The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

最大エントロピー原理に基づく逆強化ダイナミクス学習フレームワークの構築

Construction of Inverse Reinforcement Dynamics Learning Framework based on Maximum Entropy Principle

中口 悠輝 Yuki Nakaguchi　江藤 力 Riki Eto　西岡 到 Itaru Nishioka

NEC データサイエンス研究所
NEC Data Science Research Laboratories

Recently, reinforcement learning (RL) has been showing an increasingly high performance in a variety of complex tasks of decision making and control, but RL requires quite careful engineering of reward functions to solve real tasks. Inverse reinforcement learning (IRL) is a framework to construct reward functions by learning from demonstration, but most of IRL algorithms require many accesses to the dynamics, though our access to and knowledge about the dynamics is often limited. To deal with this uncertainty of the dynamics, we propose a novel mathematical framework for constructing reward and dynamics by extending the celebrated maximum entropy framework of IRL.

1. はじめに

近年、意思決定や制御の手法を学習するフレームワークである強化学習の研究が著しく進展し、計算機の性能向上も相まって、ロボットの制御やビデオゲームの攻略、囲碁や将棋といった複雑な意思決定や制御の問題において幅広く高い性能を示すようになった。しかし、より幅広い実問題に適用するにあたっては、適切な報酬関数を指定するのが困難でしばしば意図しない振る舞いが生じてしまうのが大きな問題となっている [1]。いかに強化学習のアルゴリズムが適切に働いていても、良いパフォーマンスを得るためには緻密な報酬関数設計が必要となってしまう。

逆強化学習 (Inverse Reinforcement Learning; IRL) [2] は、強化学習の逆問題を解くことでこの人手による緻密な報酬関数設計を回避する [*1]。即ち、強化学習では与えられた報酬関数のもとで意思決定主体が良い振る舞いを模索するが、逆強化学習では、その強化学習問題のエキスパートの振る舞いのデモンストレーションを与え、それを生成した報酬関数を推測する。しかし、殆どの逆強化学習の定式化はダイナミクス（状態がいかに遷移していくか）へ幾らでもアクセスできることを仮定している一方で、現実にはしばしばダイナミクスへのアクセスは限られており、不確実な知識しか持ち合わせていない。

そこで本研究では、ダイナミクスに対する知識の不確実さに対処するため、逆強化学習にて最も主流の最大エントロピー法をダイナミクスの推測を含む形へ拡張することで、報酬関数とダイナミクスを同時に構成するフレームワークを提案する。また、逆強化学習および最大エントロピー法を簡潔にレビューする。

2. 逆強化学習

2.1 強化学習の定式化

強化学習は通常、Markov 決定過程 (Markov Decision Process; MDP) と呼ばれる、各時刻にて意思決定を行う主体 (エージェント、agent) を導入することで Markov 過程を一般化した数理モデルとして定式化される。Markov 過程において状態 (state) $s \in \mathcal{S}$ の時系列 $S_t = (s_1, \ldots, s_t)$ を生成していく Markov 的ダイナミクス $p_t(s_{t+1}|s_t)$ は現在の状態 s_t にしか依らなかったが、Markov 決定過程におけるダイナミクス $p_t(s_{t+1}, r_{t+1}|s_t, a_t)$ では、現在の状態 s_t においてエージェントが取った行動 $a_t \in \mathcal{A}$ にも依存する上、次の時刻の状態 s_{t+1} のみならず報酬 (reward) と呼ばれる実数値 $r_{t+1} \in \mathbb{R}$ を確率的に返す [*2]。通常、関数形が時刻 t に依らない斉時的 (time homogeneous) な Markov 的ダイナミクス $p = p_t$ を用いる。

