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

Multi-agent maximum discounted causal entropy逆強化学習による報酬推定

Estimation of agent's rewards with multi-agent maximum discounted causal entropy inverse reinforcement learning

浪越圭一　荒井幸代
Keiichi Namikoshi　Sachiyo Arai

千葉大学大学院融合理工学府
Graduate School of Science and Engineering, Chiba University

We propose a entropy-base multi-agent inverse reinforcement learning method for constructing a multi-agent simulation. By using multi-agent inverse reinforcement learning, we can estimate the agent's behavior rule and the reward reflecting the purpose of the agent. In this paper, we extend mximum discounted causal entropy to markov game environment. Experimental results showed that the proposed method can estimate valid reward at small grid world.

1. はじめに

群衆，交通流，金融など，複数の行動主体が各自の目的に従い相互作用する現実の環境は，マルチエージェント系と呼ばれる．マルチエージェント系の振舞いを再現することで，各自の目的を理解し，行動を予測する研究は古くから取り組まれており，災害誘導や交通政策の評価法として重要な研究分野である．

マルチエージェント系の再現法の一つにマルチエージェントシミュレーション (MAS) がある．MAS は，行動主体をエージェントとして扱い，エージェントの観測から行動のマッピングを行動ルールとして記述する．そのため，エージェントの意思決定過程や行動目的を比較的容易に解釈可能である．一方，MAS は行動ルールから全体の振舞いをボトムアップに再現するため，行動ルールの設計に多くの試行錯誤と妥当性の説明を要する．著者らはこの問題に対し，全体の振舞いを観測した行動ログから，各エージェントの行動ルールを推定する枠組みを提案してきた [Namikoshi 18]．しかし，エージェントの目的を理解するには，推定した行動ルールを解析する必要がある．

そこで，マルチエージェント逆強化学習によるエージェントの目的推定に着目する．マルチエージェント逆強化学習は，マルコフ決定過程をマルチエージェント系へ拡張した Markov game において，行動ログからエージェントの報酬を推定する枠組みである．報酬は一般に状態・行動の価値を表すため，推定報酬の高い状態・行動がエージェントの目的を表すといえる．つまり，マルチエージェント逆強化学習により，報酬から各エージェントのもつ目的を容易に解釈できる可能性がある．

本論文では，エントロピー最大化原理を用いた新たなマルチエージェント逆強化学習を提案する．具体的には，infinit-horizon のマルコフ決定過程を対象とする Maximum discounted causal entropy 逆強化学習を Markov game へ拡張し，その解法を示した．実験では，2 人エージェントの簡易な GridWorld を対象に，決定的な Nash 均衡解の方策から生成した行動ログから妥当な報酬が推定可能なことを示す．

連絡先: 浪越圭一，千葉大学大学院融合理工学府都市環境システムコース，千葉県千葉市稲毛区弥生町 1-33，acka2158@chiba-u.jp

2. 対象問題

Markov game(MG) を $< \mathcal{N}, \mathcal{S}, \{\mathcal{A}_n\}_{n\in\mathcal{N}}, T, \{R_n\}_{n\in\mathcal{N}} >$ の組で表す．$\mathcal{N}$ はエージェント集合 ($|\mathcal{N}| \geq 2$)，$\mathcal{S}$ は有限離散状態空間，$\mathcal{A}_n$ はエージェント n の有限離散行動空間，$T : \mathcal{S} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_{|\mathcal{N}|} \times \mathcal{S} \to [0, 1]$ は状態遷移確率，$R_n : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ はエージェント n の報酬をそれぞれ表す．また，エージェント n の行動は $a_n \in \mathcal{A}_n$，全エージェントの結合行動は $a \in \mathcal{A}$ と記す．全エージェントは状態 $s \in \mathcal{S}$ にアクセス可能と仮定し，不完全知覚は扱わない．

