szepesvari 本: 数式補足

2.4 Dynamic programming algorithms for solving MDPs

Q^k に対してグリーディな方策に関するバウンドの証明

$$V^{\pi}(x) \geq V^{*}(x) - \frac{2}{1-\gamma} \|Q - Q^{*}\|_{\infty} \tag{1}$$

4.2 Closed-loop interactive learning

4.2.3 Active learning in Markov Decision Processes

MDP における active learning の理論研究は希少．以下，“Deterministic MDP” という単語が結構出てくるが，何が決定論的なのか曖昧．ひとまず状態遷移が決定論的，という意味で捉える．

決定論的な MDP における一様バウンド

- 遷移構造を知る
  - Efficient Exploration In Reinforcement Learning [Thrun 1992]
    主張: $n = |\mathcal{X}|, m = |\mathcal{A}|$ として，l を状態空間の深さ（とは？），d を各状態で許される行動の最大数としたとき，状態遷移構造を知るには $O(n^2 ld)$ 回の探索が必要（たぶん）
  - この本における改善の指摘
    主張: 各状態行動において，知らない行動を伴う状態の中で最も近い状態に到達するには高々 n-1 回かかる．そのあと知らない行動をとる．これを nm 回行えばよいので，$n^2 m$ 回の探索で十分（超自明）．これが漸近的にタイト（探索回数がこの式のオーダーで漸近的に上下から抑えられる）になる例を作れる．勉強会中のホワイトボードの例は $n(n+1)(m-1)/2 + n$ 回の探索が必要だったが、$k1 = 1/2, k2 = 1$ として上下から抑えられるので漸近的にタイト．
- 遷移構造が既知のときに，報酬関数（確率的）を推定する
  - Hoeffding の不等式により報酬関数の推定精度をバウンド
    主張: 各状態，行動の組について $k = \log(nm/\delta)/\epsilon^2$ 回以上訪問すれば，$1-\delta$ 以上の確率で，ϵ の精度で報酬関数を推定できる．
    本の中では明言していないが，報酬関数の値域が $[0, 1]$ であることを仮定しているはず．このと

き，サンプルサイズ $k \geq \log(2nm/\delta)/2\epsilon^2$ で，union bound と Hoeffding の不等式より

$$\begin{aligned}\Pr[\cap_{x,a}\{|\hat{r}(x,a) - r(x,a)| < \epsilon\}] &= 1 - \Pr[\cup_{x,a}\{|\hat{r}(x,a) - r(x,a)\}| \geq \epsilon] \\ &\geq 1 - \sum_{x,a} \Pr[|\hat{r}(x,a) - r(x,a)| \geq \epsilon] \\ &\geq 1 - \sum_{x,a} \delta/nm \\ &= 1 - \delta\end{aligned}$$

疑問: 本の主張と k の値が定数倍違うので著者に確認したい.

- 遷移構造が分かり，報酬関数を推定したあとで，方策を学習する
  - 方策の精度をバウンド

    主張 1: 各状態で最適価値関数と推定方策の価値関数の差が $4\gamma\epsilon/(1-\gamma)^2$ 以下になる（why?）.

    主張 2: 最終的に, ϵ-最適な方策を得るには, $\gamma \geq 0.5$ のとき, 高々 $n^2m+4e\log(nm/\delta)/((1-\gamma)^2\epsilon)^2$ のステップが必要.

    k の値に本の値を使うと，$\epsilon_r = (1-\gamma)^2\epsilon/2$ の精度で報酬関数を推定するには，各状態，行動の組について $4\log(nm/\delta)/((1-\gamma)^2\epsilon)^2$ 回の探索が必要で，このとき方策の精度は $2\gamma\epsilon$ なので，$\gamma \leq 0.5$ であれば ϵ-最適．全状態，行動を一回ずつ見るのに高々 e 回かかるとすると，単純に e 倍すればいい（第二項目）．第一項目は遷移構造を知る部分.