強化学習の問題は、各時刻 t において今までの状態の時系列 S_t と行動の時系列 $A_{t-1} = (a_1, \ldots, a_{t-1})$ の実現値に基づいて行動 a_t を選択する確率分布 (方策、policy) $\pi(a_t|S_t, A_{t-1})$ のうち、報酬の時間和 $R \equiv \sum_{t'} \gamma^{t'} r_{t'+1}$ (利得、return) の期待値を最大化する最適方策を求める問題として定式化される。ここで、割引因子 (discount factor) $\gamma \in [0, 1]$ は将来の報酬の価値を現在価値に割り引く係数である。利得 R の期待値は報酬の期待値である報酬関数 (reward function) $r(s, a) \equiv E_p[r|s, a]$ にしか依らないため、通常は報酬の分布は考えずに報酬関数のみ取り扱う。その場合、次の時刻の状態 s' に対する周辺分布 $p(s'|s, a) = \int p(s', r|s, a)\, dr$ を単にダイナミクスという。斉時 Markov 的なダイナミクス $p(s'|s, a)$ に対しては最適な斉時 Markov 的な方策 $\pi(a|s)$ が存在するため、通常は斉時 Markov 的な方策のみ考える [*3]。結局、エージェントの軌跡 (trajectory) $\zeta = (S, A)$ はダイナミクス $p(s'|s, a)$ と方策 $\pi(a|s)$ の 2 つの斉時 Markov 的分布に従って生成される。

2.2 逆強化学習の定式化とその流派

前述のとおり、逆強化学習においては、強化学習問題のエキスパートの軌跡のサンプル（デモンストレーション、demonstration）$D = \{\zeta_n\}_n$ を与え、この D を生成した報酬関数 $r(s, a)$ を推測するという逆問題を解く。しかし、一般に一つのデモンストレーションは無限に多くの報酬関数によって説明されうるため、逆強化学習はこのままでは不良設定問題である。どのように良設定問題として定式化するかに応じて、逆強化学習は主に 3 つの流派、最大マージン法 [4–6]、最大エントロピー法 [7–9]、Bayes 的アプローチ [10, 11] に分かれる。

歴史的に最も初期のアプローチである最大マージン法では、

連絡先: y-nakaguchi@cj.jp.nec.com

*1 最適制御問題における損失関数の推定に用いるという文脈では、逆最適制御 (Inverse Optimal Control; IOC) とも呼ばれる [3]。

*2 状態集合 $\mathcal{S}$ および行動集合 $\mathcal{A}$ は離散でも連続でも良いが、時間はふつう離散時間 $t = 1, 2, \ldots$ を考える。

*3 Markov 的ダイナミクス $p_t(s_{t+1}|s_t, a_t)$ に対しては最適な Markov 方策 $\pi_t(a_t|s_t)$ が存在するが、一般のダイナミクス $p(s_{t+1}|S_t, A_t)$ では Markov 方策が最適となるとは限らない。

報酬関数に対する何らかの目的関数（その報酬関数のもとでのエキスパートと非エキスパートとの利得の差など）を設定し、それを最大化する報酬関数を選び出す。しかし、最大マージン法はデモンストレーションの乱雑さに対して頑健ではないという欠点がある。現実にはエキスパートといえども唯一の最適な行動を取っておらず最適に近い行動を取っており、行動は乱雑に揺らいでいる。あるいは仮にエキスパートが常に最適な行動を取っていたとしても、エキスパートが意思決定に用いている特徴量の全てを我々が観測できるわけではない場合、その軌跡は我々には乱雑に見えることとなる。

第 3 章で後述する最大エントロピー法では、最大エントロピー原理 [12] に基づいてエキスパートの乱雑な行動を確率的にモデル化する。Bayes 的アプローチでは、エキスパートを確率的にモデル化するのみならず、さらに報酬関数に対する事前分布を設定し、エキスパートの確率モデルから従う尤度関数によって Bayes 更新することで報酬関数に対する事後分布を得て、その事後平均 (posterior mean) や最大事後確率 (MAP) 推定を報酬関数の推定値とする。しかし、関数に対する分布の自由度が巨大なためそのままでは適切な更新に大きなサンプルが必要であること、事前分布や方策のモデルの任意性、Markov 連鎖モンテカルロなどのコストが高い計算が必要となるなどの欠点がある。MAP 推定値を取る定式化の場合、最大エントロピー法を特殊なケースとして含むと解釈できる [13]。

