NEURON AS AN AGENT

Shohei Ohsawa
The University of Tokyo
7 Chome-3-1 Hongo, Bunkyo, Tokyo
ohsawa@weblab.t.u-tokyo.ac.jp

ABSTRACT

The reason why swarm of agents solve real-world problem well is, interestingly, same as the principle of representation learning: good representation improves perforamance of machine learning. Most of the problem in real-world is not Markov decision process (MDP) but partially observed MDP (POMDP). On the POMDP environment, good observation yields good action. In this paper, we optimise a deep neural network as a multi-agent system as a natural extension from representation learning to multi-agent reinforcement learning for POMDP. To achieve that, we propose a novel learning framework, neuron as an agent (NaaA). In NaaA, an individual unit is considered as an agent, and they maximizing profit instead of minimizing error. To prevent dillemma, we borrow idea from machanism design, a field of game theory. To this end, we show all the unit have valid price reflecting their contribution to performance at convergence. We confirm the result by numerical experiment using Atari and VizDoom.

1 INTRODUCTION

マルチエージェントによる強化学習が現実の課題に対して有効である理由は、興味深いことに表現学習の原理と一致する：有用な表現は予測性能を向上させる。現実にある問題の多くは、DQN が前提としているマルコフ決定過程 (MDP) ではなく部分的観測マルコフ決定過程 (POMDP) である (Sutton & Barto, 1998, p.258)。POMDP において真の状態は完全には不可視であり、良質な観測結果が良質な行動に結びつくため、アテンション、すなわち、有限のリソースを使って何を観測するかの設計が必要である。したがって、観測に用いられるエージェントが多いほど、将来的に獲得できる報酬も高くなる。本論文は、観測という行動においてエージェントとユニットは等価であるということを主張する。

本研究のゴールは、すべてのユニットが自律的に動作すると仮定した場合に、システム全体が獲得する累積報酬を最大化することである。そのために、NaaA は微分可能 (differentiable) な関数で構成されるニューラルネットワークを、報酬系の概念を用いて拡張する。報酬の分配のためにオークション理論を用いて、ジレンマ問題を解決した上で、レイヤーの入札価格が、ユニットが存在する場合と存在しない場合の差である counterfactual reward (Agogino & Tumer, 2006) に等しくなることを示す。以下では報酬分配のフレームワークである profit maximization について述べ、Valuation Net について説明を行う。

NaaA は、神経細胞の持つエネルギー消費のメカニズムをモデリングする。神経回路に含まれるニューロンは一つの細胞であるため、エネルギーを消費する。通常の細胞と同様に酸素や ATP がエネルギー源となり、これらはニューロンと接続した、アストロサイトから供給される。アストロサイトは脳の構造を支えるグリア細胞の一種であり、血管からニューロンへの栄養供給を行う。エネルギー量は有限であるため、不要なニューロンはアポトーシスによって死滅する。アポトーシスは NGF (nerve growth factor), BDNF (brain derived neurofactor) などの神経栄養因子 (neurotorophin; NTF) によって制御されるため、より多くの NTF を獲得できたニューロンが生存する。

これまでのニューラルネットワークの目的は、予測誤差 e の最小化にあった。そのため、各ユニットはバックプロパゲーションによって出力 y による微分 $\partial e/\partial y$ を計算し、e が小さくなる方向に y の大きさを制御していた。NaaA において、ユニットの目的は大域的誤差の最小化ではなく、ユニット自身の期待リターンの最大化である。

実験では、標準的な強化学習のタスクによる数値実験を用いて NaaA が POMDP の問題の精度を高めることを示す。具体的に、Atari および VisDoom における環境を用いて、既存研究が DQN や A3C を上回ることを示す。