    疑問 1: 主張 1 の不等式の根拠がわからない

    疑問 2: $\gamma \geq 0.5$ だと ϵ-最適性を示せないと思う．不等号逆では？

確率的な MDP における一様バウンド

決定論的な MDP のときに示したような一様最適性を示した研究は，筆者の知る限りなかったらしい．Even-Dar らは，MDP の状態を任意の状態にリセットできるという（強い）仮定のもとで，有限 MDP における active learning 学習問題を考察したらしい.

ランダムな探索の難しさ

ランダムに探索を行うと，状態空間の全状態を訪問するまでに，状態空間のサイズ n に対して指数関数的な時間がかかる場合がある．例として，$1,\ldots,n$ のノードが並んだ一本鎖の状態空間を持ち，$\{L_1, L_2, R\}$ からなる行動空間を持つ MDP を考える．ここで，L_1, L_2 は左方向に一つ進み，R は右方向に一つ進む行動であり，端点から外に出ようとする行動をとったときには状態は動かないとする（状態遷移は決定論的）．このとき，端から端まで（1 から n まで）の移動にかかる時間の期待値は $3(2^n - n - 1)$ であると (Howard 1960) で示されたらしい．元論文にアクセスできないのでここで証明する (イメージは http://dopal.cs.uec.ac.jp/okamotoy/lect/2014/dme/handout13.pdf を参照).

Proof. X_t を時刻 t における状態 n までの距離とする ($0 \leq X_t \leq n-1$)．また, $T_k = \mathbb{E}[\min\{t|X_t = 0\}|X_0 = k]$ と定義する．これは状態 $n-k$ ($0 \leq k \leq n-1$) から状態 n までの移動にかかる時間の期待値を意味するので，T_{n-1} が求めるべき期待値の値である．漸化式は以下のように書ける.

$$T_k = \begin{cases} 0 & (k=0) \\ 1 + \frac{2}{3}T_{k+1} + \frac{1}{3}T_{k-1} & (1 \leq k \leq n-2) \\ T_{n-2} + 3 & (k = n-1) \end{cases}$$

ここで，$U_k = T_{k+1} - T_k$ $(0 \leq k \leq n-2)$ とおくと，

$$U_k = \frac{1}{2}U_{k-1} - \frac{3}{2}$$

が成り立つ．これを用いて漸化式をぐっと眺めると，以下のように展開できる．

$$\begin{aligned} T_k &= \Big(\sum_{i=0}^{k-1}(1/2)^i\Big)U_0 - 3\sum_{i=1}^{k-1}\sum_{j=1}^{i}(1/2)^j \\ &= 2(1-\frac{1}{2^k})U_0 - 3(k-2+\frac{1}{2^{k-1}}) \end{aligned}$$

ここに，$k = n-1$ の場合の式を代入して計算すると，

$$U_0 = 3(2^{n-1}-1)$$

となる．よって，求める期待値は，

$$\begin{aligned} T_{n-1} &= 2(1-\frac{1}{2^{n-1}})3(2^{n-1}-1) - 3(n-3+\frac{1}{2^{n-2}}) \\ &= 3(2^n - 2 - n + 3 - \frac{1}{2^{n-2}} - 2 + \frac{1}{2^{n-2}}) \\ &= 3(2^n - n - 1) \end{aligned}$$

□

もしこの MDP における報酬が状態 n でしか正にならなかったとすると，ランダムな探索を通して全状態行動対への十分な訪問を行ってから活用を行うような戦略では，報酬ゼロが続いて regret が大きくなる．なお，行動価値関数に基づいた素朴な探索戦略をとってもこの問題はそれほど改善しないらしい．一方，系統的に探索を行う戦略をとれば，$O(n)$ で全状態（ないしは全状態行動対）を訪問できるので，尊い．