2.3 逆強化学習の諸課題

一方で、逆強化学習は幅広い現実の問題に適用するにあたって種々の課題を抱えている。

適切な報酬関数のクラスの設定

多くの研究では報酬関数が特徴量 $\phi(s,a)$ に関して線形な形 $r(s,a)=\theta^T\phi(s,a)$ に限っており、真の報酬関数が非線形の場合に表現できない。Gauss 過程を用いた Gauss 過程逆強化学習 (GP IRL) [14] や深層学習を用いた逆強化学習 [15, 16] などのように、表現力が高い関数近似器を用いれば真の報酬関数をよく近似できるようになるものの、バイアス-バリアンスのトレードオフによってより大きなサンプルが必要となる。また、仮に線形関数でよく近似できるとしても、特徴量 $\phi(s,a)$ をどう設計すべきかという問題が残る。Feature construction for IRL (FIRL) [17] のように特徴量設計を取り扱う手法は数少ない。

計算複雑性

典型的な逆強化学習のアルゴリズムでは、報酬関数を収束するまで逐次的に更新し、その更新において毎回、現在の報酬関数のもとで非常に計算複雑性が高い（計算量が大きい）何らかの計画問題（強化学習、動的計画法、モデル予測制御など）を解く。相対エントロピー逆強化学習 (Relative Entropy IRL; RE IRL) [18] では軌跡のサンプリングを活用し計画問題を回避するが、そのぶん多数のサンプルを必要とする。線形可解 Markov 過程 (Linearly solvable MDP; LMDP) に基づく逆最適制御 [19] では、計画問題が解析的に解けるような巧妙な問題設定に制限することで計算複雑性を回避している。誘導コスト学習 (Guided Cost Learning; GCL) [16] では、報酬関数の更新において計画問題を完全には解かず、報酬関数と同時に方策も保持して方策の小さな更新に留めることで計算複雑性を減らしている。

未知のダイナミクスとサンプル複雑性

報酬の更新における計画問題においては、ダイナミクスが既知の場合はそれを用いた動的計画法やモデル予測制御などを解けばよいが、ダイナミクスが未知の場合は強化学習などダイナミクスからのサンプリングが必要となり、デモンストレーションのみならずそのダイナミクスからのサンプル複雑性（学習に必要なサンプルのサイズ）も問題となる。RE IRL では報酬関数の更新においてはサンプリングを必要としないが、そのぶんはじめに十分なサンプルを用意する必要がある。GCL では計画問題を解ききらずに現在の方策によるサンプリングに留めることで、ダイナミクスからのサンプル複雑性も下げている。

現実には、ダイナミクスについての限られた知識や推測しか無く、ダイナミクスからのサンプリングもできない場合も考えられる [20]。そのような場合、デモンストレーションが持つダイナミクスの情報をうまく活用する技術が必要である。本研究では、ダイナミクスについての限られた知識や推測を活用しながらデモンストレーションが含むダイナミクスの情報をうまく活用する定式化として、最大エントロピー法のフレームワークを拡張し、報酬関数とダイナミクスを同時に構築する新しいフレームワークを提案する。

3. 最大エントロピー逆強化学習

3.1 最大エントロピー原理

最大エントロピー原理は、確率変数 x について何らかの特徴量ベクトル $\phi(x)$ の期待値 $E_p[\phi(x)]\equiv\sum_x p(x)\phi(x)$ の値のみ $\langle\phi\rangle$ であると観測できているがその分布 $p(x)$ 自体が未知である状況において、分布 $p(x)$ を構成する標準的な方法を与える。即ち、特徴量 $\phi(x)$ の期待値 $E_p[\phi(x)]$ を観測値 $\langle\phi\rangle$ に固定する拘束条件のもと、エントロピー $H[p]\equiv E_p[-\log p(x)]=-\sum_x p(x)\log p(x)$ を最大にする分布

$$p^*=\arg\max_{p\in\Delta} H[p] \qquad \text{s.t.} \qquad E_p[\phi(x)]=\langle\phi\rangle$$