2 RELATED WORK

現在成功している深層強化学習のモデルの多くは、単一のエージェントが環境の観測から認知、行動決定といった一連のプロセスを担う。DQN (Mnih et al., 2015; Silver et al., 2016) は、Atari のスクリーン系列から最適な行動を決定したり、AlphaGo のモジュールとして囲碁の盤面から勝利に最も近い一手を選ぶ。DDPG (Lillicrap et al., 2015) は物理空間において摩擦や重力係数などの条件を考慮した多関節の制御を実現する。単一のエージェントを用いて強化学習を解くという試みは、人間の持つ身体性のアナロジーから考えると一見して妥当であるように思えるが、現実世界は open world であり、単一のエージェントが完全に情報が観測することが難しい。そのためマルチエージェントによるアプローチが求められている。

マルチエージェントシステムによるアプローチにはいくつかの方法がある。一つは、学習の効率を高めるために、同一のモデルに従う複数のエージェントを用いて探索を行う方法であり、Gorilla, A3C (**?**) などで採用されている。二つ目は、サッカーゲームのようにアクチュエーターを増やすことによって行動の量を増やす方法であり、三つ目は、センサーを増やすことによって観測の量を増やす方法であり、自動運転や IoT などで行われている。本研究が対象にするのは、三つ目のアプローチである。

深層強化学習を POMDP 環境に適用する場合、観測困難な環境からいかに真の状態を推定するかが重要になる。Deep Recurrent Q-Network (DRQN) (Sorokin et al., 2015) は、隠れマルコフ連鎖を想定し、リカレントニューラルネットワーク (RNN) を用いて真の状態を推定している。他にも、エレベーター制御 (Crites & Barto, 1998)、センサーネットワーク (Fox et al., 2000)、ロボットサッカー (Stone & Veloso, 1998) などがマルチエージェントによって解かれている（あとで修正: 専門家に確認中）。

マルチエージェントによる観測および認知の枠組みは、センサー処理の分野ではエッジコンピューティング (Bonomi et al., 2012) としても知られている。エッジコンピューティングは分散環境を前提とした信号処理のモデルであり、一つの処理系がすべてのデータを処理するのではなく、複数のセンサーの情報を一つのエッジサーバが集約し、複数のエッジサーバが次元削減したデータをデータセンターに送るという階層的な構造をしている。

各神経細胞を独立した生物として捉える見方はニューラルダーウィズム (Edelman, 1987) と呼ばれる。実際、人間の脳を観察しても、各神経細胞は独立して動作する。神経細胞は胚細胞からの発達段階において Nerve Growth Factor、BDNF といった神経栄養因子 (NTF) を追求することが知られており、十分な NTF を受け取ることができなかったニューロンはアポトーシスを引き起こして自死することが知られている (Almeida et al., 2005)。

マルチエージェントで強化学習の問題を解決する場合には、信頼度割り当て問題の解決が重要になる。そこで、エージェントの信頼度を、そのエージェントがいた場合と、いなかったと仮定した場合の差として定量化する研究が行われている。QUICR-learning (Agogino & Tumer, 2006) では、エージェント i が reward $R(a_t)$ の代わりに、そのエージェントがある行動 a_{ti} をとった場合 a_t と取らなかった場合 $a_t - a_{ti}$ の差、counterfractual reward $R(a_t) - R(a_t - a_{ti})$ の減衰和を最大化している。COMA (Foerster et al., 2017) は、actor-critic において critic が共通しており、actor がマルチエージェントであるという actor-critic の仕組みを考え、それぞれの actor が counterfractual reward を最大化するような仕組みを考えている。

これらの研究の問題は、情報がすべて共有されているという前提に立っており、信頼度を割り当てるという性善説に基づいている点にある。そのため、裏切ることが考えられる人物がいると、予想外の内容が学習されてしまう。たとえば、IoT のような実環境における問題を考えると、センサーを持っている主体は異なる人物であるために、協力行動をとるとは考えにくい。

本研究では、すべてのニューロンをエージェントとみなす方法を提案している。

3 NEURON AS AN AGENT

ニューラルネットワークのユニット間のトポロジーを有向グラフ $\mathfrak{G} = (V, E)$ で表す。$V = \{v_1, \ldots, v_N\}$ はユニットの集合であり、$E \subset V^2$ はユニットの接続関係を表すエッジの集合