本論文では，報酬が未知の MG$\backslash\{R_n\}_{n\in N}$ において，全エージェントの行動ログ，すなわち可変長の軌跡集合 $\mathcal{D} = \{\{s_t, a_t\}_{t=0}^{t_d}\}_{d=1}^{D}$ から報酬 $\{R_n\}_{n\in\mathcal{N}}$ を推定する．以下，行動ログの行動主体をエキスパート，$\mathcal{D}$ をエキスパート軌跡と記す．報酬推定時，状態遷移確率 T を直接知ることはできないが，シミュレーション環境は利用可能とする．また，エキスパート軌跡を生成したエキスパート方策 $\boldsymbol{\pi}^E$ も得られないものとする．

3. 関連研究

3.1 模倣学習における逆強化学習の位置づけ

模倣学習 (Imitation Learning) とは，エキスパート軌跡からエキスパートの振舞いを再現する枠組みである．模倣学習は主に二つのアプローチに大別される．一つめの Behavioral Cloning は，エキスパート軌跡からエキスパートの振舞いを直接模倣する．そのため，実装が比較的容易な反面，エキスパート軌跡が十分に得られない場合，軌跡に含まれない状態行動への汎化性能が問題となる．二つめの逆強化学習 (IRL: Inverse reinforcement learning) は，エキスパート軌跡を生成したエキスパート方策を学習する．そのため，軌跡に含まれない状態行動においても適切な学習が期待できる．しかし，最適な報酬が複数存在する ill-posed 問題や，推定報酬から方策を計算する強化学習の計算コストが高いといった課題がある．

IRL は，エキスパート報酬の推定を目的とする Reward learning と，エキスパート方策の推定を目的とする Policy learning に分けられる．Reward learning は，エキスパート方策と他の方策とのマージンを最大化する手法 [Ng 00, Abbeel 04] や，最大エントロピー法に基づく手法 [Ziebart 10, Zhou 18] などが提案されている．一方 Policy learning は，主に敵対的学習を用いた手法 [Ho 16] などが提案されている．

3.2 マルチエージェント逆強化学習の分類

表 1 に，マルチエージェント逆強化学習 (MAIRL: Multi-agent IRL) の分類を示す．MAIRL は，定式化の目的関数と推定する報酬の構造によって分類できる．

表 1: マルチエージェント逆強化学習の分類

Reward structure	Objectives		
	Max-margin	Max-entropy	others
homogeneous	[Šošić 17]		
zerosum			[Lin 18] [Wang 18]
others	[Natarajan 10] [Reddy 12]	[Ziebart 10] [Bogert 18] [Song 18]	[Le 17]

[Šošić 17] は，swarm system において homogeneous なエージェントの報酬推定を提案している．[Lin 18, Wang 18] は，ゼロ和ゲームを対象とする MAIRL を提案している．しかしどちらの提案も，報酬に特殊な構造を仮定する必要があり，一般的な Markov game への適用は難しい．[Natarajan 10, Reddy 12] は [Ng 00] を拡張した MAIRL であり，[Natarajan 10] は中央制御器，[Reddy 12] は分散制御器を仮定しそれぞれ解いている．しかし前者は状態遷移確率を陽に必要とし，後者は Inner-Loop において Nash 均衡解を求める Nash Q-learning を用いる必要がある．[Ziebart 10, Bogert 18] は最大エントロピー法に基づいた MAIRL である．[Ziebart 10] は状態遷移が確率的な場合に有効な Maximum causal entropy IRL を定式化し，3 体の pursuit-evasion において有効性を示している．[Bogert 18] はエキスパート軌跡の観測に隠れ (Occlusuion) が生じる環境下の MAIRL を提案している．しかし前者は finit-horizon を対象とし，後者はエージェントが相互作用する状態での利得行列を必要とすることから，適用範囲が限られる．[Le 17, Song 18] はいずれも Policy learning を目的とした提案であり，エージェントの報酬を陽に推定しない．

本論文では，特徴ベクトル f と重み θ の線形和で報酬関数が表されると仮定し，最大エントロピー法に基づく推定法を提案する．[Ziebart 10, Bogert 18] とは infinit-horizon を扱える点，利得行列を必要としない点で異なる．