を選ぶ。ここで、$\Delta\equiv\{p\mid p(x)\geq 0,\sum_x p(x)=1\}$ は規格化を満たす確率分布全体（確率単体）である。エントロピー $H[p]$ が凹関数で実行可能領域も凸集合なため、この問題は凸最適化問題である。よって、緩い条件下で強双対性が成り立ち、その解は Lagrange 関数 $L[p,\theta,\lambda]$ を

$$L\equiv H[p]+\theta^T(E_p[\phi(x)]-\langle\phi\rangle)+\lambda\Big(\sum_x p(x)-1\Big) \quad (1)$$

とする Lagrange 双対問題 $\min_{\theta,\lambda}\max_p L[p,\theta,\lambda]$ を解くことにより、指数型分布

$$p_\theta(x)=\frac{e^{\theta^T\phi(x)}}{Z_\theta} \quad (2)$$

として得られる。$H[p]$ が狭義凹関数なので解は一意的である。規格化因子 $Z_\theta=\sum_x e^{\theta^T\phi(x)}$ は Bayes 統計と同様、統計力学の用語を流用して分配関数 (partition function) と呼ばれる。

Lagrange 未定乗数 θ はエントロピー $H[p_\theta]=E_{p_\theta}\left[-\theta^T\phi(x)+\log Z_\theta\right]=-\log(e^{\theta^T\langle\phi\rangle}/Z_\theta)$ を最小にするように、特徴量の期待値のマッチング

$$0=-\frac{\partial H[p_\theta]}{\partial\theta}=\langle\phi\rangle-E_{p_\theta}[\phi(x)] \quad (3)$$

により定まる。ここで、指数型分布の恒等式 $E_{p_\theta}[\phi(x)]=-\frac{\partial}{\partial\theta}\log Z_\theta$ を用いた。とくに、観測値 $\langle\phi\rangle$ が何らかの観測サンプル X のサンプル平均 $\langle\phi\rangle=\frac{1}{|X|}\sum_{x\in X}\phi(x)$ として与えられている場合には、エントロピー $H[p_\theta]$ は単に負の対数尤度 $-\log p(X|\theta)$ であるため、この特徴量マッチングによる θ の決定は指数型分布 (2) に対する最尤法と解釈できる。

直感的には、エントロピーはその変数のもつ不確実さ、その変数の値に対する我々の無知の度合いを表すと解釈できるため、既知の知識と整合する確率分布のうち、未知の部分に対しては最も謙虚な、最も無情報で “無難” な分布を選ぶということであ

る。事実、最大エントロピー分布は、最悪時予測対数誤差

$$\max_{\tilde{p}\in\Delta}\left[-\sum_x \tilde{p}(x)\log p(x)\right] \quad \text{s.t.} \quad E_{\tilde{p}}[\phi(x)]=\langle\phi\rangle$$

を最小化するという点で最も “無難” である [21]。

3.2 ダイナミクスが決定論的な場合

Ziebart はエキスパートの行動の確率モデルの構築においてこの最大エントロピー原理を適用し、最大エントロピー逆強化学習 (Maximum Entropy IRL; ME IRL) を提案した。最初の論文 [7] では、ダイナミクスが決定論的な場合、即ち、次の時刻の状態 s' が現在の状態 s と現在の行動 a の関数として表せる場合 $s'=f(s,a)$ を取り扱っている。その場合、軌跡 $\zeta=(S,A)$ は行動の系列 $A=A_T=(a_1,\ldots,a_T)$ によって一意的に ζ_A と定めることができるので、取り扱われるべき確率変数は本質的に行動の系列 A のみである。

軌跡に対する特徴量 $\phi(\zeta)$ の期待値がデモンストレーション D のサンプル平均 $\langle\phi\rangle\equiv\frac{1}{|D|}\sum_{\zeta\in D}\phi(\zeta)$ と一致するという拘束条件を取り、最大エントロピー原理に基づいてエントロピー $H[\pi]=-\sum_A\pi(A)\log\pi(A)$ を最大化する分布

$$\pi^*=\arg\max_{\pi\in\Delta} H[\pi] \qquad \text{s.t.} \qquad E_\pi[\phi(\zeta_A)]=\langle\phi\rangle$$