である。$(v_i, v_j) \in E$ であるとき、$v_i \to v_j$ という接続関係が成立し、v_j は v_i から値を入力する。ユニットの v_i の時刻 t における出力を $x_{it} \in \mathbb{R}$ で表す。ユニット i の入力元の集合を $N_i^{\text{out}} = \{j|(v_i, v_j) \in E\}$、出力先の集合を $N_i^{\text{in}} = \{j|(v_j, v_i) \in E\}$ で表現する。

NaaA は $\mathfrak{G}$ をマルチエージェントシステムであるととらえ、v_i に以下の前提を加える。

- N1: （利己性）v_i は、各時点 t において、汎化誤差の最小化ではなく、自身の期待リターン G_{it} の最大化を目的として行動する。
- N2: （保存則）v_i が受け取る報酬 R_{it} の総和は、マルチユニットシステム全体が外的環境から得る報酬 R_0 に等しい。
- N3: （取引）v_i は信号 x_i を $v_j \in V$ に伝達する際に、信号と引き換えに報酬 ρ_{jit} を受け取る。
- N4: （NOOP）v_i は、期待リターンが 0 の NOOP (no operation) という行動をオプションとして持つ。NOOP では、ユニットは何も入力せず、何も出力しない。

N1 はユニットがエージェントとして振る舞うことを述べている。N2, N3 は NTF の分配に、N4 はニューロンのアポトーシスに相当する。NOOP が選択されるのは、それ以外のすべての行動の期待報酬が負であった場合である。以下ではこれらの前提から出発して、NaaA の仕組みを構築していく。

3.1 Cumulative Profit Maximization Framework

通貨は $\mathfrak{G}$ 上を流れる。ユニット i がユニット j に対して、時刻 t で支払う通貨を $\rho_{ijt} \in [0, \infty)$ で表す。ユニット i が時刻 t で外部から得る報酬を R_i^{ex} と書き、消費エネルギーを α_{it} で表す。時刻 t に i が獲得する報酬 R_i の報酬は次のように表現される。

$$R_{it} = \left[R_{it}^{\text{ex}} + \sum_{j \in N_i^{\text{out}}} \rho_{jit} \right] - \left[\sum_{j \in N_i^{\text{in}}} \rho_{ijt} + \alpha_{it} \right] \tag{1}$$

この式は、符号が正の項と負の項の二つに分解される。前者を収益 (revenue)、後者をコスト (cost) と呼び、それぞれ r_{it}, c_{it} で表す。

ユニット v_i が最大化する対象のリターン G_{it} は次のように書くことができる。

$$G_{it} = \sum_{k=0}^{T} \gamma^k R_{i,t+k} = \sum_{k=0}^{T} \gamma^k (r_{i,t+k} - c_{i,t+k}) \tag{2}$$

$R_{it} > 0$、すなわち、$r_{it} > c_{it}$ であれば、ユニットは得たデータに対して付加価値を与えていることになる。もし、すべての t に対して $R_{it} < 0$ であれば、$y_{it} < 0$ であるから、ユニットは NOOP になる。

4 Optimization

ここでは、NaaA のフレームワークにおいて、ニューラルネットワークを最適化する手法について説明する。最適化の方法はいくつかあるが、practical な方法として、envy-free auction と valuation net の二つを紹介する。

実際には、このフレームワークは期待通りに動作しない。なぜなら、この式を最適化すると、コスト最小化によってエージェントは他のエージェントに対して金額を支払わないことがナッシュ均衡解として得られるためである。

Theorem 4.1. *The Nash equivalem of the game* (r_{ijt}) *is* 0.

