2I3-2

分散秘密情報源からの強化学習 -時間分割モデルへの適用-

Reinforcement Learning from Distributed Private Information in Partitioned-by-Time Model

佐久間 淳[*1] Jun Sakuma　小林 重信[*1] Shigenobu Kobayashi　Rebecca N. Wright[*2]

[*1]東京工業大学大学院 総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

[*2]Rutgers University, DIMACS

We consider a problem of distributed reinforcement learning (DRL) from private perceptions. In our setting, agents' perceptions, such as states, rewards, and actions, are not only distributed but also are desired to be kept private. This can occur when agents' perceptions include private or confidential information. Conventional DRL algorithms could be applied to such problems, but do not necessarily guarantee privacy preservation. Additionally, DRL which learns only from local perceptions often sacrifice optimality. In this work, we design solutions that achieve optimal policies without requiring the agents to share their private information by means of well-known cryptographic tools, secure function evaluation.

1. はじめに

分散強化学習 (distributed reinforcement learning, DRL) とは分散エージェント・環境間の相互作用を通じて環境における最適な振舞を獲得する枠組みであり, センサーネットワークや移動ロボット群の制御などへの応用が期待されている. 既存の DRL には分散価値関数 [10] や政策勾配法 [7] などが知られている. これらのアプローチは環境について主に二種類の物理的制約を想定する. 一つは不安定なネットワーク環境や通信経路における制限など, 通信上の制約であり, もう一つは巨大な状態行動空間の扱いに伴うメモリ上の制約である. これらの制約に対処するため, 既存の DRL は分散エージェント間の知覚の共有をなるべく少なくした上で, 準最適な政策を獲得することを主な目的としてきた. 本稿では DRL の新しい枠組みとして, 通信帯域など物理的計算資源は潤沢に用意されているが, 分散エージェントの知覚に秘密情報が含まれるために情報共有ができないといった社会的制約が課せられるケースを扱う. 以下にエージェントのプライバシ保護が重要になる具体的な例を示す.

Optimized Marketing [1]: 顧客の購買行動のマルコフ決定過程 (MDP) によるモデル化を考える. タイムスタンプが付された顧客の状態とカタログ配送の履歴を状態変数, 顧客の購買行動を行動変数として考え, 事業者の長期利益の最大化を目的関数として価値関数を学習させる. 目的は事業者の最適なカタログ配送戦略を価値関数から獲得することである. もしこれらの履歴が二つ以上の事業者に管理されており, 個人情報保護のため共有が許されない場合, 価値関数を学習することは可能だろうか?

Load Balancing [2]: 工場間の負荷分散問題を考える. 負荷の高い工場から低い工場へタスクを転送することによって各工場の負荷を互いに平滑化することが目標である. 各工場は自身の未処理タスク数を観測できるが, 他の工場の未処理タスク数は機密保護のため観測できない. 各工場はタスクを他の工場に転送するかどうかを決定するが, 互いに情報を交換しなくとも, 最適な意思決定は可能だろうか?

連絡先: 佐久間 淳, 東京工業大学大学院 総合理工学研究科, 神奈川県横浜市緑区長津田町 4259, 045-924-5677, 045-924-5442, jun@fe.dis.titech.ac.jp

従来の DRL はこれらの問題に適用可能ではあるが, プライバシ保護は必ずしも保証されず, また情報交換を制限した上で学習を行う DRL では最適性も保障されない場合がある.

同様の目的意識で発展した分野にプライバシ保護型データマイニングがあげられる. Lindell らは分散秘密情報からの ID3 決定木学習法 [5] を提案し, その後, サポートベクターマシン [13], k-means クラスタリング [8] など様々な分散秘密情報源のための学習アルゴリズムが提案されてきた. Zhang らは [14] にてプライバシを保護した平均報酬型強化学習を提案した. ただし更新式が標準的な形式を取らないため, 強化学習が MDP について一般的に保障する最適性は保持されない. これらを考慮し, 本研究の目的は, 各エージェントのプライバシ保護を理論的に保障しつつ, 最適政策の獲得を保証する DRL アルゴリズムの提案とする.

秘匿関数計算 (secure function evaluation, SFE) を用いることによって分散秘密情報源を用いた任意の計算が実行可能になることが知られている [12, 4]. SFE は広い汎用性を持つ, よく知られた方法論であるが, 強化学習で扱われるような大規模な入出力を想定していない. そのため, DRL の SFE による直接実行は非現実的な時間を要する. 我々は提案アルゴリズムの一部に既存の SFE を用い, その他の部分では加法的準同型性公開鍵暗号を用いるより効率的な計算法を示す.