4. 提案法

4.1 定式化

マルコフ決定過程を対象とする Maximum discounted causal entropy IRL[Zhou 18](以下 MDCE IRL と記す) を，Markov game へ拡張した Multi-agent MDCE IRL(以下 M-MDCE IRL と記す) を定式化する．式 (1) から式 (5) に M-MDCE IRL の定義を示す．

$$\max_{\boldsymbol{\pi}_t, t\geq 0} \sum_{n\in N} H_{\pi_{t,n},\pi^E_{-n}}(\pi_{t,n}) \tag{1}$$

$$\text{s.t.} \quad \overline{f}_{n,\boldsymbol{\pi}^E} = \overline{f}_{n,\pi_{t,n},\pi^E_{-n}} \quad \forall n\in\mathcal{N}, t\geq 0 \tag{2}$$

$$\pi_{t,n}(a_n|s) \geq 0 \quad \forall a_n\in\mathcal{A}_n, s\in\mathcal{S}, n\in\mathcal{N}, t\geq 0 \tag{3}$$

$$\sum_{a_n\in\mathcal{A}_n} \pi_{t,n}(a_n|s) = 1 \quad \forall s\in\mathcal{S}, n\in\mathcal{N}, t\geq 0 \tag{4}$$

$$\pi_{t,n}(a_n|s) = \pi_{t',n}(a_n|s) \quad \forall s\in\mathcal{S}, a_n\in\mathcal{A}_n, n\in\mathcal{N}, t,t'\geq 0 \tag{5}$$

ここで，式 (1) はエージェント i の方策に対するエントロピーであり式 (6) で定義される．式 (2) は式 (7) の特徴期待ベクトルを一致させる制約，式 (3) から式 (5) は方策に関する制約を表す．また，$f_n : \mathcal{S}\times\mathcal{A}\to\mathbb{R}^k$ はエージェント n の特徴ベクトル，$\overline{f}_n$ は特徴期待ベクトルである．

$$H_{\pi_{t,n},\pi^E_{-n}}(\pi_{t,n}) = \mathbb{E}\left[\sum_{t=0}^{\infty} -\gamma^t \log \pi_{t,n}\left(A_{t,n}|S_t\right)\right] \tag{6}$$

$$\overline{f}_{n,\pi_n,\pi_{-n}} = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}\left[f_n\left(S_t, A_t\right)\right] \tag{7}$$

$$\boldsymbol{\pi}\left(A_t|S_t\right) = \prod_{n\in\mathcal{N}} \pi_n\left(A_{t,n}|S_t\right) \tag{8}$$

4.2 Single-agent 系への分解と解法

M-MDCE 問題の解法は大きく 2 つ考えられる．一つめは，エントロピーに基づくマルチエージェント強化学習を Inner-Loop に用いることで，M-MDCE を Multi-agent 系のまま扱う方法である．しかし，マルチエージェント強化学習は状態行動空間の爆発や同時学習問題を扱う必要があり，方策が適切に学習できない可能性がある．

二つめは，各エージェントごとの目的関数へ M-MDCE を分解し，Single-agent 系として扱う方法である．具体的には，以下の M-MDCE のラグランジュ緩和問題を，各エージェントごとに解く．

$$\max_{\boldsymbol{\pi}_t, t\geq 0} \sum_{n\in N} H_{\pi_{t,n},\pi^E_{-n}}(\pi_{t,n}) + \theta_n(\overline{f}_{n,\boldsymbol{\pi}^E} - \overline{f}_{n,\pi_n,\pi^E_{-n}})$$

$$\text{s.t.} \quad \pi_{t,n}(a_n|s) \geq 0 \quad \forall a_n\in\mathcal{A}_n, s\in\mathcal{S}, n\in\mathcal{N}, t\geq 0$$

$$\sum_{a_n\in\mathcal{A}_n} \pi_{t,n}(a_n|s) = 1 \quad \forall s\in\mathcal{S}, n\in\mathcal{N}, t\geq 0$$

$$\pi_{t,n}(a_n|s) = \pi_{t',n}(a_n|s) \quad \forall s\in\mathcal{S}, a_n\in\mathcal{A}_n, n\in\mathcal{N}, t,t'\geq 0$$