ただちに、次の系が成立する。

Corollary 4.1. *Cumulative Maximization Framework* では、外的環境から報酬を受け取らないニューロンはすべて *NOOP* になる。

すなわち、単純な累積利益最大化フレームワークでは、すべてのニューロンが活動せず、マルチエージェントシステムは無情報でアクションを選択する必要が生じ、これはランダムなア

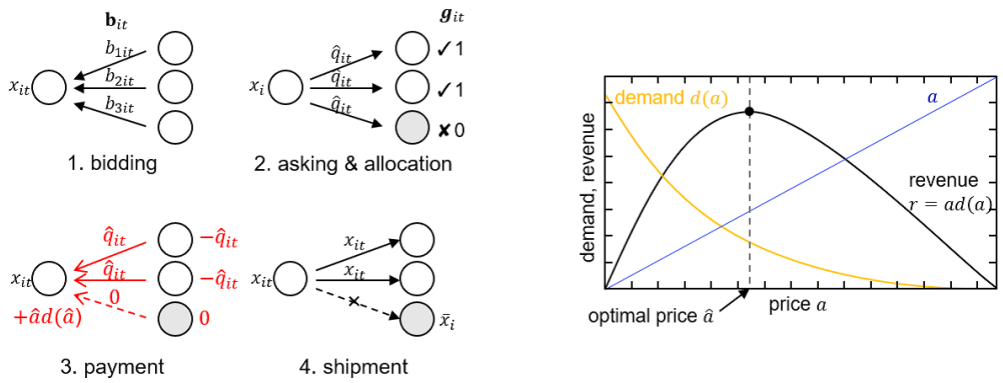

Figure 1: Left: NaaA による取引の流れ。Right: ユニットの価格決定方法。ユニットの収益は単調減少な需要と価格の積となり、これを最大化する価格が最適価格となる。

クションを取っている状況に等しい。したがって、明らかに外的環境からの報酬 R_t の大きさは小さくなる。これは、ナッシュ均衡がパレート最適と一致しないパレート劣位な状況が発生する。

4.1 Digital Goods Auction

パレート効率な仕組みを作るために、我々はオークション理論における digital goods auction からアイデアを借りる。オークション理論は、ゲーム理論におけるメカニズムデザインという分野に属しており、複数のエージェントの利害を一致させ、全体としてパレート最適を目指すことを目指している。digital goods auction は、本や音楽などの、複製可能な財を割り当てる仕組みを作っている。

Digital goods auction によってジレンマを防ぐ方法はいくつかあり、どれも本問題へ適用可能ではあるが、本研究では単純な前提のみを設けるだけでよいという理由から envy-free auction を用いる。これは、同じ時点取引において、一つのユニットの価格を同じにするというものである。NaaA において、これは次のシンプルな前提によって表現できる。

N5: （一物一価）$\rho_{j_1,i,t}, \rho_{j_2,i,t} > 0$ であれば $\rho_{j_1,i,t} = \rho_{j_2,i,t}$

これは、ユニット v_i は同じ時間 (timing) t に個有の価格を持つことを意味する。この価格を q_{it} で表す。

Envy-free auction の流れを Figure 1 の左に示す。まず、信号を送信する側を売り手、受信する側を買い手と呼ぶ。買い手はユニットに対して入札 b_{jit} を行う (1)。次に、入札額をもとに、売り手は価格 $\mathbf{q}_{it}$ を決定し、割当を行う (2)。このとき、$b_{jit} \geq \mathbf{q}_{it}$ であれば割当を行って $g_{jit} = 1$ とし、そうでなければ $g_{jit} = 0$ とする。割当を行った後は、$\rho_{jit} = g_{jit}\mathbf{q}_{it}$ として、送金を行い (3)、売り手は割当を行ったノードに対してのみ信号 x_i を送付する (4)。信号を受け取れなかったノードは、x_i の期待値 $\mathbb{E}_\pi[x_i]$ によって x_i を近似する。

ここで、売上とコストの選定方法について分けて説明を行う。まず、売上について考える。エージェントは利益を最大化したいため、利益は次のようにして与えられる。

$$r_{it} = \sum_{j \in N_i^{\text{in}}} g(b_{jit}, q_i) q_i + R_i^{\text{ex}} = q_i \sum_{j \in N_i^{\text{in}}} g(b_{ji}, q_i) + R_i^{\text{ex}}$$
$$= q_i d_t(q_t) + R_i^{\text{ex}} \tag{3}$$