2. 問題設定

S を状態集合, A を行動集合とする。政策 π は状態行動対 (s,a) から, 状態 s において行動 a を取る確率 $\pi(s,a)$ への写像と定義される。政策 π の Q 関数は期待利得

$$Q^{\pi}(s,a) = E_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^{k} r_{t+k+1} | s_t = s, a_t = a\right\} \qquad (1)$$

として定義される。ここで γ は割引率 $(0 \leq \gamma < 1)$ である. 目標は最適 Q 関数 $Q^{*}(s,a) = \max_{\pi} Q(s,a)$ をすべての (s,a) において求めることである。SARSA 学習では Q 値を以下のように更新する.

$$\begin{aligned} \Delta Q(s_t,a_t) &\leftarrow \alpha(r_t + \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)), \\ Q(s_t,a_t) &\leftarrow \Delta Q(s_t,a_t) + Q(s_t,a_t), \end{aligned} \qquad (2)$$

ここで α は学習率である。Q 学習では ΔQ の更新を以下のように行う.

$$\Delta Q(s_t, a_t) \quad \leftarrow \quad \alpha(r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)).$$

環境が離散 MDP であるとき, これらの更新をある適切な条件の下で繰り返すことによって, Q 値は確率 1 で最適収束する [11]. 最適政策は最適 Q 関数から直ちに得られる。

2.1 DRL における秘密情報のモデル化

エージェントの知覚を $h_t = (s_t, a_t, r_t, s_{t+1}, a_{t+1})$, その集合を $H = \{h_t\}$ とする. エージェントの秘密情報は, H がどのように分割され, どの分割が各エージェントに観測・選択されるかによって定義される. DRL では主に二つの分割モデルを考えることができる.

一つは**時間分割モデル**である. このモデルでは, 任意の時間ステップ t において, m エージェントの中の一つのみが環境とインタラクションする。T^i を i 番目のエージェントが環境とインタラクションする時間ステップの集合としよう。ただし $T^i \cap T^j = \emptyset, (i \neq j)$ である. $H^i = \{h_t | t \in T^i\}$ とすると, この H^i が i 番目のエージェントの秘密情報と定義される. このモデルは 1 章で述べた optimized marketing の例に相当する.

もう一つは**観測分割モデル**である. このモデルでは全ての時間ステップにおいてすべてのエージェントが環境と同時にインタラクションすることを想定する。状態および行動は複数の状態変数あるいは行動変数の集合として定義され, 各エージェントは環境から状態変数の一部のみを観測し, 行動変数の一部のみを環境に行使できる。このモデルは Load balancing の例に相当する. 本稿では時間分割モデルのみを扱い, 観測分割モデルについては [9] を参照されたい.

時間分割モデルにおいて, 各エージェントの知覚 H^i を特定のエージェントに集約させた上で強化学習を実行し獲得された政策を π^c とする. このとき, 秘匿関数計算の標準的な定式化 [4] に基づけば, プライバシーを保護した強化学習 (privacy-preserving reinforcement learning, PPRL) は, 以下のように定義できる。

Statement 1. 時間分割モデルにおける i 番目のエージェントの入力を H^i とする. *PPRL* の実行後, 全エージェントは政策 π^c と等価な政策 π および, π から i 番目のエージェントが推測可能な情報を獲得するが, それ以外には何も得ない。

本稿では, エージェントは *semi-honest* に振舞うものとする. つまり, エージェントはプロトコルを常に正しく実行するが, そのエージェントが計算途中で得た情報はすべて保持し, そこから他のエージェントの秘密情報を獲得しようとする.

3. 要素技術

強化学習における操作は主に (1) Q 値の更新, (2) 行動選択, の交互繰り返しである. PPRL では, 互いの知覚を相手には解読できないように暗号でやり取りすることによってプライバシを保護する. Q 値も常に公開鍵暗号によって暗号化され, この暗号化された Q 値と暗号化された知覚を用いて, この二つの操作を実行する.

プロトコルでは加法的準同型性を持つ暗号系を用いる. 加法的準同型性暗号によって暗号化された二つの値は, 秘密鍵の知識を用いずにその和の暗号文を計算することができる。この性質を利用し, 真の Q 値はエージェントに知られることなく, Q 値を暗号化したままで, 通常の強化学習と同様の更新を実行する. 準同型性を利用して実行できない計算については, SFE をプリミティブとして利用する.

