathbb{A}^{\bar{\pi}}_{\pi, d^{\pi}} - \frac{2\alpha\gamma \max_s (A^{\bar{\pi}}_{\pi}(s))}{(1-\gamma)(1-\gamma(1-\alpha))} \right)
\end{aligned}
$$

最適な α の導出 (元論文)

報酬の最大値を R とすると, $\max_s \mid A^{\bar{\pi}}_{\pi}(s) \mid \leq R$ が成立する[2].
また, $0 \leq \alpha \leq 1$ なので,

$$\alpha \mathbb{A}^{\bar{\pi}}_{\pi, d^{\pi}} - \frac{2\alpha^2 \gamma \epsilon}{(1-\gamma(1-\alpha))} \geq \alpha \mathbb{A}^{\bar{\pi}}_{\pi, d^{\pi}} - \frac{2\alpha^2 R}{(1-\gamma)^2}$$

右辺は α についての二次式になっているので, $\mathbb{A}^{\bar{\pi}}_{\pi, d^{\pi}} \geq 0$ である場合,

$$\alpha = \frac{\mathbb{A}^{\bar{\pi}}_{\pi, d^{\pi}} (1-\gamma)^2}{4R} \tag{8}$$

で下界の最大値 $\frac{{\mathbb{A}^{\bar{\pi}}_{\pi, d^{\pi}}}^2 (1-\gamma)^2}{8R}$ を得る.

[2] 元論文ではこの記述なんですが, どう頑張ってもこの不等式の証明ができない... $\max_s \mid A^{\bar{\pi}}_{\pi}(s) \mid \leq \frac{R}{1-\gamma}$ になると思うんですよね. 実際 SPI で出てくる CPI の記述はこれになってない気がする.

最適な α の導出 (後で出てくる SPI で言及されてる版)

報酬の絶対値の最大値を R とすると, $\max_s \mid A_\pi^{\bar{\pi}}(s) \mid \leq \frac{2R}{1-\gamma}$ が成立する[3].
また, $0 \leq \alpha \leq 1$ なので,

$$\alpha \mathbb{A}_{\pi, d^\pi}^{\bar{\pi}} - \frac{2\alpha^2\gamma\epsilon}{(1-\gamma(1-\alpha))} \geq \alpha \mathbb{A}_{\pi, d^\pi}^{\bar{\pi}} - \frac{2\alpha^2\gamma\epsilon}{1-\gamma} \geq \alpha \mathbb{A}_{\pi, d^\pi}^{\bar{\pi}} - \frac{4\alpha^2\gamma R}{(1-\gamma)^3}$$

右辺は α についての二次式になっているので, $\mathbb{A}_{\pi, d^\pi}^{\bar{\pi}} \geq 0$ である場合,

$$\alpha = \frac{\mathbb{A}_{\pi, d^\pi}^{\bar{\pi}} (1-\gamma)^3}{8\gamma R} \tag{9}$$

で下界の最大値 $\frac{{\mathbb{A}_{\pi, d^\pi}^{\bar{\pi}}}^2 (1-\gamma)^3}{16\gamma R}$ を得る.

[3]報酬の最小値が 0 なら分子の 2 は不要.SPI の論文では報酬の最小値が 0 の時で比較している.

目次

表記 I

SPI は行列を使うと証明しやすいので, 行列メインの表記:

- $\mathbf{\Pi}^\pi$: $\mathbf{\Pi}^\pi(s,(s,a)) = \pi(a \mid s)$ なる $(|\mathcal{S}| \times |\mathcal{S}||\mathcal{A}|)$ の行列.
- $\mathbf{P}$: $\mathbf{P}\left((s,a),s'\right) = P(s' \mid s,a)$ なる $(|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|)$ の行列.
- $\mathbf{P}^\pi = \mathbf{\Pi}^\pi \mathbf{P}$: 方策 π に従う時の状態から状態への遷移行列. サイズ $|\mathcal{S}| \times |\mathcal{S}|$.
- $\mathbf{v}^\pi, \mathbf{r}^\pi, d^\pi, A^{\bar{\pi}}_\pi$: サイズ $|\mathcal{S}|$ のベクトル.