Single-agent 系へ分解した場合，MDCE IRL[Zhou 18] と同様に各エージェントごとに MDCE IRL を解けばよく，Multi-agent 系における問題は生じない．しかし，対象問題の仮定から，報酬を推定するエージェント n 以外のエキスパート方策 π^E_{-n} は得られない．

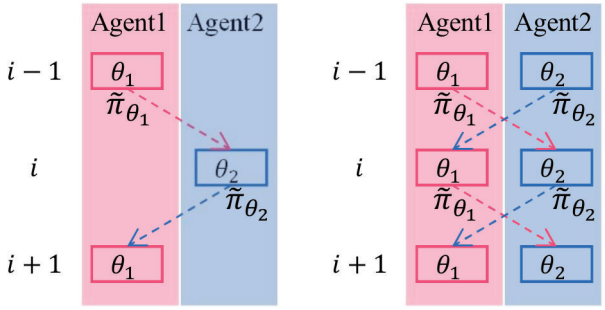
そこで本提案では，エキスパート方策 π^E_{-n} を代替方策 $\tilde{\pi}_{-n}$ に置き換え推定する方法を提案する．Algorithm 1 に提案法のアルゴリズムを示す．まず，各 iteration において，報酬の重みと代替方策を更新するエージェント集合 $\tilde{\mathcal{N}}$ を選択する．次に，MDCE IRL により報酬の重み θ_n を更新する．更新の打ち切りは，特徴期待ベクトルが十分一致した場合か，打ち切り回数に達した場合とする．最後に，Soft Q-Learning[Zhou 18] により方策 π_n を更新したのち，$\tilde{\pi}_{-n}$ を π^E_{-n} へ近づける「補完」を実行する．以下では，エージェントの選択法と，代替方策の更新及び補完法について述べる．

Algorithm 1 Multi-agent MDCE

Input: Markov Game$\backslash\{R_n\}_{n\in\mathcal{N}}$
Input: Expert trajectories $\mathcal{D}$
Output: reward weight $\{\theta_n\}_{n\in N}$
Initialize policies $\{\pi_n\}_{n\in N}$ and $\{\theta_n\}_{n\in N}$
for iteration $= 1, 2\ldots$ **do**
  $\tilde{\mathcal{N}} \leftarrow$ Selector($\mathcal{N}$)
  $\theta_n \leftarrow$ MDCE($\mathcal{D}^n, \tilde{\pi}_{-n}$) $\quad \forall n\in\tilde{\mathcal{N}}$
  $\pi_n \leftarrow$ SoftQ($\theta_n, \tilde{\pi}_{-n}$) $\quad \forall n\in\tilde{\mathcal{N}}$
  $\tilde{\pi}_n \leftarrow$ Completation($\pi_n, \mathcal{D}$) $\quad \forall n\in\tilde{\mathcal{N}}$

■エージェントの選択法

エージェントの選択法は二種類ある．図 1 にエージェントが 2 体の場合の更新手順と代替方策の流れを示す．一つめの選択法は，エージェントを 1 体ずつ選択して更新する Cyclic である．Cyclic は，1 体ずつ報酬をエキスパートへ少しずつ近づけていき，更新した報酬に対する方策を代替方策として用いる．二つめの選択法は，全エージェントの報酬を同時に更新する Parallel である．Parallel は，全エージェントの報酬更新を並列に実行しつつ，更新した報酬に対する代替方策を定期的に交換する．



(a) Cyclic　　(b) Parallel

図 1: 報酬の更新順と代替方策の流れ：$|\mathcal{N}| = 2$ の場合．$i-1, i, i+1$ は iteration

■代替方策の更新・補完法

MDCE IRL において報酬 θ に対する方策 π_θ は式 (9)，式 (10) の Soft Bellman 方程式を満たすことが知られている [Zhou 18]．ここで $\mathrm{softmax}_{a\in\mathcal{A}} Q_\theta^{\mathrm{soft}}(s,a) = \log \sum_{a\in\mathcal{A}} \exp\left(Q_\theta^{\mathrm{soft}}(s,a)\right)$ とする．よって代替方策の更新には，Algorithm 2 に示す TD-base の Soft Q-Laerning を用いる．最後に，更新した代替方策に対し，エキスパート軌跡に含まれる状態・行動を確率 1 で取るよう，代替方策を補完する．この操作は，推定中の方策とエキスパート方策の確率分布を近づけることを意図している．

