逐次意思決定における諸問題設定と問題に関する事前知識が性能保証に及ぼす影響について

On Various Problem Settings of Sequential Decision Making and How Prior Knowledge of Problems Affects Theoretical Guarantees

小津野 将 [*1] Tadashi Kozuno　北村 俊徳 [*2] Toshinori Kitamura　市原 有生希 [*3] Yuki Ichihara　萩原 誠 [*4] Makoto Hagiwara

[*1]オムロンサイニックエックス Omron Sinic X (OSX)　[*2]東京大学 Univ. of Tokyo　[*3]NAIST, ATR NAIST, ATR　[*4](株)pluszero pluszero

Recently, various problem settings of sequential decision making are proposed, and various types of performance guarantees are given. Examples are non-stationary MDPs and constrained MDPs. In this paper, we summarize and report recent development in performance guarantees for those settings, and we explain what kind of prior knowledge is useful for a better performance guarantee. Then, finally, we discuss open problems and interesting future study directions in the theory of sequential decision making.

1　序論

近年，様々な逐次意思決定の問題設定が考えられ，問題設定に対して様々な性能保証が示されている．非定常 MDP や制約付き MDP などがその例となる．本論文では，諸設定に対する性能保証の近年の発展をまとめ，どのような事前知識（問題設定に依存するパラメータ）が性能向上に有益となるかを説明する．そして最後に，現在未解決の問題と将来の逐次意思決定理論の方向に関して議論する．本論文が理論家の方には将来の研究テーマの検討の助けに，その他の方にはアルゴリズムの設計における事前知識の入れ方の参考になれば幸いである．[*1]

記法．任意の正の整数 N に対し，$[N]$ を 1 から N までの整数の集合とする．任意の有限集合 $\mathbf{S}$ に対し，$\Delta(\mathbf{S})$ を $\mathbf{S}$ 上の確率分布全体の集合とする．また，$[0,1]^{\mathbf{S}}$ や $\mathbb{R}^{\mathbf{S}}$ はそれぞれ $\mathbf{S}$ から $[0,1]$ や $\mathbb{R}$ への写像全体を表す．本論文では問題に関係ない定数などを無視する．そのための表記として $\mathcal{O}$ や Ω を用いる．$\mathcal{O}(x)$ はサンプル複雑性やリグレットの上界が定数を無視すると x となることを意味する．一方，$\Omega(x)$ は下界がそうなることを意味する．また，$\widetilde{\mathcal{O}}$ と $\widetilde{\Omega}$ は問題に関係した量の内，対数の多項式項（例えば $(\ln x)^2$ 等）を無視する．

2　マルコフ決定過程（MDP）

逐次意思決定の諸問題設定はマルコフ決定過程（MDP）を基にしていることが多い．MDP には有限時間 MDP と無限時間 MDP が存在するが，前者の方が多くの理論解析の結果が得られているため，本論文では有限時間 MDP の結果のみをまとめ，有限時間 MDP を単に MDP と呼ぶ．

MDP は $(\mathbf{S}, \mathbf{A}, \mathbf{r} \coloneqq \{r_h\}_{h=1}^H, \mathbf{P} \coloneqq \{P_h\}_{h=1}^{H-1}, \mu, H)$ の五つの要素で定義される．ここで，$\mathbf{S}$ は要素数が S の有限状態集合，$\mathbf{A}$ は要素数が A の有限状態集合，$r_h \in [0,1]^{\mathbf{S}\times\mathbf{A}}$ は報酬関数，$P_h \in \Delta(\mathbf{S})^{\mathbf{S}\times\mathbf{A}}$ は遷移確率関数，$\mu \in \Delta(\mathbf{S})$ は初期状態分布，$H \in \mathbb{N}$ は時間長となる．r_h と P_h が時刻 h に依存しない MDP を均一 MDP と呼び，その他を不均一 MDP と呼ぶ．

これらの要素は次のように用いられる．時刻 $h\,(\neq H)$ において，エージェントは S_h を観測し，S_h に基づいて行動 A_h を決定，実行する．その結果，報酬 $r_h(S_h, A_h)$ がエージェントへと与えられ，次の状態 S_{h+1} が $P_h(S_h, A_h)$ からサンプルされる．時刻が 1 のときは S_1 が μ よりサンプルされ，時刻が H のときは，$r_H(S_H, A_H)$ がエージェントへと与えられた後に，時刻が 1 へとリセットされ，再び上記のことを繰り返す．時刻 1 から時刻 H までの一連の MDP・エージェントの相互作用をエピソードと呼ぶ．[*2]

エージェントが行動を決定するルールを方策と呼び，$\pi \coloneqq \{\pi_h \in \Delta(\mathbf{A})^{\mathbf{S}}\}_{h=1}^H$ で表す．方策 π と任意の $\mathbf{g} \coloneqq \{g_h \in \mathbb{R}^{\mathbf{S}\times\mathbf{A}}\}_{h=1}^H$ に対し，$\mathbf{g}$ についての行動価値関数 $Q_{\mathbf{g},h}^\pi$ と状態価値関数 $V_{\mathbf{g},h}^\pi$ を以下のように定める．時刻 H に対しては $Q_H^\pi = g_H$ で，その他の時刻に関しては，$Q_{\mathbf{g},h}^\pi(s,a) \coloneqq g_h(s,a) + \sum_{s'\in\mathbf{S}} P_h(s'|s,a) V_{\mathbf{g},h+1}^\pi(s')$．ただし，$V_{\mathbf{g},h}^\pi(s) \coloneqq \sum_{a\in\mathbf{A}} \pi_h(a|s) Q_{\mathbf{g},h}^\pi(s,a)$．また期待リターンを $\rho_{\mathbf{g}}^\pi = \sum_{s\in\mathbf{S}} \mu(s) V_{\mathbf{g},1}^\pi(s)$ と定める．$\mathbf{g}$ についての最適方策の価値関数および期待リターンを $Q_{\mathbf{g},h}^*$，$V_{\mathbf{g},h}^*$ および $\rho_{\mathbf{g}}^*$ と記す．