SPI: 概要

基本的なアイディア

CPI で出てきた

$$J^{\pi'} - J^{\pi} = \mathbb{A}^{\pi'}_{\pi, d^{\pi'}}$$

について, CPI とは違う下界:

$$J^{\pi'} - J^{\pi} \geq \mathbb{A}^{\pi'}_{\pi, d^{\pi}} - \frac{\gamma}{(1-\gamma)^2} \left\| \mathbf{\Pi}^{\pi'} - \mathbf{\Pi}^{\pi} \right\|_{\infty} \frac{\Delta \mathbf{A}^{\pi'}_{\pi}}{2} \tag{10}$$

を最大化することを考える.

式 (10) の導出 I

次の不等式 [Haviv and Heyden, 1984] を使って下界を導く:

$$\left|\mathbf{c}^T\mathbf{d}\right| \leq \|\mathbf{c}\|_1 \frac{\Delta\mathbf{d}}{2} \tag{11}$$

ここで, $\mathbf{d}$, $\mathbf{c}$ は任意の $\mathbf{c}^T\mathbf{e} = 0$ を満たすようなベクトルであり, $\Delta\mathbf{d} = \max_{i,j} |\mathbf{d}_i - \mathbf{d}_j|$.
これを使うと:

$$\begin{aligned} J^{\pi'} - J^{\pi} &= \mathbb{A}^{\pi'}_{\pi, d^{\pi'}} \\ &= \mathbf{d}^{\pi\,T}\mathbf{A}^{\pi'}_{\pi} + \left(\mathbf{d}^{\pi'} - \mathbf{d}^{\pi}\right)^T \mathbf{A}^{\pi'}_{\pi} \\ &\geq \mathbf{d}^{\pi\,T}\mathbf{A}^{\pi'}_{\pi} - \left\|\mathbf{d}^{\pi'} - \mathbf{d}^{\pi}\right\|_1 \frac{\Delta\mathbf{A}^{\pi'}_{\pi}}{2} \end{aligned} \tag{12}$$

これは $\mathbf{d}^{\pi'} - \mathbf{d}^{\pi}$ が平均 0 のベクトルのため.

式 (10) の導出 II

さらに, $\mathbf{d}^{\pi'}$ と $\mathbf{d}^{\pi}$ について, 以下の不等式が成立する:

$$\left\|\mathbf{d}^{\pi'} - \mathbf{d}^{\pi}\right\|_1 \leq \frac{\gamma}{1-\gamma} \left\|\mathbf{P}^{\pi'} - \mathbf{P}^{\pi}\right\|_\infty \left\|\left(\mathbf{I} - \gamma\mathbf{P}^{\pi'}\right)^{-1}\right\|_\infty \tag{13}$$

上の不等式を導くため, まず $\mathbf{d}^{\pi'} - \mathbf{d}^{\pi}$ を変形する[4]

$$\begin{aligned}
\mathbf{d}^{\pi'} - \mathbf{d}^{\pi} &= \gamma\mathbf{d}^{\pi'}\mathbf{P}^{\pi'} - \gamma\mathbf{d}^{\pi^\top}\mathbf{P}^{\pi} \\
&= \gamma\left(\mathbf{d}^{\pi'} - \mathbf{d}^{\pi}\right)\mathbf{P}^{\pi'} + \gamma\mathbf{d}^{\pi'}\left(\mathbf{P}^{\pi'} - \mathbf{P}^{\pi}\right) \\
&= \gamma\mathbf{d}^{\pi^\top}\left(\mathbf{P}^{\pi'} - \mathbf{P}^{\pi}\right)\left(\mathbf{I} - \gamma\mathbf{P}^{\pi'}\right)^{-1} \\
&\leq \gamma\left\|\mathbf{d}_\mu^{\pi\tau}\right\|_\infty \left\|\mathbf{P}^{\pi'} - \mathbf{P}^{\pi}\right\|_\infty \left\|\left(\mathbf{I} - \gamma\mathbf{P}^{\pi'}\right)^{-1}\right\|_\infty \\
&= \frac{\gamma}{1-\gamma}\left\|\mathbf{P}^{\pi'} - \mathbf{P}^{\pi}\right\|_\infty \left\|\left(\mathbf{I} - \gamma\mathbf{P}^{\pi'}\right)^{-1}\right\|_\infty
\end{aligned}$$