$$Q_\theta^{\mathrm{soft}}(s,a) = \theta^\top f(s,a) + \gamma \sum_{s'\in\mathcal{S}} T(s'|s,a) V_\theta^{\mathrm{soft}}(s') \tag{9}$$

$$V_\theta^{\mathrm{soft}}(s) = \mathrm{softmax}_{a\in\mathcal{A}} Q_\theta^{\mathrm{soft}}(s,a) \tag{10}$$

$$\pi(a|s) = \exp\left(Q_\theta^{\mathrm{soft}}(s,a) - V_\theta^{\mathrm{soft}}(s)\right) \tag{11}$$

Algorithm 2 Soft Q-Learning

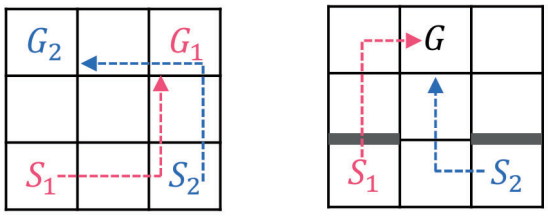
Input: reward weight θ_n, explore policy π, other agent's policy $\tilde{\pi}_{-n}$

1: **for** $t = 0, 1, 2, \cdots$ **do**
2: Generate sample (s_t, a_t, s_{t+1}) from $\pi, \tilde{\pi}_{-n}$
3: $Q_n^{\mathrm{soft}}(s_t, a_{t,n}) \leftarrow Q_n^{\mathrm{soft}}(s_t, a_{t,n}) + \eta_t(s_t, a_{t,n}) \cdot$
4: $\left[\theta_n^\top f_n(s_t, a_t) + \gamma V_n^{\mathrm{soft}}(s_{t+1}) - Q_n^{\mathrm{soft}}(s_t, a_{t,n})\right]$

5. 計算機実験

5.1 実験設定

図 2 に二つの実験環境を図示する．どちらの環境も 3 × 3 の GridWorld であり，2 体のエージェントが各自のスタート座標 (S_1, S_2) からゴール G へ最短 step で到達することを目的とする．状態集合は全エージェントの座標の組み合わせ，各エージェントの行動集合は $A_1 = A_2 = \{\mathrm{up}, \mathrm{down}, \mathrm{right}, \mathrm{left}\}$ であり，1step で隣接する四方向のセルに移動できる．ただし，壁に移動する場合と，2 体のエージェントが同じセルへ移動する場合は，遷移前の座標に留まる．加えて，後者の条件は，ゴール座標へ移動する場合を除く．GW2 では，スタート座標のセルと 1 つ上のセルの間に障壁があるため，S_1, S_2 で up がとられた場合は 1/2 の確率で遷移に失敗する．いずれか，もしくは両方のエージェントがゴールへ到達した状態は吸収状態として扱う．



(a) GW1：決定的遷移　　(b) GW2：確率的遷移

図 2: 環境とエキスパート軌跡：エージェント 1, 2 に対して s_1, s_2 は初期座標，g_1, g_2 はゴール．太線は障壁

各環境のエキスパート軌跡には Nash 均衡解の一つを与える．図 2 にエキスパート方策を矢印で示す．エキスパートは，矢印に沿った座標でそれぞれの行動を決定的にとるものとする．この軌跡は，ゴールに到達したエージェントに +100，同じセルに移動しようとした場合 −1 の報酬を与えたときの Nash 均衡解である [Hu 03]．特徴ベクトルは全状態・行動対に対するバイナリベクトルとし，特徴期待ベクトルは最大ステップ数 50 の軌跡を 100 本サンプリングから式 (7) で求める．初期の重みは 0 ベクトル，方策は一様分布とし，各 iteration ではステップ幅 0.1 の最急降下法で MDCE IRL を最大 100 回更新する．また，Inner-Loop の Soft Q-Learning は 100episode 学習する．報酬の重み更新には正則化なしの場合と L2 正則化の場合をそれぞれ実験した．