ここで、a は価格、$d_t(a)$ はユニット i の信号に対する価値を a 以上と評価しているエージェントの数であり、需要 (demand) と呼ぶ。同様の式で、右辺を最大化する a を最適価格と呼び、$\hat{a}_{it}$ で表す。したがって、最適価格 $\hat{q}_{it}$ は次のようにして与えられる。

$$\hat{q}_{it} = \text{argmax} q_{it} d_t(q_{it}) \tag{4}$$

この仕組みを、Figure 1 の右に図示する。$d_t(q_{it})$ は単調減少な関数であり、q_{it} との積によって表現される。

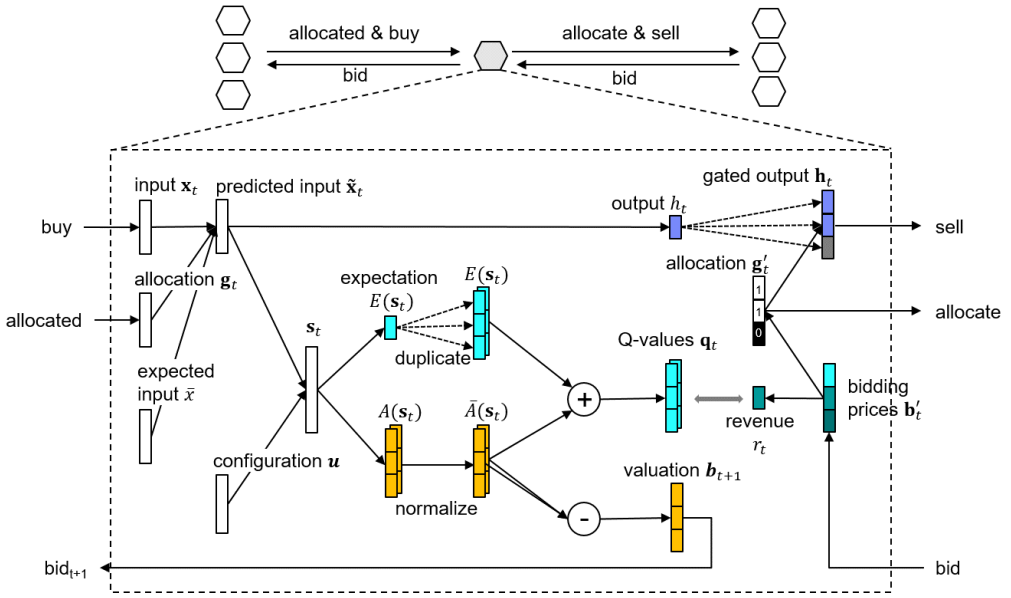

Figure 2: Valuationn Net は情報の価値を評価し、bidding price を決定する。下位のニューロンに対して入札し、信号を購入する。購入したデータを用いて、データを次のニューロンに対して売る。

次に、コストについて考える。これは、正しいデータを購入した場合とそうでない場合の期待リターン G_{it} の差に等しい。すなわち、

$$\begin{aligned} o_t &= \mathbb{E}_\pi \left[r_{i,t+1} \mid a_t = 1 \right] - \mathbb{E}_\pi \left[r_{i,t+1} \mid a_t = 0 \right] \\ &= \mathbb{E}_\pi \left[R_{i,t+1} \mid a_t = 1 \right] - \mathbb{E}_\pi \left[R_{i,t+1} \mid a_t = 0 \right] \\ &= Q(s_t, 1) - Q(s_t, 0), \end{aligned} \tag{5}$$

ただし、Q は状態行動価値関数であり、一手先のコストはどちらの行動を選んでも一定であると仮定している。o_{it} を counterfractual state-action value と呼ぶ。これは QUICR (Agogino & Tumer, 2006) と導出は異なるが等価である。すなわち、エージェントが支払うコストは、データの購入に成功した場合は $\hat{a}_{it}$ であり、それ以外は o_{it} となる。