3　通常の問題設定における結果

通常の問題設定における学習アルゴリズムの性能指標を以下に定義し [5]，後ほど，それらに関する説明を加える．以下の定義においては，アルゴリズムが各エピソード k の初めに方策 π_k を出力し，エージェントは π_k に従うとする．

定義 1 (リグレット)**.** 正の定数 $\delta \in (0,1)$ を固定する．このとき（擬）リグレットを $R(K) \coloneqq \sum_{k=1}^K (\rho_{\mathbf{r}}^* - \rho_{\mathbf{r}}^{\pi_k})$ と定義する．アルゴリズムが任意の MDP に対して以下の不等式を満たす多項式 F_{HPR} を持つ場合，その多項式を性能の指標とする：確率 $1-\delta$ で

$$R(K) \leq F_{\mathrm{HPR}}\left(S, A, H, K, \ln\frac{1}{\delta}\right).$$

定義 2 (PAC-MDP)**.** 正の定数 $\delta \in (0,1)$ と $\varepsilon \in (0, H)$ を固定する．このとき最適性ギャップを $\Delta(k) \coloneqq \rho_{\mathbf{r}}^* - \rho_{\mathbf{r}}^{\pi_k}$ で定義し，失敗回数を $N_\varepsilon \coloneqq \sum_{k=1}^\infty \mathbb{I}\{\Delta_k > \varepsilon\}$ とする．アルゴリズムが任意の MDP に対して以下の不等式を満たす多項式 F_{PAC}

連絡先: 小津野 将, OSX, 東京都文京区本郷五丁目 24 番 5 号 ナガセ本郷ビル 3F, tadashi.kozuno@sinicx.com

*1　詳細版は次の URL より利用可能：http://bit.ly/3YYr319

*2　理論解析においては確率的報酬を考えないことが多い．学習においては価値推定の分散が問題となるが，確率的報酬に起因する分散よりも確率的状態遷移に起因する分散の方が大きく，報酬が確率的であっても問題の難しさは本質的に変わらないためだ．

表 1: さまざまな問題設定のまとめ．報酬関数と遷移確率関数の空欄は単一の固定されたものを使うことを意味する．

設定名称	報酬関数 $\mathbf{r}$	遷移確率関数 $\mathbf{P}$	目的関数	文献
敵対的 MDP	エピソード依存		定義 5	[1, 2]
制約付き MDP			定義 9 or 定義 10	[3, 4]
時間長非依存強化学習			定義 1	[5]
非定常 MDP	エピソード依存	エピソード依存	定義 7	[6]
無報酬強化学習	全報酬関数		定義 8	[7]

を持つ場合，その多項式を性能の指標とする：確率 $1-\delta$ で

$$N_\varepsilon \leq F_{\mathrm{PAC}}\left(S, A, H, \frac{1}{\varepsilon}, \ln\frac{1}{\delta}\right).$$

定義 3 (一様 PAC-MDP)．正の定数 $\delta \in (0,1)$ を固定する．アルゴリズムが任意の MDP に対して以下の不等式を満たす多項式 F_{UPAC} を持つ場合，その多項式を性能の指標とする：確率 $1-\delta$ で

$$\forall \varepsilon \in (0,H) \text{ に対して } N_\varepsilon \leq F_{\mathrm{UPAC}}\left(S, A, H, \frac{1}{\varepsilon}, \ln\frac{1}{\delta}\right).$$

定義 4 (最適方策識別)．正の定数 $\delta \in (0,1)$ と $\varepsilon \in (0,H)$ を固定する．アルゴリズムが任意の MDP に対して以下の条件を満たす非負整数値確率変数 τ を持つ場合，τ の期待値を性能の指標とする：アルゴリズムは確率 1 で有限の τ エピソード後に停止して方策 π を出力し，確率 $1-\delta$ で $\rho^*_{\mathbf{r}} - \rho^{\pi}_{\mathbf{r}} \leq \varepsilon$．

F_{HPR} が $o(K)$ となるアルゴリズムを無後悔アルゴリズムと呼ぶ．最悪ケースを考える場合，情報理論的に達成可能な F_{HPR} の下界については均一 MDP で $\widetilde{\Omega}(\sqrt{SAH^2K})$，不均一 MDP で $\widetilde{\Omega}(\sqrt{SAH^3K})$ であることが知られており [8]，実際に F_{HPR} がそれらに一致するアルゴリズムも存在する [9].[*3] [*4]

一方，F_{PAC} と F_{UPAC} は存在しない場合がある．それらが存在するアルゴリズムをそれぞれ，PAC-MDP アルゴリズム，一様 PAC-MDP アルゴリズムと呼ぶ．一様 PAC アルゴリズムは PAC アルゴリズムとなり，一様 PAC アルゴリズムは無後悔アルゴリズムにもなる [10]．最悪ケースを考える場合，達成可能な F_{PAC} の下界（よって F_{UPAC} の下界）については均一 MDP で $\widetilde{\Omega}(SAH^2/\varepsilon^2)$，不均一 MDP で $\widetilde{\Omega}(SAH^3/\varepsilon^2)$ であることが知られており [5]，均一 MDP では実際にそれを達成するミニマックス最適アルゴリズムが存在する [11]．