5.2 実験結果

GW1 のおける 10 試行平均および標準偏差の推移を図 3 と図 4 に示す．図 3 はエージェントの選択が Cyclic の場合，図 4 は Parallel の場合である．横軸は iteration，縦軸はエキスパート軌跡との特徴期待ベクトルの差のノルム表し，エキスパートと完全に一致するとき 0 となる．結果から，GW1 ではエキスパート軌跡と完全に一致する報酬が推定したことがわかる．また，エキスパート軌跡による補完は収束までの iteration を減らし，Cyclic に比べ Paralell のほうが早く収束している．この結果の原因としては，代替方策 $\tilde{\pi}_n$ がエキスパート方策 π_n^E に一致したとき分解された問題が元の M-MDCE IRL に一致することや，補完により推定対象のエージェントがエキスパート軌跡上を遷移することを邪魔しなかったことが考えられる．

次に，図 5 に GW2 に Cyclic を適用した結果を示す．GW2 では，代替方策を補完しない場合，エキスパートと一致する解が得られず，補完を用いてもエキスパートと完全に一致しなかった．この結果は，Parallel においても同様である．しかし，推定報酬のうち最も値の大きな状態行動上位 5 組 (図 6) を確認したところ，エキスパート軌跡上に大きな報酬が置かれていることから，結果が妥当であると確認できた．エキスパートと完全に一致しない理由としては，環境の確率的遷移により，特徴期待ベクトル計算のためのサンプル分布が一致しずらいことが考えられる．

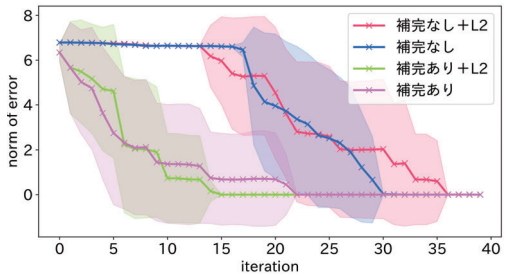


図 3: GW1 + Cyclic

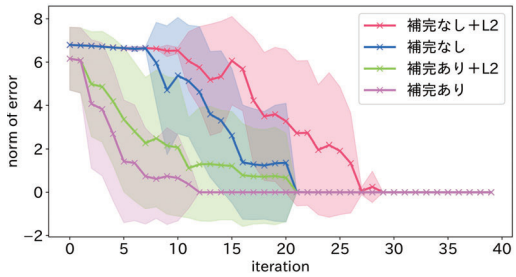


図 4: GW1 + Parallel

6. 結論と今後の課題

本論文では，MAS におけるエージェントの行動ルール設計，およびエージェントの目的理解のため，新たな MAIRL を提案した．具体的には，infinit-horizon の Markov game において，最大エントロピー法に基づく MAIRL を提案した．提案法は，決定的なエキスパート軌跡から報酬が推定できることを，3x3 の GridWorld の実験から示した．今後の課題として，異なる環境や連続状態空間への適用，収束・最適性の考察や，獲得する均衡解概念の特定を挙げる．

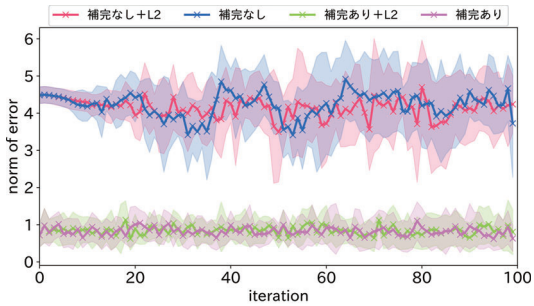


図 5: GW2 + Cyclic

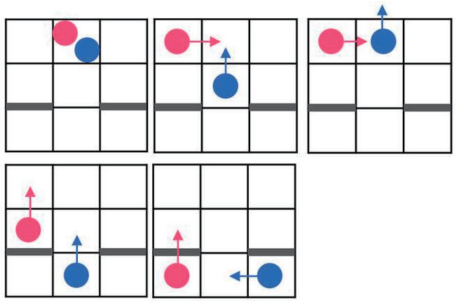


図 6: 推定報酬値の大きな状態行動 (GW2 + Cyclic)