3.1 準同型性公開鍵暗号

公開鍵暗号では, 暗号化にはだれもが知ることができる**公開鍵**を用い, 復号化にはメッセージの受信者のみが保持する対応した**秘密鍵**を用いる. 公開鍵と秘密鍵のペア (s_k, p_k) とメッセージ m が与えられた時, m の暗号化を $c = \mathrm{Enc}_{p_k}(m)$, その復号化を $m = \mathrm{Dec}_{s_k}(c)$ と表す. **加法的準同型性暗号系**では, 暗号化された値の和を, 秘密鍵の知識なしに計算できる. つまり, 任意のメッセージ m_1, m_2 について, 以下の式を満足するような秘密鍵の知識を必要としない演算"$\cdot$"が存在する,

$$\mathrm{Enc}_{p_k}(m_1 + m_2) \quad = \quad \mathrm{Enc}_{p_k}(m_1) \cdot \mathrm{Enc}_{p_k}(m_2).$$

これに基づき, ある定数 c と暗号化された値 $\mathrm{Enc}_{p_k}(m_1)$ について, 積 $\mathrm{Enc}_{p_k}(cm_1)$ を計算することができ, これを

$$\mathrm{Enc}_{p_k}(cm_1) = \mathrm{Enc}_{p_k}(m_1)^c$$

と表す. 後に示す実験では, semantically secure な加法的準同型性暗号として [3] を用いる.

3.2 秘匿関数計算による比較と除算

ある秘密の値 $x \in \mathbb{Z}_N$ を A, B が共有することを考える. x を二つのランダムな値 (**ランダムシェア**)$x^A, x^B \in \mathbb{Z}_N$ に分割し, それぞれを A, B が持つことにする. ただし, x^A と x^B は $x \equiv (x_i^A + x_i^B) \mod N$ を満足しつつ $\mathbb{Z}_N$ において一様ランダムに分布しているものとする. 両者がランダムシェアを保持している場合, シェア自体からは x について知ることができないが, 両者が協力することによって x を復元することができるため, これを**秘密分散**と呼ぶ. 提案プロトコルでは, 暗号の準同型性を用いて実行できない比較と除算操作を, ランダムシェアを経由し秘匿関数計算 (SFE) で行う. 秘匿関数計算 (SFE) とは, 二つ以上のエージェントが互いの入力を明かさずに, それらを入力とした任意の関数を評価することができる暗号学的プリミティブである [4, 12].

ランダムシェアの比較: 秘密 $\boldsymbol{x} = (x_1, ..., x_d) \in \mathbb{Z}_N^d$ について, $i^* = \arg\max_i (x_i^A + x_i^B \mod N)$ とする. A と B がランダムシェア $\boldsymbol{x}^A = (x_1^A, ..., x_d^A)$, $\boldsymbol{x}^B = (x_1^B, ..., x_d^B)$ をそれぞれ持っているとき, i^* を SFE によって互いの入力を明かさずに決定する.

ランダムシェアの除算: 秘密 $x \in \mathbb{Z}_N$ について, x を K で除算した商を $Q \in \mathbb{Z}_N$ とする. ただし $x = (QK + R) \mod N, R \in \mathbb{Z}_N (0 \leq R < K)$ である. A と B がシェア $x^A, x^B \in \mathbb{Z}_N$ をそれぞれ持ち, 整数 K を両者が持つときに, SFE によって $Q \equiv Q^A + Q^B \mod N$ なるランダムシェア Q^A, Q^B を, 互いの入力を明かさずに計算する.

4. 時間分割型モデルにおける強化学習

PPRL は Q 値の更新と行動選択の二つから構成される. 前者では Q 値は暗号化され, 更新は主に準同型性を利用して実行される. 後者は, ランダムシェアの比較により実行される. 以降では, 報酬 r_t, 学習率 α および割引率 γ は非負の有理数であることを仮定する. また $\sum_{t=1}^{\infty} (\gamma^t L r_{\max}) < N$ なる整数 L が与えられるものとする. ただし $r_{\max}$ はエージェントが観測可能な最大の報酬である. まず max 操作を含まない SARSA 学習とランダム行動選択による PPRL を次節にて説明する. その後に, Q 学習と (ϵ-)greedy 行動選択による PPRL へと拡張する.

4.1 プライベートな Q 値更新

エージェント A は鍵ペア (p_k^A, s_k^A) を生成する. A および B の公開鍵による $m \in \mathbb{Z}_N$ の暗号化を, 簡単のため, それぞれ $e^A(m), e^B(m)$ と記述する. また $T^A = \{1, ..., t-1\}, T_B = \{t\}$ とする. 期間 T^A では A の知覚のみを用いて, A は通常の更新則を適用し Q 値を計算することができる. Fig. 1 に示すプロトコルは, 時刻 t において B が s_t, a_t, r_t を知覚し, これによって A が持つ Q 値を SARSA 学習によってプライベートに更新するプロトコルである. まず Step 1 で A は全ての (s, a) について $c(s, a) = e^A(Q(s, a))$ を計算し, B へ送信する. 以下, $c(s, a)$ が $Q(s, a)$ の代わりに更新される. 続いて B が行動 a_t を実行, r_t, s_{t+1} を観測し (step 2), a_{t+1} をランダム選択する (step 3). SARSA 学習の更新式は

$$\Delta Q(s_t, a_t) \leftarrow \alpha(r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)), \quad (3)$$
$$Q(s_t, a_t) \leftarrow \Delta Q(s_t, a_t) + Q(s_t, a_t). \quad (4)$$