では、コストを最小化するためのエージェントの入札額 b_{it} は何か。これについては次の定理が成立する。

Theorem 4.2. コストを最小化する最適な入札額は $b_{it} = o_{it}$ である。

証明については Appendix を参照。

すなわち、エージェントは自身の機会損失のみを問題にすればよい (!) したがって、NaaA のメカニズムでは、エージェントはあたかも他のエージェントを価値評価 (valuation) し、その価値を正直に申告していることを意味する。

系として次の解が得られる。

Corollary 4.2. *The Nash equivalem of the envy-free game* $(\mathbf{b}, q)$ *is* $(\mathbf{o}_t, \max_q q d_{\mathbf{o}_t}(q))$.

4.2 Valuation Net

残る問題は、$\mathbf{o}_t$ をいかに推定するかである。この推定には様々な方法が存在しており、多くのメソッドを使うことができるが、本論文では Q の推定に Q-learning を採用する。ただし、SARSA や actor-critic などの on-policy な方法も使うことができることを補足する。

図 2 に示す Valuation Net は、通常のニューラルネットワークのユニットに、Q-learning による valuation を組み合わせたネットワークである。まず、上部はエージェント間の通信について示したものである。ニューラルネットワークではユニットを円で表現するのが通例である

が、ここではユニットをエージェントとしてみなすことを強調して、六角形で一つのユニットを表現している。エージェント間では、通常のニューラルネットワークと同様の信号の通信以外に、取引に関する通信 (allocate, buy, sell & bid) が発生する。

Valuation Net では、状態 $\mathbf{s}_t$ として、予測後入力 $\tilde{\mathbf{x}}_t$ および入力に依存しない構成情報 $\mathbf{u}$ を横につなげたベクトル $(\tilde{\mathbf{x}}_t^{\mathrm{T}}, \mathbf{u}^{\mathrm{T}})^{\mathrm{T}}$ を用いる。構成情報の一例としてはユニットのパラメータがあげられ、たとえば重みやバイアスの情報を用いることができる。

状態からの Q 関数の予測にニューラルネットワークを用いる。エージェントが受け取った売上に基づき時間差分 (TD)-誤差 が計算され、ネットワークが訓練される。ネットワークの構成にはこれまでの deep Q-learning で用いられている二重化ネットワーク (dualing network) (Wang et al., 2015) のテクニックを用いる。オリジナルの文献 (Wang et al., 2015) で述べられている二重化ネットワークは、学習を加速するために、状態関数と、Q 関数との差分を別々に予測する手法である。Dosovitskiy & Koltun (2016) はこれに対して、差分の要素の総和が 0 になるように正規化するよう改良している。本研究では Dosovitskiy & Koltun (2016) の手法に従い、期待値 $E(\mathbf{s}_t)$ と正規化差分 $\tilde{A}(\mathbf{s}_t)$ を別々に求める。

Q 関数は次のように表現される。

$$Q(\mathbf{s}_t, a_t) = E(\mathbf{s}_t) + \tilde{A}(\mathbf{s}_t, a_t)$$
$$\sum_{i+1}^{k} \tilde{A}_i(\mathbf{s}_t, a_t) = 0 \tag{6}$$

第 2 式を満たすために、まず、$\mathbf{s}_t$ に基づいた予測を行い、次のような正規化を行う。

$$\tilde{A}_i(\mathbf{s}_t, a_t) = A_i(\mathbf{s}_t, a_t) - \frac{1}{k}\sum_{j=1}^{k} A_j(\mathbf{s}_t, a_t) \tag{7}$$

次に、valuation を行い、bidding price $\mathbf{b}_t$ を求める。b_{it} の値は式 5 および式 6 より、最適な入札価格 $\hat{b}_{it}$ は次のように計算できる。

$$\hat{b}_{it} = \tilde{A}(\mathbf{s}_t, 1) - \tilde{A}(\mathbf{s}_t, 0) \tag{8}$$