参考文献

[Abbeel 04] Abbeel, P. and Ng, A. Y.: Apprenticeship Learning via Inverse Reinforcement Learning, in *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pp. 1–, New York, NY, USA (2004), ACM

[Bogert 18] Bogert, K. and Doshi, P.: Multi-robot inverse reinforcement learning under occlusion with estimation of state transitions, *Artificial Intelligence*, Vol. 263, pp. 46–73 (2018)

[Ho 16] Ho, J. and Ermon, S.: Generative Adversarial Imitation Learning, in Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. eds., *Advances in Neural Information Processing Systems 29*, pp. 4565–4573, Curran Associates, Inc. (2016)

[Hu 03] Hu, J. and Wellman, M. P.: Nash Q-learning for general-sum stochastic games, *Journal of machine learning research*, Vol. 4, No. Nov, pp. 1039–1069 (2003)

[Le 17] Le, H. M., Yue, Y., Carr, P., and Lucey, P.: Coordinated Multi-Agent Imitation Learning, in *International Conference on Machine Learning*, pp. 1995–2003 (2017)

[Lin 18] Lin, X., Beling, P. A., and Cogill, R.: Multiagent Inverse Reinforcement Learning for Two-Person Zero-Sum Games, *IEEE Transactions on Games*, Vol. 10, No. 1, pp. 56–68 (2018)

[Namikoshi 18] Namikoshi, K. and Arai, S.: Estimation of the heterogeneous strategies from action log, in *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 1310–1317ACM (2018)

[Natarajan 10] Natarajan, S., Kunapuli, G., Judah, K., Tadepalli, P., Kersting, K., and Shavlik, J.: Multi-Agent Inverse Reinforcement Learning, in *2010 Ninth International Conference on Machine Learning and Applications*, pp. 395–400, Washington, DC, USA (2010), IEEE

[Ng 00] Ng, A. Y. and Russell, S. J.: Algorithms for Inverse Reinforcement Learning, in *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 663–670, San Francisco, CA, USA (2000), Morgan Kaufmann Publishers Inc.

[Reddy 12] Reddy, T. S., Gopikrishna, V., Zaruba, G., and Huber, M.: Inverse reinforcement learning for decentralized non-cooperative multiagent systems, in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1930–1935, Seoul, Korea (South) (2012), IEEE

[Song 18] Song, J., Ren, H., Sadigh, D., and Ermon, S.: Multi-Agent Generative Adversarial Imitation Learning, in Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. eds., *Advances in Neural Information Processing Systems 31*, pp. 7471–7482, Curran Associates, Inc. (2018)

[Šošić 17] Šošić, A., KhudaBukhsh, W. R., Zoubir, A. M., and Koeppl, H.: Inverse Reinforcement Learning in Swarm Systems, in *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '17, pp. 1413–1421, Richland, SC (2017), International Foundation for Autonomous Agents and Multiagent Systems

[Wang 18] Wang, X. and Klabjan, D.: Competitive Multi-agent Inverse Reinforcement Learning with Sub-optimal Demonstrations, in Dy, J. and Krause, A. eds., *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 5143–5151, Stockholmsm?ssan, Stockholm Sweden (2018), PMLR

[Zhou 18] Zhou, Z., Bloem, M., and Bambos, N.: Infinite Time Horizon Maximum Causal Entropy Inverse Reinforcement Learning, *IEEE Transactions on Automatic Control*, Vol. 63, No. 9, pp. 2787–2802 (2018)

[Ziebart 10] Ziebart, B. D., Bagnell, J. A., and Dey, A. K.: Modeling Interaction via the Principle of Maximum Causal Entropy, in *ICML* (2010)