実は無後悔学習アルゴリズムがあれば，PAC-MDP アルゴリズムに近いことが出来る．$R(K) \leq CK^{1-1/\alpha}$ とすると，$(\pi_k)_{k=1}^K$ の平均リターンは $\sum_{k=1}^K \rho^{\pi_k}_{\mathbf{r}}/K \geq \rho^*_{\mathbf{r}} - CK^{-1/\alpha}$ となる．そのため $K = (C/\varepsilon)^\alpha$ とすれば，平均リターンと $\rho^*_{\mathbf{r}}$ との差が ε となる [12]．これは batch-to-online conversion としてオンライン学習やバンディット理論では知られていた方法である [13]．しかし，それらのような一ステップだけの意思決定問題においては平均的な方策に相当する方策が存在するが，MDP のような逐次意思決定問題においては存在すると限らない．そのため，上記の方法ではよい方策を出力できない．

最適方策識別においても定義 4のような τ が存在しない場合がある．存在する場合，そのアルゴリズムを (ϵ,δ)-PAC アルゴリズムと呼ぶ．最適方策識別は他の指標とは異なり，学習中にエージェントがどのような方策に従っているかは気にしない．その代わり，最終的に出力された方策が最適方策に近いことを要求する．最悪ケースを考える場合，情報理論的に達成可能な $\mathbb{E}[\tau]$ の下界については均一 MDP で $\widetilde{\Omega}(SAH^2/\varepsilon^2)$，不均一 MDP で $\widetilde{\Omega}(SAH^3/\varepsilon^2)$ であることが知られており [5, 10]，不均一 MDP では上界がこの下界と一致するミニマックス最適アルゴリズムが存在する [14]．なお，PAC-MDP アルゴリズムは (ϵ,δ)-PAC アルゴリズムでもあるため，上記の Dann et al. [11] によるアルゴリズムが均一 MDP でミニマックス最適となる．

上記の結果は最悪ケースを想定した解析であり，問題依存的な解析も存在するが，本論文では割愛する．

4 様々な問題設定における結果

本章では，様々な問題設定とそれらにおける結果をまとめる．特に，問題設定に関する事前知識を利用して性能を向上させた結果を紹介する．ただし本論文では，“基礎となる問題設定に追加された情報であり，かつアルゴリズムで使用されている情報” を事前知識と呼ぶ．例えば4.1 節で紹介する Ghasemi et al. [15] は遷移確率関数 $\mathbf{P}$ を事前知識としてアルゴリズムで使用している．表 1では，本論文でまとめる問題設定を報酬関数と遷移確率関数がエピソード毎に変化するか，目的関数が何かという二つの観点から類別している．

4.1 敵対的 MDP（Adversarial MDP）

敵対的 MDP は報酬関数がエピソード毎に変化する状況をモデル化するために提案された [1, 2]．アルゴリズムの性能指標としては定義 1に類似した以下のものが用いられる．エピソード k の報酬関数を $\mathbf{r}_k$ とする．

定義 5 (敵対的 MDP におけるリグレット)．ある正の定数 $\delta \in (0,1)$ を固定する．このときリグレットを $R_{\mathrm{adv}}(K) := \max_\pi \sum_{k=1}^K \left(\rho^{\pi}_{\mathbf{r}_k} - \rho^{\pi_k}_{\mathbf{r}_k}\right)$ と定義し，アルゴリズムが任意の MDP に対して以下の不等式を満たす多項式 F_{AdvHPR} を持つ場合，その多項式を性能の指標とする：確率 $1-\delta$ で，任意の報酬関数列 $(\mathbf{r}_k)_{k=1}^K$ に対し，

$$R_{\mathrm{adv}}(K) \leq F_{\mathrm{AdvHPR}}\left(S, A, H, K, \ln\frac{1}{\delta}\right).$$

定義 1とはリグレットの定義が異なることに注意してほしい．ここで，$\sum_{k=1}^K \left(\rho^*_{\mathbf{r}_k} - \rho^{\pi_k}_{\mathbf{r}_k}\right)$ でリグレットを定義する方が

*3 Azar et al. [9] は均一 MDP を考えているが，彼らの結果を不均一 MDP に拡張すると $\sqrt{H}$ 倍の F_{HPR} となる．より一般に，均一 MDP 用アルゴリズムのリグレットを $\sqrt{H}$ 倍すると不均一 MDP に拡張したアルゴリズムのリグレットになると考えられている．F_{PAC}, F_{UPAC}, $\mathbb{E}[\tau]$ などのサンプル複雑性の指標は H 倍すると，不均一 MDP に拡張したアルゴリズムのサンプル複雑性が得られると考えられている．

*4 性能指標の最悪ケースを考えた場合の情報理論的下界とアルゴリズムの上限が一致する場合，そのようなアルゴリズムをミニマックス最適と呼ぶ．単にミニマックス最適と呼んだ場合は，文脈によってどの性能指標が考えられているかに注意．また，K が大きい場合や ε が小さい場合などにしか成立しない場合もあるが，そういった点に関しては割愛する．

表 2: 敵対的 MDP における事前知識と性能保証の比較.

	事前知識	F_{AdvHPR}	文献
下界	-	$\widetilde{\Omega}\left(\sqrt{SAHK}\right)$	[16]
上界	$\mathbf{P}$	$\widetilde{\mathcal{O}}\left(\sqrt{SAHK}\right)$	[15]
	-	$\widetilde{\mathcal{O}}\left(SH\sqrt{AK}\right)$	[17]

表 3: 非定常 MDP における事前知識と性能保証の比較.