Valuation Net ではこの式に基づき、advantage の出力を引き算することで入札価格を計算している。

5 Experiment

6 Discussion

7 Conclusion and Future Works

本論文では、POMDP の問題設定において良質な特徴表現を得るために、ニューラルネットワーク上の各ユニットをエージェントとして扱うフレームワーク、NaaA について述べた。NaaA のフレームワークでは、ジレンマ問題を解決し、それぞれのエージェントの持つ付加価値がナッシュ均衡として得られ、全体としてパレート最適になることを示した。入札価格の決定アルゴリズムの一つとして、Q-learning に基づくネットワーク Valuation Net を示した。評価実験では、Atari と VizDoom を用いた実験を行い、実験結果が既存手法よりもよくなることを示した。

今後の方向性として、高速化、Valuation Net を A3C などの on-policy な手法で置き換えるといった方向性の他、神経科学的な説明を可能にしていくといった方法、遺伝的アルゴリズムとの組み合わせが挙げられる。

Appendix

A.1 定理 4.2 の証明

買い手の獲得する生涯報酬 G は次で与えられる。

$$G(b, q) = g(b, q) \cdot (v - q) + G_0, \tag{9}$$

ただし、g は割当 (allocation) であり、G_0 はユニットを購入しなかった場合の生涯報酬である。割当は、$g(b,q) = H(b-q)$ が成立する。H はステップ関数である。

買い手にとって売り手が提示する asking price q は未知であるため、q を台 $[0,\infty)$ の上の確率変数であるとして扱い、期待値 $\mathbb{E}_q[G(b,q)]$ を最大化することを考える。このとき、

$$\begin{aligned}\frac{\partial}{\partial b}\mathbb{E}_q[G(b,q)] &= \frac{\partial}{\partial b}\int_0^\infty (H(b-q)\cdot(v-q)+G_0)p(q)dq \\ &= \frac{\partial}{\partial b}\left[\int_0^b (v-q)p(q)dq + G_0\int_0^\infty p(q)dq\right] \\ &= \frac{\partial}{\partial b}\int_0^b (v-q)p(q)dq \\ &= (v-b)p(q=b)\end{aligned}$$

したがって、 $\mathbb{E}_q[G(b,q)]$ が最大となるための条件は $b=v$ ある。

REFERENCES

A. K. Agogino and K. Tumer. QUICR-learning for multi-agent coordination. AAAI'06, 2006.

R. D. Almeida, B. J. Manadas, C. V. Melo, J. R. Gomes, C. S. Mendes, M. M. Graos, R. F. Carvalho, A. P. Carvalho, and C. B. Duarte. Neuroprotection by bdnf against glutamate-induced apoptotic cell death is mediated by erk and pi3-kinase pathways. *Cell death and differentiation*, 12(10): 1329, 2005.

F. Bonomi, R. Milito, J. Zhu, and S. Addepalli. Fog computing and its role in the internet of things. In *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*, pp. 13–16. ACM, 2012.

R. H. Crites and A. G. Barto. Elevator group control using multiple reinforcement learning agents. *Machine learning*, 33(2):235–262, 1998.

A. Dosovitskiy and V. Koltun. Learning to act by predicting the future. *ICLR'17*, 2016.

G. M Edelman. *Neural Darwinism: The theory of neuronal group selection.* Basic books, 1987.

J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson. Counterfactual multi-agent policy gradients. *arXiv:1705.08926*, 2017.

D. Fox, W. Burgard, H. Kruppa, and S. Thrun. A probabilistic approach to collaborative multi-robot localization. *Autonomous robots*, 8(3):325–344, 2000.

T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *ICLR'16*, 2015.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

I. Sorokin, A. Seleznev, M. Pavlov, A. Fedorov, and A. Ignateva. Deep attention recurrent q-network. *arXiv:1512.01693*, 2015.

P. Stone and M. Veloso. Towards collaborative and adversarial learning: A case study in robotic soccer. *International Journal of Human-Computer Studies*, 48(1):83–104, 1998.

R. S. Sutton and A. G Barto. *Reinforcement learning: An introduction*. A Bradford Book, 1998.

Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas. Dueling network architectures for deep reinforcement learning. *arXiv:1511.06581*, 2015.