を採用することで、指数型分布 $\pi_\theta(A)=e^{\theta^T\phi(\zeta_A)}/Z_\theta$ を得る。さらに Ziebart は、指数 $r_\theta(\zeta)\equiv\theta^T\phi(\zeta)$ をこの軌跡 ζ に対する利得 $\sum_t r(s_t,a_t)$ と解釈することを提案した。実際、この解釈のもとで、この分布による利得の期待値 $E_{\pi_\theta}[r_\theta]$ は観測値 $\langle r_\theta\rangle=\theta^T\langle\phi\rangle$ に一致する。この提案の数理的な正当性については、後に [22] で与えられた。

3.3 非線形な報酬関数への一般化

未定乗数としてパラメータ θ を導入するこの手法では線形な報酬関数しか構成できないが、拘束条件を取り込んだ Lagrange 関数 (1) $L[\pi,\theta]=H[\pi]+E_\pi[r_\theta(\zeta_A)]-\langle r_\theta\rangle$ から出発することで、報酬関数 r_θ をパラメータ θ に関して非線形の関数へと一般化できる [15, 16]。解は全く同様に $\pi_\theta(A)=e^{r_\theta(\zeta_A)}/Z_\theta$ として得られ、パラメータ θ はエントロピー $H[\pi_\theta]=-\log(e^{\langle r_\theta\rangle}/Z_\theta)$（あるいは負の対数尤度）を最小化するように

$$0=-\frac{\partial H[p_\theta]}{\partial\theta}=\frac{\partial\langle r_\theta\rangle}{\partial\theta}-E_{\pi_\theta}\left[\frac{\partial r_\theta(\zeta_A)}{\partial\theta}\right] \tag{4}$$

により定まる。これは特徴量マッチング (3) の一般化である。

3.4 一般のダイナミクスの場合

より一般の確率的なダイナミクスへと最大エントロピー法を一般化するためには、行動の系列 A のみならず状態の系列 S も確率変数として扱う必要があるが、エントロピーの定義をどう一般化するかは非自明である。いま、ダイナミクスによる状態の系列 S の確率モデルは構築しないため、両者の不確実さである同時エントロピー (joint entropy) $H[S,A]\equiv-\sum_{S,A}p(S,A)\log p(S,A)$ を最大化するのは不自然である。状態の系列 S のエントロピーは適切に除外し、行動の系列 A のエントロピーのみ用いたい。

しかし、同時分布の分解 $p(S,A)=p(A|S)p(S)$ から従うエントロピーの分解 $H[S,A]=H[A|S]+H[S]$ において、第一項の条件付エントロピー $H[A|S]\equiv-\sum_{S,A}p(S,A)\log p(A|S)$ のみ用いると、条件付分布 $p(A|S)=\prod_t p(a_t|S,A_{t-1})$ では行動 a_t の分布が未来の状態にも条件付けられており因果律と整合せず、条件付エントロピー $H[A|S]=\sum_t H[a_t|S,A_{t-1}]$ も因果律と整合せず不適切である。

一般に、ある時系列 A を別の時系列 S によって時々刻々と条件付けたい場合、因果律と整合するように条件付けた因果的条件付分布 (causally conditioned probability)

$$p(A||S)\equiv\prod_t p(a_t|S_t,A_{t-1})$$

を用いるのが適切である。このとき同時分布は $p(S,A)=p(A||S)p(S||A_{T-1})$ と 2 つの因果的条件付分布の積に分解でき、それに応じて同時エントロピーは 2 つの因果的エントロピー (causal entropy)

$$H[A||S]\equiv-\sum_{S,A}p(S,A)\log p(A||S)=\sum_t H[a_t|S_t,A_{t-1}]$$

の和 $H[S,A]=H[A||S]+H[S||A_{T-1}]$ に分解できる。Ziebart はこの因果的エントロピーを採用することで、ダイナミクスが確率的な場合へ最大エントロピー逆強化学習を一般化した [8, 9]。