	事前知識	F_{DynaHPR}	文献
下界	-	$\widetilde{\Omega}\left(\sqrt{SAHK}\right)$	[6]
上界	-	$\widetilde{\mathcal{O}}(\sqrt[3]{\Delta_K SAH^7K^2})$	[18]
	Δ_K	$\widetilde{\mathcal{O}}(\sqrt[3]{\Delta_K SAH^5K^2})$	[6]

自然に思うかもしれない．実はこの定義だと，どんなアルゴリズムも線形のリグレットしか持ちえない状況を簡単に作り出せるため，リグレットとしては適切ではない．(より正確に述べると，報酬関数を知っている “ズルい” アルゴリズムを比較対象に選ぶことが適切ではない.)

具体的な例として，$H=1$ の場合を考える．また，各エピソード毎に一様に行動を選び，その行動のみ報酬が 1，その他の行動は報酬が 0 とする．このとき，学習アルゴリズムは行動をどのように選んでも確率 $1/A$ でしか最適行動を選べない．そのため，$\sum_{k=1}^{K}\left(\rho^*_{\mathbf{r}_k}-\rho^{\pi_k}_{\mathbf{r}_k}\right)\approx K-K/A$ となる．

実は，敵対的 MDP における理論解析の結果は少ない．遷移確率関数 $\mathbf{P}$ が事前知識として既知な場合は，F_{AdvHPR} が $\widetilde{\mathcal{O}}(\sqrt{SAHK}+H\sqrt{K})$ となるミニマックス最適なアルゴリズムが提案されている [16, 15].[*5] 一方，遷移確率関数が未知の場合は F_{AdvHPR} が $\widetilde{\mathcal{O}}(SH\sqrt{AK})$ となるアルゴリズムが提案されているが [17]，このアルゴリズムがミニマックス最適であるかは不明である．

敵対的 MDP における結果を表 2にまとめた．

4.2 非定常 MDP （Non-stationary MDP）

非定常 MDP では，報酬関数および遷移確率関数がエピソード毎に変化する状況を取り扱う．敵対的 MDP との違いは，遷移確率関数が変化する点および MDP の変化量に依存したリグレットを考える点である．変化量は以下の定義が用いられる．

定義 6 (非定常性指標 [18])**.** エピソード k における方策 π のリターンを ρ^{π}_k とする．このとき $\Delta_K \coloneqq \sum_{k=1}^{K-1}\max_{\pi}|\rho^{\pi}_k-\rho^{\pi}_{k+1}|$ および $L_K \coloneqq \sum_{k=1}^{K-1}\mathbb{I}\{\max_{\pi}|\rho^{\pi}_k-\rho^{\pi}_{k+1}|\neq 0\}$ を非定常性指標とする．[*6]

これらを用い，性能指標を次のように定義する．なお，エピソード k における最適方策のリターンを ρ^*_k とする．

定義 7 (ダイナミックリグレット)**.** ある正の定数 $\delta\in(0,1)$ を固定する．このときダイナミックリグレットを $R_{\mathrm{Dynamic}}(K) \coloneqq \sum_{k=1}^{K}\left(\rho^*_k-\rho^{\pi_k}_k\right)$ と定義し，アルゴリズムが任意の MDP に対して以下の不等式を満たす多項式 F_{DynaHPR} を持つ場合，その多項式を性能の指標とする：確率 $1-\delta$ で

$$R_{\mathrm{Dyna}}(K)\leq F_{\mathrm{DynaHPR}}\left(S,A,H,K,\Delta_K,L_K,\ln\frac{1}{\delta}\right).$$

F_{DynaHPR} の下界に関しては $\widetilde{\Omega}(\sqrt[3]{\Delta_K SAHK^2})$ が知られている [6]．一方，上界に関しては事前知識なしで $\widetilde{\mathcal{O}}(\min\{\sqrt{L_K SAH^5K},\sqrt[3]{\Delta_K SAH^7K^2}+\sqrt{SAH^5K}\})$ を達成するアルゴリズムが存在する [18]．Δ_K を知っている場合には $\widetilde{\mathcal{O}}(\sqrt[3]{\Delta_K SAH^5K^2}+\sqrt{SAH^3K})$ の上界を達成するアルゴリズムが存在する [6]．事前知識なしで $\widetilde{\mathcal{O}}(\sqrt[3]{\Delta_K SAHK^2})$ の上界を達成できるかは今のところ不明である．これらの結果を表 3にまとめた．

非定常 MDP における新たな研究の方向性として，後ほど説明する制約付き MDP との組み合わせが研究されている [19]．また，定義 7のダイナミックリグレットは適切ではないという指摘がバンディット理論の文脈で指摘されており，上記の結果についても同様に適切ではない可能性がある [20]．そのため，Liu et al. [20] が提唱するリグレットで上記のアルゴリズムを解析しなおす必要があるかも知れない．

4.3 無報酬強化学習 （Reward-Free 強化学習）

無報酬強化学習では，任意の報酬関数に対して最適方策に近い方策を出力できるように学習することを目標とする [7]．そのため，無報酬強化学習アルゴリズムの性能指標には定義 4を次のように拡張した指標が用いられることが多い．

定義 8 (無報酬強化学習での最適方策識別)**.** ある正の定数 $\delta\in(0,1)$ と $\varepsilon\in(0,H)$ を固定する．アルゴリズムが任意の無報酬 MDP, $(\mathbf{S},\mathbf{A},H,\mathbf{P},\mu)$, に対して以下の条件を満たす非負整数値確率変数 τ を持つ場合，τ の期待値を性能の指標とする：アルゴリズムは確率 1 で有限の τ エピソード後に停止し，報酬関数を入力として方策を出力する関数 Θ を出力し，少なくとも確率 $1-\delta$ で $\max_{\mathbf{r}}(\rho^*_{\mathbf{r}}-\rho^{\Theta(\mathbf{r})}_{\mathbf{r}})\leq\varepsilon$.