式 (10) の導出 III

目的の $\left\|\mathbf{d}_\mu^{\pi'} - \mathbf{d}_\mu^{\pi}\right\|_1$ について変形しなおして,

$$\begin{aligned}
\left\|\mathbf{d}_\mu^{\pi'} - \mathbf{d}_\mu^{\pi}\right\|_1 &\leq \frac{\gamma}{1-\gamma}\left\|\mathbf{P}^{\pi'} - \mathbf{P}^{\pi}\right\|_\infty \left\|\left(\mathbf{I} - \gamma\mathbf{P}^{\pi'}\right)^{-1}\right\|_\infty \\
&\leq \frac{\gamma}{1-\gamma}\left\|\mathbf{\Pi}^{\pi'} - \mathbf{\Pi}^{\pi}\right\|_\infty \|\mathbf{P}\|_\infty \sum_{t=0}^{\infty}\gamma^t \left\|\mathbf{P}^{\pi'}\right\|_\infty^t \\
&= \frac{\gamma}{(1-\gamma)^2}\left\|\mathbf{\Pi}^{\pi'} - \mathbf{\Pi}^{\pi}\right\|_\infty \qquad (14)
\end{aligned}$$

ここで, $\|\mathbf{P}\|_\infty = 1$ を使ってる.
あとは式 (12) に代入して終わり.

---

[4] 導出では $\sum_{t=0}^{\infty}(\gamma\mathbf{P}^{\pi'})^t$ がノイマン級数であるため, $\sum_{t=0}^{\infty}(\gamma\mathbf{P}^{\pi})^t = \left(\mathbf{I} - \gamma\mathbf{P}^{\pi'}\right)^{-1}$ が成立することを利用.

最適な α の導出

$$J^{\pi'} - J^{\pi} \geq \mathbb{A}^{\pi'}_{\pi,d^{\pi}} - \frac{\gamma}{(1-\gamma)^2} \left\| \mathbf{\Pi}^{\pi'} - \mathbf{\Pi}^{\pi} \right\|_{\infty} \frac{\Delta \mathbf{A}^{\pi'}_{\pi}}{2}$$

について, π' に $\alpha\bar{\pi} + (1-\alpha)\pi$ を代入すると,

$$J^{\pi'} - J^{\pi} \geq \alpha \mathbb{A}^{\bar{\pi}}_{\pi,d^{\pi}} - \alpha^2 \frac{\gamma}{(1-\gamma)^2} \left\| \Pi^{\bar{\pi}} - \Pi^{\pi} \right\|_{\infty} \frac{\Delta \mathbf{A}^{\bar{\pi}}_{\bar{\pi}}}{2} \tag{15}$$

なる不等式が出てくる. 右辺は $\alpha = \frac{(1-\gamma)^2 \mathbb{A}^{\bar{\pi}}_{\pi,d^{\pi}}}{\gamma \left\| \mathbf{\Pi}^{\bar{\pi}} - \mathbf{\Pi}^{\pi} \right\|_{\infty} \Delta \mathbf{A}^{\bar{\pi}}_{\bar{\pi}}}$ の時最大になり, この時の下界は

$$J^{\pi'} - J^{\pi} \geq \frac{(1-\gamma)^2 {\mathbb{A}^{\bar{\pi}}_{\pi,d^{\pi}}}^2}{2\gamma \left\| \Pi^{\pi} - \Pi^{\pi} \right\|_{\infty} \Delta \mathbf{A}^{\bar{\pi}}_{\pi}} \tag{16}$$