リセットのない active learning

これまでの話とよく似ているが，推定方策の “一様な” 最適性ではなく，trajectory の上で訪問した状態における最適性を評価するところがおそらく違っている．このように評価方法を緩めると，確率的な MDP でも色々なバウンドが出せる．model-based な手法をいくつか紹介．

- E^3 (Explicit Explore or Exploit) アルゴリズム
  https://www.cis.upenn.edu/ mkearns/papers/reinforcement.pdf の P22 参照
  以下の戦略をとると，discounted な場合は多項式オーダーで探索が終わる．
  – 一定回数以上訪問されていない状態からは，過去に最もとっていない行動を選ぶ
  – 一定回数以上訪問された状態からは，そのときに推定された方策が一定以上良ければ，活用を行う．
  疑問 1: 論文見てると，活用フェーズで，undiscounted な場合は T（mixing time）回活用を行う（停止はしない）が，discounted な場合は停止するように見える．理解合ってる？ respectively が乱用されてて読めない．
  疑問 2: 翻訳の “そこへの探索は止める” という表現がよくわからない．“そこ” が良い方策が見つかった状態を意味するとしたら，“その状態に向かう探索を止める” というよりは，“その状態からは探索ではなく活用を行う” というイメージに見える．“そこへの” がない方が良いかも？

- R-max アルゴリズム: E^3 の改良
  以下の戦略をとると，（active learning における最適性を満たすまでの）探索回数が $\tilde{O}\left(\frac{n^2 m V_{max}^3}{\epsilon^3 (1-\gamma)^3}\right)$
  - 観測が十分にない状態行動対に対しては，報酬を最大値に固定する
  - 観測が十分たまったものについては，価値反復して方策を学習
- Domingo による適応サンプリング
  状態遷移がほぼ決定論的なときには効率が良いらしい
- Simsek と Barto による実問題でのパフォーマンス評価もあるらしい

実用性の評価はあまり行われていない．E^3 や R-max は undiscounted な場合では ϵ-mixing time が分からないといつアルゴリズムを停止させるべきかわからない．

4.2.4 Online learning in Markov Decision Processes

MDP における探索活用並行学習に話を戻す．regret により評価する UCRL2 と，PAC-MDP という基準で評価する遅延 Q 学習と MORMAX の概要を説明する．おまけとして，KWIK とベイジアンアプローチの紹介も行う．

regret 最小化と UCRL2

ざっくり言うと，新しい情報が十分収集できるまでは今の推定方策に従って情報を集め，収集できたらモデルと方策を更新する Model-based な手法である．方策の更新部分が本のアルゴリズム 11（OPTSOLVE）に相当し，肝になる．これは元論文（http://www.jmlr.org/papers/volume11/jaksch10a/jaksch10a.pdf）では Extended Value Iteration（拡張価値反復）と呼ばれている．

理論保証については，報酬の値域が $[0, 1]$ で，すべての決定論的方策が全状態を確率 1 で訪問する（unichain な）MDP のクラスを考えている．また，MDP の直径 D（ある状態から他の状態にたどり着くために要する平均ステップ数の最大値）という概念を導入している．示しているバウンドは二通り．

- 最適方策と次善方策の性能の差を g を使う regret バウンド

$$O(D^2 n^2 m \log T / g)$$

- g を使わない regret バウンド

$$O(Dn\sqrt{mT \log T})$$

後者は T については悪化するが，前者は g が小さいとき緩いバウンドになることに注意．なお，直径 D が無限大のときは無意味なバウンドになる．MDP のいくつかのパラメータが与えられたという仮定のもとでは，Bartlett and Tewari (2009) により D が無限大のときの上界が示されている．

理論はつらいので，以降では拡張価値反復のアルゴリズムを説明する．よくある価値反復は，報酬関数 $r(x, a)$ と確率遷移カーネル $\mathcal{P}(\cdot, x, a)$ を用いて，各状態 $x \in \mathcal{X}$ に対して，適当な終了条件のもとで以下の反