無報酬強化学習で達成可能な $\mathbb{E}[\tau]$ の下界は $\widetilde{\Omega}\left(S^2AH^3/\varepsilon^2\right)$ であることが指摘されている [7].[*7] この下界に対して，Zhang et al. [21] は $\mathbb{E}[\tau]$ の上界が $\widetilde{\mathcal{O}}\left(S^2AH^3/\varepsilon^2\right)$ となるミニマックス最適なアルゴリズムを提案した．さらに，$\sum_{h=1}^{H}r_h\leq 1$ a.s を仮定した MDP を考え，この仮定を事前知識として利用した場合には，Zhang et al. [21] のアルゴリズムは均一 MDP では $\widetilde{\mathcal{O}}\left(S^2A/\varepsilon^2\right)$，不均一 MDP では $\widetilde{\mathcal{O}}\left(S^2AH/\varepsilon^2\right)$ の上界となることが知られている．

この下界は前述の通常の最適方策識別問題の下界と比較すると S 倍大きい．これは無報酬強化学習が任意の報酬関数の最悪ケースを考えるためである．これを回避するために Zhang et al. [22] は任意の報酬関数の代わりに有限な N_{rew} 個の報酬関数に対しての最適方策識別を考える問題設定を提案した．つまり，定義 8において $\max_{\mathbf{r}\in\{\mathbf{r}_n\}_{n\in N_{\mathrm{rew}}}}\left(\rho^*_{\mathbf{r}}-\rho^{\mathcal{A}(\mathbf{r})}_{\mathbf{r}}\right)$ を上から抑えることを考える．N_{rew} が事前知識として与えられている場合に $\mathbb{E}[\tau]$ の上界が $\widetilde{\mathcal{O}}(SAH^5\log(N_{\mathrm{rew}})/\varepsilon^2)$ となるアルゴリズムが存在する [22]．このアルゴリズムが事前知識がある場合にミニマックス最適であるかは不明である．

*5 これらの結果ではループがない環境を考えており，均一 MDP のリグレットの下界よりも $\sqrt{H}$ 倍小さいリグレットの下界となる．

*6 Mao et al. [6] は報酬関数および遷移確率関数を用いて非定常性指標を定義しているが，Wei and Luo [18] が導入したこちらの定義の方が弱いため，これを導入する．Mao et al. [6] の定義に比べ，H 倍大きくなることに注意.

*7 Jin et al. [7] は均一 MDP の下界が $\widetilde{\Omega}\left(S^2AH^2/\varepsilon^2\right)$ であると証明した．均一 MDP は不均一 MDP の特殊な場合であるため，この下界は不均一 MDP のでも成立するが，最適ではない．Ménard et al. [14] と Zhang et al. [21] はこの下界が最適ではなく，不均一 MDP では $\widetilde{\Omega}\left(S^2AH^3/\varepsilon^2\right)$ の下界が成立すると主張している．

表 4: 無報酬強化学習における事前知識と性能保証の比較.

	事前知識	$\mathbb{E}[\tau]$	文献
下界	-	$\widetilde{\Omega}\left(\frac{S^2AH^3}{\varepsilon^2}\right)$	[7]
上界	-	$\widetilde{\mathcal{O}}\left(\frac{S^2AH^3}{\varepsilon^2}\right)$	[21]
	$\sum_{h=1}^{H} r_h \le 1$	$\widetilde{\mathcal{O}}\left(\frac{S^2AH}{\varepsilon^2}\right)$	[21]
	N_{rew}	$\widetilde{\mathcal{O}}\left(\frac{SAH^5}{\varepsilon^2}\right)$	[22]
	c_{gap}	$\widetilde{\mathcal{O}}\left(\frac{S^2AH^3}{c_{\mathrm{gap}}\varepsilon}+\frac{S^3AH^4}{\varepsilon}\right)$	[23]

また，Wu et al. [23] は最適な行動とそれ以外の行動の価値の差が一定値以上であるという問題設定を導入した．つまり，$\mathbf{r}$ における最適価値の差を $\mathrm{gap}(s,a,h,\mathbf{r}) := V^{\pi_*}_{\mathbf{r},h}(s) - Q^{\pi_*}_{\mathbf{r},h}(s,a) \ge 0$ と定義して，$0 < c_{\mathrm{gap}} \le \min_{(s,a,h,\mathbf{r})} \mathrm{gap}(s,a,h,\mathbf{r})$ が成立すると仮定する．そして，事前知識として c_{gap} が与えられているとき，上界が $\widetilde{\mathcal{O}}\left(S^2AH^3/(c_{\mathrm{gap}}\varepsilon) + S^3AH^4/\varepsilon\right)$ となるアルゴリズムを提案しており，これは事前知識がない場合の上界よりも $1/\varepsilon$ に対する性能が改善されている [23].

無報酬強化学習における上記の結果を表 4にまとめた.

4.4 制約付き MDP（Constrained MDP）

制約付き MDP は六つの要素 $(\mathbf{S}, \mathbf{A}, \mathbf{r}, \mathbf{P}, \mu, H, \mathbf{c})$ で定義される．ただし，$c_h \in [0,1]^{\mathbf{S}\times\mathbf{A}}$ を時刻 h でのコスト関数として，$\mathbf{c} := \{c_h\}_{h=1}^{H}$ とする．制約付き MDP では，制約の閾値 $\alpha \in [0, H]$ について，$\rho^{\pi}_{\mathbf{c}} < \alpha$ のもとで $\rho^{\pi}_{\mathbf{r}}$ を最大化させるような π を学習することを目標とする．Efroni et al. [3] をもとに，次の制約付き MDP についての性能指標を定義する.