CPI との比較

SPI の下界について, $\Delta \mathbf{A}_{\pi}^{\bar{\pi}} < \frac{2R}{1-\gamma}$, $\|\Pi^{\pi} - \Pi^{\pi}\|_{\infty} < 2$ を使って CPI と比較すると:

- CPI の下界[5]: $J^{\pi'} - J^{\pi} \geq \frac{(1-\gamma)^3 \mathbb{A}_{\pi,d^{\pi}}^{\bar{\pi}}{}^2}{16\gamma R}$
- SPI の下界: $J^{\pi'} - J^{\pi} \geq \frac{(1-\gamma)^2 \mathbb{A}_{\pi,d^{\pi}}^{\bar{\pi}}{}^2}{2\gamma \|\Pi^{\pi} - \Pi^{\pi}\|_{\infty} \Delta \mathbf{A}_{\pi}^{\bar{\pi}}} \geq \frac{(1-\gamma)^3 \mathbb{A}_{\pi,d^{\pi}}^{\bar{\pi}}{}^2}{8\gamma R}$

よって SPI は CPI よりマシな下界になることがわかる.

[5] SPI の論文は報酬の最小値を 0 と仮定している. その場合分母の 16 は 8 になる.

目次

TRPO: 概要

基本的なアイディア

CPI の更新は α を使って方策同士を混ぜる: 扱いづらい...
$\rightarrow$ TRPO はもっと汎用的な形の下界に: 以下の不等式を利用
任意の方策 π' と π について,

$$
\begin{aligned}
J^{\pi'} - J^{\pi} &\geq \mathbb{A}^{\pi'}_{\pi, d^{\pi}} - \frac{4\epsilon\gamma}{(1-\gamma)} D^{\max}_{TV}(\pi, \pi')^2 \\
&\geq \mathbb{A}^{\pi'}_{\pi, d^{\pi}} - \frac{4\epsilon\gamma}{(1-\gamma)} D^{\max}_{KL}(\pi, \pi') \qquad (17)
\end{aligned}
$$

ここで, $\epsilon = \frac{1}{(1-\gamma)} \max_s \max_a |A^{\pi}(s,a)|$.
$D^{\max}_{KL}(\pi, \pi') = \max_s D_{KL}(\pi(s,\cdot) \| \pi'(s,\cdot))$,
$D^{\max}_{TV}(\pi, \pi') = \max_s D_{TV}(\pi(s,\cdot) \| \pi'(s,\cdot))$
式 (17) より, π と π' の KL ダイバージェンスを小さくしながら $\mathbb{A}^{\pi'}_{\pi, d^{\pi}}$ を最大化すると良さそう.

式 (17) の導出 I

以下, 不等式 (17) の導出を行う[6].
導出の前に, 以下の概念を導入する.

α-coupled

2 つの方策 π と π' について, $P(a \neq a' \mid s) \leq \alpha$ が全ての状態 s について成立しているなら, 2 つの方策は α-coupled であるという.
つまり, α-coupled な方策について, α が小さいほど 2 つの方策は近くなる.

式 (17) の導出 II

α-coupled な方策についての補題 1

α-coupled な方策 π と π' で, 全ての状態 s について, 次が成立:

$$\left|A_\pi^{\pi'}(s)\right| \leq 2\alpha \max_{s,a} |A_\pi(s,a)| \tag{18}$$

証明:

$$\begin{aligned} A_\pi^{\pi'}(s) &= \mathbb{E}_{a'\sim\pi'}\left[A_\pi(s,a')\right] = \mathbb{E}_{(a,a')\sim(\pi,\pi')}\left[A_\pi(s,a') - A_\pi(s,a)\right] \\ &= P(a \neq a' \mid s)\mathbb{E}_{(a,a')\sim(\pi,\pi')|a\neq a'}\left[A_\pi(s,a') - A_\pi(s,a)\right] \\ |A_\pi^{\pi'}(s)| &\leq \alpha \cdot 2 \max_{s,a} |A_\pi(s,a)| \end{aligned} \tag{19}$$

1 行目: 方策について期待値取るとアドバンテージが 0 になる.
2 行目: $a = a'$ のときはそもそも期待値の中身が 0 になる.