因果的エントロピー $H[A||S]$ を最大化する因果的条件付分布 $\pi(A||S)$ は、その自由度を時間に関して分解 $\pi(A||S)=\prod_t\pi(a_t|S_t,A_{t-1})$ し、各 $\pi(a_t|S_t,A_{t-1})$ について最大化することで求めることができる。Lagrange 関数 (1)

$$L[\pi,\theta]=H[A||S]+E_\pi[r_\theta(S,A)]-\langle r_\theta\rangle \tag{5}$$

を変分し解は $\pi_\theta(a_t|S_t,A_{t-1})\propto e^{q_\theta^{(t)}(S_t,A_t)}$ と求まる。指数 $q_\theta^{(t)}(S_t,A_t)\equiv E_\pi\left[r_\theta(S,A)-\log p(A_{t+1:T}||S_{t+1:T})|S_t,A_t\right]$ は初項 $q_\theta^{(T)}(S,A)=r_\theta(S,A)$ および漸化式 $q_\theta^{(t)}(S_t,A_t)=E_{s_t}[\log\sum_{a_{t+1}}e^{q_\theta^{(t+1)}(S_{t+1},A_{t+1})}]$ によって再帰的に求まり、パラメータ θ は因果的エントロピー $H[A||S]=-\log(e^{\langle r_\theta\rangle}/e^{q_\theta^{(0)}})$ を最小化するように再び特徴量マッチング (4) によって定まる。この特徴量マッチングによる θ の決定は因果的尤度 (causal likelihood)

$$\sum_{S,A}\pi_e(A||S)p(S||A_{T-1})\log\pi_\theta(A||S)\simeq\sum_{\zeta\in D}\log\pi_\theta(A||S)$$

の最大化と解釈できる。ここで、$\pi_e(A||S)$ はエキスパートのデモンストレーション D を生成した真の方策である。

さらに、利得が報酬の時間和で書けること $r(S,A)=\sum_t r(s_t,a_t)$ およびダイナミクスの斉時 Markov 性を仮定すると、$Q^{(t)}(S_T,A_T)\equiv q^{(t)}(S_T,A_T)-\sum_{t'<t}r(s_{t'},a_{t'})$ が (s_t,a_t) のみにしか依存しないことを帰納的に示すことができ、その結果として方策も斉時 Markov 的 $\pi(a_t|s_t)\propto e^{Q(s_t,a_t)}$ であると分かる。規格化 $V_\theta(s)\equiv\log\sum_a e^{Q_\theta(s,a)}$ も導入すると、解は

$$\pi_\theta(a|s)=\frac{e^{Q_\theta(s,a)}}{e^{V_\theta(s)}} \tag{6}$$

と書ける。$q^{(t)}$ の漸化式は Q に対するソフト Bellman 方程式

$$Q_\theta(s,a)=r_\theta(s,a)+\gamma E_p[V_\theta(s')|s,a] \tag{7}$$

となる。これは、強化学習における価値関数 Q,V に対する最適 Bellman 方程式の max を softmax にしたものである [*4][*5]。事実、目的関数 (5) の方策 π での最適化は、エントロピー正則化つきの利得の期待値 $E_\pi[r_\theta(S,A)]+H[A||S]$ を最大化する計画問題と解釈できる。このように利得にエントロピー正則化項を入れて一般化した強化学習の定式化は最大エントロピー強化学習

---

*4 割引因子 γ の導入は、因果的エントロピーを一般化した割引因果的エントロピー $H_\gamma[A||S]\equiv\sum_t\gamma^t H[a_t|S_t,A_{t-1}]$ および割引利得 $r(S,A)=\sum_t\gamma^t r(s_t,a_t)$ を用いることで実現できる [23]。