定義 9 (制約違反ありのリグレット)．ある正の定数 $\delta \in (0,1)$ を固定する．このとき報酬についての（擬）リグレットを $R_{\mathbf{r}}(K) := \sum_{k=1}^{K}(\rho^*_{\mathbf{r}} - \rho^{\pi_k}_{\mathbf{r}})$，制約違反に関する（擬）リグレットを $R_{\mathbf{c}}(K) := \sum_{k=1}^{K}(\rho^{\pi_k}_{\mathbf{c}} - \alpha)$ と定義し，アルゴリズムが任意の MDP に対して以下の不等式を満たす多項式 $F^{\mathrm{rew}}_{\mathrm{CHPR}}$ と多項式 $F^{\mathrm{cost}}_{\mathrm{CHPR}}$ を持つ場合，それぞれの多項式を性能の指標とする：確率 $1-\delta$ で

$$R_{\mathbf{r}}(K) \le F^{\mathrm{rew}}_{\mathrm{CHPR}}\left(S, A, H, K, \ln\frac{1}{\delta}\right)$$
$$\text{かつ } R_{\mathbf{c}}(K) \le F^{\mathrm{cost}}_{\mathrm{CHPR}}\left(S, A, H, K, \ln\frac{1}{\delta}\right).$$

制約付き MDP での達成可能なリグレットの下界についての理論解析の結果はあまり得られていないが，$F^{\mathrm{rew}}_{\mathrm{CHPR}}$ の上界が $\widetilde{\mathcal{O}}(H^2\sqrt{S^3AK})$ かつ $F^{\mathrm{cost}}_{\mathrm{CHPR}}$ の上界が $\widetilde{\mathcal{O}}(H^2\sqrt{S^3AK})$ であるアルゴリズムが提案されている [3].

制約違反についてのリグレットを改善するため，$\rho^{\pi_{safe}}_{\mathbf{c}} = \alpha_{safe} < \alpha$ を満たす安全な方策 π_{safe} の存在を仮定した問題設定が提案されている．事前知識として α_{safe} が既知である場合には，$F^{\mathrm{rew}}_{\mathrm{CHPR}}$ の上界が $\widetilde{\mathcal{O}}\left(\frac{H^3}{\alpha-\alpha_{\mathrm{safe}}}\sqrt{S^3AK}\right)$ かつ $F^{\mathrm{cost}}_{\mathrm{CHPR}}$ の上界が $\widetilde{\mathcal{O}}(1)$ であるアルゴリズムが提案されている [24].

定義 9では制約違反を許していたが，実世界の応用では全てのエピソードで一度も制約を違反しないアルゴリズムが好まれる場合がある.[*8] そこで，Bura et al. [4] をもとに次の性能指標を定義する.

*8 定義 9の性能指標では，どこかのエピソード $k \in [K]$ で制約を違反しても $R_{\mathbf{c}}(K)$ が 0 になり得ることに注意.

表 5: 制約付き MDP における事前知識と性能保証の比較.

事前知識	$F^{\mathrm{rew}}_{\mathrm{CHPR}}$	$F^{\mathrm{cost}}_{\mathrm{CHPR}}$	文献
-	$\widetilde{\mathcal{O}}(H^2\sqrt{S^3AK})$	$\widetilde{\mathcal{O}}(H^2\sqrt{S^3AK})$	[3]
α_{safe}	$\widetilde{\mathcal{O}}\left(\frac{H^3\sqrt{S^3AK}}{\alpha-\alpha_{\mathrm{safe}}}\right)$	$\widetilde{\mathcal{O}}(1)$	[24]
$\alpha_{\mathrm{safe}}, \pi_{\mathrm{safe}}$	$\widetilde{\mathcal{O}}\left(\frac{H^3\sqrt{S^2AK}}{\alpha-\alpha_{\mathrm{safe}}}\right)$	0	[4]

表 6: 時間長非依存強化学習における性能保証の比較.　上界の行にある 3 つのアルゴリズムでは全て $\sum_{h=1}^{H} r_h \le 1$ a.s を事前知識として利用している.

	リグレット	文献
下界	$\widetilde{\Omega}\left(\sqrt{SAK}\right)$	[8]
上界	$\widetilde{O}\left(\sqrt{SAK}+S^2A\right)$	[25]
	$\widetilde{O}\left(\sqrt{S^9A^3K}\right)$	[26]
	$\widetilde{O}\left(\sqrt{\min\left\{\mathrm{Var}^{\sum}_{K}, \mathrm{Var}^{*}\,K\right\}SA}+\Gamma SA\right)$	[27]

定義 10 (制約違反なしのリグレット)．ある正の定数 $\delta \in (0,1)$ を固定する．アルゴリズムが任意の MDP に対して以下の不等式を満たす多項式 F^{*}_{CHPR} 持つ場合，その多項式を性能の指標とする：確率 $1-\delta$ で

$$R_{\mathbf{r}}(K) \le F^{*}_{\mathrm{CHPR}}\left(S, A, H, K, \ln\frac{1}{\delta}\right) \text{ かつ } \max_{k\in[K]} \rho^{\pi_k}_{\mathbf{c}} < \alpha\,.$$

定義 10の性能指標について，事前知識として π_{safe} と α_{safe} が既知であるとき，F^{*}_{CHPR} の上界が $\widetilde{\mathcal{O}}\left(\frac{H^3S}{\alpha-\alpha_{\mathrm{safe}}}\sqrt{AK}\right)$ であるアルゴリズムが提案されている [4].