式 (17) の導出 III

α-coupled な方策についての補題 2

α-coupled な方策 π と π' について, 次が成立:

$$\begin{aligned}\left|\mathbb{E}_{s_t\sim\pi'}\left[A_\pi^{\pi'}(s_t)\right]-\mathbb{E}_{s_t\sim\pi}\left[A_\pi^{\pi'}(s_t)\right]\right| &\leq 2\alpha\max_s A_\pi^{\pi'}(s)\\ &\leq 4\alpha\left(1-(1-\alpha)^t\right)\max_s|A_\pi(s,a)|\end{aligned} \tag{20}$$

証明: 変数 n_t を時刻 $i<t$ で $a_i\neq a'_i$ が生じた回数とする. このとき, 方策についての期待値を n_t が 0 である場合と 0 以上の場合に分けると,

$$\mathbb{E}_{s_t\sim\pi'}\left[A_\pi^{\pi'}(s_t)\right]$$
$$=P(n_t=0)\,\mathbb{E}_{s_t\sim\pi'|n_t=0}\left[A_\pi^{\pi'}(s_t)\right]+P(n_t>0)\,\mathbb{E}_{s_t\sim\pi'|n_t>0}\left[A_\pi^{\pi'}(s_t)\right]$$
$$\mathbb{E}_{s_t\sim\pi}\left[A_\pi^{\pi'}(s_t)\right]$$
$$=P(n_t=0)\,\mathbb{E}_{s_t\sim\pi|n_t=0}\left[A_\pi^{\pi'}(s_t)\right]+P(n_t>0)\,\mathbb{E}_{s_t\sim\pi|n_t>0}\left[A_\pi^{\pi'}(s_t)\right]$$

式 (17) の導出 IV

ここで, $n_t = 0$ のときの期待値は方策 π から出てると考えて良いので, $\mathbb{E}_{s_t \sim \pi'}\left[A_\pi^{\pi'}(s_t)\right]$ と $\mathbb{E}_{s_t \sim \pi}\left[A_\pi^{\pi'}(s_t)\right]$ の差分は

$$
\begin{aligned}
&\mathbb{E}_{s_t \sim \pi'}\left[A_\pi^{\pi'}(s_t)\right] - \mathbb{E}_{s_t \sim \pi}\left[A_\pi^{\pi'}(s_t)\right] \\
=&P(n_t > 0)\left(\mathbb{E}_{s_t \sim \pi' | n_t > 0}\left[A_\pi^{\pi'}(s_t)\right] - \mathbb{E}_{s_t \sim \pi | n_t > 0}\left[A_\pi^{\pi'}(s_t)\right]\right) \qquad (21)
\end{aligned}
$$

式 (21) について,

- α の定義から, $P(n_t > 0) \leq 1 - (1-\alpha)^t$
- 式 (20) から,
  $\left|\mathbb{E}_{s_t \sim \pi' | n_t > 0}\left[A_\pi^{\pi'}(s_t)\right] - \mathbb{E}_{s_t \sim \pi | n_t > 0}\left[A_\pi^{\pi'}(s_t)\right]\right| \leq 4\alpha \max_{s,a} |A_\pi(s,a)|$

式 (17) の導出 V

これらを合わせると,

$$\begin{aligned}&\left|\mathbb{E}_{s_t\sim\pi'|n_t>0}\left[A_\pi^{\pi'}(s_t)\right]-\mathbb{E}_{s_t\sim\pi|n_t>0}\left[A_\pi^{\pi'}(s_t)\right]\right|\\&\leq 4\alpha\left(1-(1-\alpha)^t\right)\max_{s,a}|A_\pi(s,a)|\end{aligned}\tag{22}$$

以上で準備が整ったので, 元々の目標である

$$\begin{aligned}J^{\pi'}-J^{\pi}&\geq \mathbb{A}_{\pi,d^\pi}^{\pi'}-\frac{4\epsilon\gamma}{(1-\gamma)}D_{TV}^{\max}(\pi,\pi')^2\\&\geq \mathbb{A}_{\pi,d^\pi}^{\pi'}-\frac{4\epsilon\gamma}{(1-\gamma)}D_{KL}^{\max}(\pi,\pi')\end{aligned}\tag{23}$$

を導出しよう.