*5 Bellman 方程式と同様に、一様ノルム $\|Q\|_\infty=\sup_{s,a}Q(s,a)$ に関して有界な Q 関数全体がなす Banach 空間において、ソフト Bellman 演算子 $T:Q(s,a)\mapsto r(s,a)+\gamma E_{s'}[\log\sum_{a'}e^{Q(s',a')}|s,a]$ も γ-縮小写像であり、Banach の不動点定理により任意の Q に対し T^nQ は唯一存在する不動点 $Q_*=TQ_*$ へ指数的に収束する [23]。

といい、従来の強化学習に比べ様々な利点があることが知られ [24]、また価値関数ベースの学習と方策ベースの学習を統一するという点で理論的にも注目を集めている [25, 26]。報酬関数のスケーリング $r \to r/\alpha$ によりエントロピー正則化項の効果は α 倍でき、ゼロ温度 $\alpha \to 0$ で消去できる。このとき softmax は max に帰着し、漸化式 (7) は最適 Bellman 方程式に帰着する。方策も決定論的な最適方策 $\pi(a|s) = \delta(a - \arg\max_a Q(s,a))$ となり、従来の強化学習に帰着する。

4. 最大エントロピー逆強化ダイナミクス学習

本研究では、限られた知識のもとで方策 $\pi(A||S)$ とダイナミクス $p(S||A_{T-1})$ の両者に関して同時エントロピー $H[S,A] = H[A||S] + H[S||A_{T-1}]$ を最大化することで報酬とダイナミクスを同時推定するフレームワークを提案する。即ち、

$$\pi^*, p^* = \arg\max_{\pi,p\in\Delta} H[S,A] \quad \text{s.t.} \quad E_{\pi,p}[\phi(S,A)] = \langle\phi\rangle$$

を解く。紙面の都合上、導出の詳細は割愛する。ダイナミクスの推定では、現在の変数のみに依る “即時” 特徴量 $\phi(s_t, a_t)$ のみならず、次の時刻の状態にも依る “遷移” 特徴量 $\phi(s_t, a_t, s_{t+1})$ も用いる必要がある。軌跡の特徴量は両者の時間和

$$\phi(S,A) = \sum_t \begin{pmatrix} \phi(s_t,a_t) \\ \phi(s_t,a_t,s_{t+1}) \end{pmatrix}$$

とし、それぞれに対する未定乗数を θ, λ とする。両者への分解は一意では無いが、結果は分解の仕方に依らない。

方策 $\pi(a|s)$ の解の形は (6) のままだが、ソフト Bellman 方程式にエントロピーボーナスが導入される：

$$Q(s,a) = r(s,a) + E_p[V(s')|s,a] + H[s'|s,a]$$

この表式は [8] の (6.21) と同じだが、報酬関数において “遷移” 特徴量が期待値で現れる点が異なる：

$$r_{\theta,\lambda}(s,a) = \theta^T\phi(s,a) + \lambda^T E_{s'}[\phi(s,a,s')|s,a]$$

即ち、この定式化では報酬関数の特徴量はダイナミクスに応じて構成されることになる。ダイナミクスの解も同様の分布

$$p(s'|s,a) = \frac{e^{-E(s,a,s')}}{e^{-F(s,a)}}$$

を取る。ここで E および $F(s,a) = \log\sum_{s'} e^{-E(s,a,s')}$ はそれぞれ Q, V の対応物であり、ダイナミクス推定だけの [27] とは異なってエントロピーボーナス付きのソフト Bellman 方程式

$$E(s,a,s') = e_\lambda(s,a,s') + E_{a'}[F(s',a')|s'] + H[a'|s']$$

をみたす。ここで、報酬関数の対応物である $e_\lambda(s,a,s') \equiv \lambda^T\phi(s,a,s')$ は “即時” 特徴量に依らず、“遷移” 特徴量のみで定まる。エントロピーボーナス項は、お互いの関数を用いて

$$H[s'|s,a] = E_{s'}[E(s,a,s')|s,a] - F(s,a)$$

$$H[a'|s'] = -E_{a'}[Q(s',a')|s'] + V(s')$$

と書くことができる。

この定式化によって、方策とダイナミクスを同時に構成することができ、真のダイナミクスへアクセスせずに推定中のダイナミクスを用いて報酬を推定できる。ダイナミクスの推定が改善していくことで報酬の推定が改善し、報酬の推定が改善していくことでダイナミクスの推定が改善するという定式化となっており、別々に推定するよりも良い性能が期待される。