制約付き MDP における上記の結果を表 5にまとめた.[*9][*10]

4.5 時間長非依存強化学習（Horizon-Free 強化学習）

時間長非依存強化学習では，アルゴリズムの性能の上界が H に対して対数の多項式的にまでしか依存しないような学習を目標とする．時間長非依存強化学習では特に F_{HPR}（定義 1）を性能の指標として扱うことが多いため，本章では F_{HPR} の上界が H について対数の多項式的に依存するアルゴリズムを“時間長非依存なアルゴリズム”と呼ぶ.

時間長非依存の目標を達成するために，ほとんどの研究では $\sum_{h=1}^{H} r_h \le 1$ a.s を仮定した MDP を考えている．以下，この仮定を事前知識として利用して時間長非依存の目標を達成した研究を紹介する．Zhang et al. [25] は F_{HPR} の上界が $\widetilde{O}(\sqrt{SAK}+S^2A)$ であるアルゴリズムを提案した．この上界の K に依存する第一項目は，$H=1$ の MDP に相当する文脈付きバンディットにおける F_{HPR} の下界 $\widetilde{\Omega}(\sqrt{SAK})$ と対数の多項式項を無視すれば一致している点で，ほぼミニマックス最適なアルゴリズムである.

Zhang et al. [25] の上界は H に対数の多項式的に依存してしまっている一方で，Zhang et al. [26] は F_{HPR} の上界が H に全く依存していないアルゴリズムを提案した．このアルゴリ

*9 Bura et al. [4] と Liu et al. [24] の結果は K が十分大きいとき，つまり $K \ge \mathrm{poly}(S,A,H)$ のときにのみ成立する．ただし，$\mathrm{poly}(S,A,H)$ は S, A, H についての多項式である.

*10 表 5では簡単のため，Bura et al. [4] の結果は定義 9の定義に従った結果に直した.

ズムの上界は $\widetilde{O}(\sqrt{S^9A^3K})$ であり，S と A に対する依存性は悪くなってしまっているが，上界が H に全く依存していない点で完全な時間長非依存なアルゴリズムである．

ここまでに導出された上界は全て K に多項式的に依存している．K に対する依存性も改善するため，Zhou et al. [27] は MDP がもつランダム性に上界が依存するアルゴリズムを提案した．Zhou et al. [27] は MDP のランダム性を表す次の二つの量を導入した．

$$\begin{aligned} \mathrm{Var}_K^{\Sigma} &:= \sum_{k=1}^{K}\sum_{h=1}^{H} \mathbb{V}\left(V_{h+1}^{\star}\right)\left(s_h^k, a_h^k, h\right) \\ \mathrm{Var}^{*} &:= \max_{\pi}\max_{s_1\in\mathbf{S}} \mathbb{E}_{\pi}\left[\sum_{h=1}^{H} \mathbb{V}\left(V_{h+1}^{\pi}\right)\left(s_h^{\pi}, a_h^{\pi}, h\right)\right] \end{aligned} \tag{1}$$

ここで，s_h^k, a_h^k はアルゴリズムが k エピソードの h ステップ目で訪れた状態行動，s_h^{π}，a_h^{π} は π に従って訪れた状態行動，そして $\mathbb{E}_{\pi}$ は π で条件付けられた期待演算子である．また，任意の $V \in \mathbb{R}^{\mathbf{S}}$ について，$\mathbb{V}(V)(s,a,h) := \sum_{s'\in\mathbf{S}} P_h(s' \mid s,a)\left(V(s')\right)^2 - \left(\sum_{s'\in\mathbf{S}} P_h(s' \mid s,a)V(s')\right)^2$ を遷移確率関数による V の分散として定義した．式 (1)で定義された二つを用いて，Zhou et al. [27] は F_{HPR} の上界が $\widetilde{\mathcal{O}}\left(\sqrt{\min\left\{\mathrm{Var}_K^{\Sigma}, \mathrm{Var}^{\star}\right\}K}SA + \Gamma SA\right)$ であるアルゴリズムを提案した．ここで，$\Gamma := \max_{h,s,a}\sum_{s'\in\mathbf{S}} \mathbb{1}\left[P_h(s' \mid s,a) > 0\right]$ は MDP における遷移確率関数の台の最大の大きさである．MDP に全くランダム性がない場合，この上界は $\widetilde{\mathcal{O}}(SA)$ となり，時間長非依存かつ K に対しても非依存なアルゴリズムになる．

時間長非依存強化学習における上記の結果を表 6にまとめた．

5 未解決問題と将来の理論研究の方向

本論文では様々な逐次意思決定の問題設定と，各問題設定でどのような事前知識が性能向上に有益となるかをまとめた．本章ではこれまでの内容を踏まえて，現在未解決の問題をいくつか提示し，将来の逐次意思決定理論の方向を三つ示す．

まず，本論文で紹介した問題設定では，ミニマックス最適なアルゴリズムが未発見のものがある．例えば4.3 節で紹介したように，無報酬強化学習では時間長非依存強化学習で利用されている事前知識を導入することで，$\mathbb{E}[\tau]$ の上界を時間長 H について改善したアルゴリズムが提案されている [21]．一方で，事前知識なしの場合にはミニマックス最適なアルゴリズムが見つかっている [21]．このような性能の良いアルゴリズムの工夫を応用して，既存の非ミニマックス最適なアルゴリズムの性能改善を試みる研究が，将来の理論研究の方向として期待される．