式 (17) の導出 VI

CPI の式 (7) で示したように, $J^{\pi'} = J^{\pi} + \mathbb{A}^{\pi'}_{\pi, d^{\pi'}}$ が成立.
一方, J^{π} は $L_{\pi}(\pi') = J^{\pi} + \mathbb{A}^{\pi'}_{\pi, d^{\pi}}$ と一次の項まで同じである.
以上を使って

$$\begin{aligned}
\left| J^{\pi'} - L_{\pi}(\pi') \right| &= \sum_{t=0}^{\infty} \gamma^t \left| \mathbb{E}_{s_t \sim \pi'} \left[A_{\pi}^{\pi'}(s_t) \right] - \mathbb{E}_{s_t \sim \pi} \left[A_{\pi}^{\pi'}(s_t) \right] \right| \\
&\leq \sum_{t=0}^{\infty} \gamma^t \cdot 4\epsilon\alpha \left(1 - (1-\alpha)^t\right) \qquad (24) \\
&= \frac{4\alpha^2\gamma\epsilon}{(1-\gamma)(1-\gamma(1-\alpha))} \leq \frac{4\alpha^2\gamma\epsilon}{(1-\gamma)^2}
\end{aligned}$$

ここで $\epsilon = \max_{s,a} |A_{\pi}(s,a)|$.
後は α を $\max_s D_{TV}(\pi(\cdot \mid s) \| \pi'(\cdot \mid s)) \leq \alpha$ で置き換えたら完成!

[6] ここでの導出は元論文 [Schulman et al., 2015] に基づいてるが, 元論文よりコロンビア大学の IEOR 8100 のLecture7の方が簡単かも

目次

CPO での performance bound

π と $\bar{\pi}$ についての performance bound

任意の関数 $f: S \to \mathbb{R}$ と任意の方策 $\bar{\pi}$, π について,

$$\delta_f(s, a, s') \doteq R(s, a, s') + \gamma f(s') - f(s),$$

$$\epsilon_f^{\pi'} \doteq \max_s \left| \mathrm{E}_{a \sim \pi', s' \sim P} \left[\delta_f(s, a, s') \right] \right|,$$

$$L_{\pi,f}(\pi') \doteq \mathop{\mathrm{E}}_{s \sim d^\pi} \left[\left(\frac{\pi'(a \mid s)}{\pi(a \mid s)} - 1 \right) \delta_f(s, a, s') \right],$$

$$D^{\pm}_{\pi,f}(\pi') \doteq \frac{L_{\pi,f}(\pi')}{1-\gamma} \pm \frac{2\gamma\epsilon_f^{\pi'}}{(1-\gamma)^2} \mathrm{E}_{s \sim d^\pi} \left[D_{TV}(\pi' || \pi)[s] \right]$$

を定義する.

π と $\bar{\pi}$ についての performance bound (続き)

このとき,

References


Achiam, J., Held, D., Tamar, A., and Abbeel, P. (2017).
Constrained policy optimization.
In *ICML*.


Haviv, M. and Heyden, L. V. D. (1984).
Perturbation bounds for the stationary probabilities of a finite markov chain.
Advances in Applied Probability, 16(4):804–818.


Kakade, S. and Langford, J. (2002).
Approximately optimal approximate reinforcement learning.
In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, page 267â 274, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.


Pirotta, M., Restelli, M., Pecorino, A., and Calandriello, D. (2013).
Safe policy iteration.
In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of