5. 今後の展望

このフレームワークに基づいてデモンストレーションの持つダイナミクスの情報を活用するアルゴリズムを構成し、最大エントロピー法に基づかないダイナミクス推定による既存手法 [20] と比較を行う。また、現実にはしばしばダイナミクスの推定モデルが手元にあるため、その事前知識を Bayes 的に取り入れてダイナミクスの推定モデルを更新する枠組みへの拡張を目指す。

参考文献

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety, arxiv:1606.06565. 2016.
[2] Stuart J. Russell. Learning agents for uncertain environments (extended abstract). In *COLT*, pages 101–103, 1998.
[3] Rudolf Emil Kalman. When is a linear control system optimal? *Journal of Basic Engineering*, 86(1):51–60, 1964.
[4] Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *ICML*, pages 663–670, 2000.
[5] Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *ICML*, 2004.
[6] Nathan D. Ratliff, J. Andrew Bagnell, and Martin Zinkevich. Maximum margin planning. In *ICML*, pages 729–736, 2006.
[7] Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, pages 1433–1438, 2008.
[8] Brian D Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. PhD thesis, 2010.
[9] Brian D. Ziebart, J. Andrew Bagnell, and Anind K. Dey. Modeling interaction via the principle of maximum causal entropy. In *ICML*, pages 1255–1262, 2010.
[10] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. In *IJCAI*, pages 2586–2591, 2007.
[11] Manuel Lopes, Francisco S. Melo, and Luis Montesano. Active learning for reward estimation in inverse reinforcement learning. In *ECML PKDD*, pages 31–46, 2009.
[12] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.
[13] Jaedeug Choi and Kee-Eung Kim. MAP inference for bayesian inverse reinforcement learning. In *NIPS*, pages 1989–1997, 2011.
[14] Sergey Levine, Zoran Popovic, and Vladlen Koltun. Nonlinear inverse reinforcement learning with gaussian processes. In *NIPS*, pages 19–27, 2011.
[15] Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. Maximum entropy deep inverse reinforcement learning, arxiv:1507.04888. 2015.
[16] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *ICML*, pages 49–58, 2016.
[17] Sergey Levine, Zoran Popovic, and Vladlen Koltun. Feature construction for inverse reinforcement learning. In *NIPS*, pages 1342–1350, 2010.
[18] Abdeslam Boularias, Jens Kober, and Jan Peters. Relative entropy inverse reinforcement learning. In *AISTATS*, pages 182–189, 2011.
[19] Krishnamurthy Dvijotham and Emanuel Todorov. Inverse optimal control with linearly-solvable mdps. In *ICML*, pages 335–342, 2010.
[20] Michael Herman, Tobias Gindele, Jörg Wagner, Felix Schmitt, and Wolfram Burgard. Inverse reinforcement learning with simultaneous estimation of rewards and dynamics. In *AISTATS*, pages 102–110, 2016.
[21] Peter D Grünwald, A Philip Dawid, et al. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *the Annals of Statistics*, 32(4):1367–1433, 2004.
[22] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *NIPS*, pages 4565–4573, 2016.
[23] Zhengyuan Zhou, Michael Bloem, and Nicholas Bambos. Infinite time horizon maximum causal entropy inverse reinforcement learning. *IEEE Trans. Automat. Contr.*, 63(9):2787–2802, 2018.
[24] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *ICML*, pages 1352–1361, 2017.
[25] John Schulman, Pieter Abbeel, and Xi Chen. Equivalence between policy gradients and soft Q-learning, arxiv:1704.06440. 2017.
[26] Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. In *NIPS*, pages 2772–2782, 2017.
[27] Xiangli Chen and Brian D Ziebart. System identification via the principle of maximum causal entropy. *ICML Workshop on Machine Learning For System Identification*, 2013.