また、まだ解析されていない事前知識と問題設定の組み合わせも存在する．例えば4.3 節で紹介したように，無報酬強化学習では時間長非依存強化学習で利用されている事前知識を導入し，$\mathbb{E}[\tau]$ の上界が時間長 H について改善されたアルゴリズムが提案されている [21]．このように，いくつかの事前知識は他の問題設定に応用可能であり，既存の性能の下界を改善し得る事前知識と問題設定の組み合わせの発見が，将来の理論研究に期待される．

最後に，複合的な問題設定の理論解析が将来的な研究として期待される．例えば実世界の応用では非定常かつ制約付きな MDP の問題設定は容易に考えられ，実際に Ding and Lavaei [19] が理論解析を行っている．しかし，他にどのような複合問題が解けるのか，そしてどのような事前知識を導入すれば効率よく解けるのかについては，あまり研究が進んでいない．

本論文で紹介した問題設定のまとめと，事前知識による性能改善の結果のまとめが，これらの未解決問題に取り組む将来の理論研究の助けになれば幸いである．

参考文献

[1] Eyal Even-Dar, Sham. M. Kakade, and Yishay Mansour. Online Markov Decision Processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

[2] Jia Yuan Yu, Shie Mannor, and Nahum Shimkin. Markov Decision Processes with Arbitrary Reward Processes. *Mathematics of Operations Research*, 34(3): 737–757, 2009.

[3] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-Exploitation in Constrained MDPs. *arXiv*, 2020.

[4] Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. DOPE: Doubly Optimistic and Pessimistic Exploration for Safe Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2021.

[5] Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic Reinforcement Learning in Finite MDPs: Minimax Lower Bounds Revisited. In *International Conference on Algorithmic Learning Theory*, 2020.

[6] Weichao Mao, Kaiqing Zhang, Ruihao Zhu, David Simchi-Levi, and Tamer Basar. Near-Optimal Model-Free Reinforcement Learning in Non-Stationary Episodic MDPs. In *International Conference on Machine Learning*, 2021.

[7] Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-Free Exploration for Reinforcement Learning. In *International Conference on Machine Learning*, 2020.

[8] Ian Osband and Benjamin Van Roy. On Lower Bounds for Regret in Reinforcement Learning. *arXiv*, 2016.

[9] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax Regret Bounds for Reinforcement Learning. In *International Conference on Machine Learning*, 2017.

[10] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and Regret: Uniform PAC Bounds for Episodic Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2017.

[11] Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy Certificates: Towards Accountable Reinforcement Learning. In *International Conference on Machine Learning*, 2019.

[12] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-Learning Provably Efficient? In *Advances in Neural Information Processing Systems*, 2018.

[13] Nicolò Cesa-bianchi, Alex Conconi, and Claudio Gentile. On the Generalization Ability of On-Line Learning Algorithms. In *Advances in Neural Information Processing Systems*, 2001.

[14] Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, 2021.

[15] Mahsa Ghasemi, Abolfazl Hashemi, Haris Vikalo, and Ufuk Topcu. Online Learning with Implicit Exploration in Episodic Markov Decision Processes. In *American Control Conference*, 2021.

[16] Alexander Zimin and Gergely Neu. Online Learning in Episodic Markovian Decision Processes by Relative Entropy Policy Search. In *Advances in Neural Information Processing Systems*, 2013.

[17] Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning Adversarial Markov Decision Processes with Bandit Feedback and Unknown Transition. In *International Conference on Machine Learning*, 2020.

[18] Chen-Yu Wei and Haipeng Luo. Non-stationary Reinforcement Learning without Prior Knowledge: An Optimal Black-box Approach. In *Conference on Learning Theory*, 2021.

[19] Yuhao Ding and Javad Lavaei. Provably Efficient Primal-Dual Reinforcement Learning for CMDPs with Non-stationary Objectives and Constraints. In *AAAI Conference on Artificial Intelligence*, 2023.

[20] Yueyang Liu, Benjamin Van Roy, and Kuang Xu. A Definition of Non-Stationary Bandits. *arXiv*, 2023.

[21] Zihan Zhang, Simon Du, and Xiangyang Ji. Near Optimal Reward-Free Reinforcement Learning. In *International Conference on Machine Learning*, 2021.

[22] Xuezhou Zhang, Yuzhe Ma, and Adish Singla. Task-agnostic Exploration in Reinforcement Learning. 2020.

[23] Jingfeng Wu, Vladimir Braverman, and Lin Yang. Gap-Dependent Unsupervised Exploration for Reinforcement Learning. In *International Conference on Artificial Intelligence and Statistics*, 2022.

[24] Tao Liu, Ruida Zhou, Dileep Kalathil, Panganamala Kumar, and Chao Tian. Learning Policies with Zero or Bounded Constraint Violation for Constrained MDPs. In *Advances in Neural Information Processing Systems*, 2021.

[25] Zihan Zhang, Xiangyang Ji, and Simon Shaolei Du. Is Reinforcement Learning More Difficult Than Bandits? A Near-optimal Algorithm Escaping the Curse of Horizon. *arXiv*, 2020.

[26] Zihan Zhang, Xiangyang Ji, and Simon Shaolei Du. Horizon-Free Reinforcement Learning in Polynomial Time: the Power of Stationary Policies. In *Conference on Learning Theory*, pages 3858–3904, 2022.

[27] Runlong Zhou, Zihan Zhang, and Simon Shaolei Du. Sharp Variance-Dependent Bounds in Reinforcement Learning: Best of Both Worlds in Stochastic and Deterministic Environments. *arXiv*, 2023.