であり, 両辺を暗号化すると以下を得る.

$$\begin{aligned} c(s_t, a_t) &\leftarrow e^A(\Delta Q(s_t, a_t) + Q(s_t, a_t)), \\ &= e^A(\Delta Q(s_t, a_t)) \cdot e^A(Q(s_t, a_t)) \end{aligned}$$

以下, 上式を以下のように記述する.

$$c(s_t, a_t) \leftarrow \Delta c(s_t, a_t) \cdot c(s_t, a_t). \quad (5)$$

もし $\Delta c(s_t, a_t)$ が B が観測した情報から B によって計算可能ならば, $c(s_t, a_t)$ は eq. 5 より A の助けを得ずに更新可能である. Step 4 ではこの $\Delta c(s_t, a_t)$ を計算する.

前述したように, γ や α など有理数の乗算は準同型性暗号では実行できない. そこで $\alpha\gamma K \in \mathbb{Z}_N$, $Lr_t \in \mathbb{Z}_N$ for all r_t なる K および L を用い, K を eq. 3 の両辺に, L を r_t に, それぞれ乗ずることによって,

$$K\Delta Q(s_t, a_t) \leftarrow K\alpha(Lr_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)),$$

を得る. この計算は $\mathbb{Z}_N$ において閉じており, 準同型性暗号による計算が可能となる. 両辺を A の公開鍵で暗号化することによって,

$$\begin{aligned} &e^A(K\Delta Q(s_t, a_t)) \\ &= e^A(Lr_t)^{\alpha K} \cdot c(s_{t+1}, a_{t+1})^{\alpha\gamma K} \cdot c(s_t, a_t)^{-\alpha K}. \end{aligned} \quad (6)$$

を得る. K, L, α, γ は公開されており, B は r_t, $c(s, a)$ を保持しているため B は $e^A(K\Delta Q(s_t, a_t))$ を eq. 6 により単独で計算可能である (step 4(a)). $c(s_t, a_t)$ を eq. 5 で更新するために, B は $e^A(K\Delta Q(s_t, a_t))$ を K で除算する必要があるが, 除算は準同型性暗号では実行できない. そこで $\Delta Q(s_t, a_t) = K\Delta Q'(s_t, a_T) + R, 0 \leq R < K$ なる $\Delta Q'(s_t, a_t)$ を代わりに用いることにする. これをランダムシェアを系酒した秘匿関数計算による除算によって計算し, B は $e^A(\Delta Q'(s_t, a_t))$ を得る (step 4(b)). 最後に B は以下の更新式を実行する.

$$c(s_t, a_t) \leftarrow e^A(\Delta Q'(s_t, a_t)) \cdot c(s_t, a_t). \quad (7)$$

式 7 は丸め誤差がランダムシェアの除算によって導入される以外は, 式 5 と等価である (step 4(c)). 丸め誤差は L を十分大きくとれば無視できるほど小さくすることができる.

Lemma 1. エージェント A および B が *semi-honest* に振舞うものとする. 時間分割モデルにおける *SARSA* 学習のプライベートな Q 値更新の後, B は暗号化された Q 値を正しく更新するが, その他には何も得ない. A も何も得ない.

- Public input; L, K, learning rate α, discount rate γ
- A's input: $Q(s, a)$ (trained by A during T^A)
- B's input: (s_t, a_t)
- A's output: Nothing
- B's output: Encryption of updated Q value $c(s_t, a_t)$

1. A: Compute $e^A(Q(s, a))$ for all (s, a) and send to B
2. B: Take action a_t and get r_t, s_{t+1}
3. B: Choose a_{t+1} randomly
4. Update Q value
   (a) B: Compute $e^A(K\Delta Q(s_t, a_t))$ by eq. 6
   (b) B: Do private division of $e^A(K\Delta Q(s_t, a_t))$ with A, then B learns $e^A(\Delta Q'(s_t, a_t))$.
   (c) B: Update $c(s_t, a_t)$ by eq. 7.

図 1: 時間分割モデルにおけるプライベートな Q 値更新 (SARSA/ランダム行動選択)

紙数の都合上, 本稿では証明を全て省略し, 直観的な説明にとどめる. Step 4(b) は SFE によって実装されているためセキュアである. Step 4(b) 以外のステップにおいて B が受け取るメッセージはすべて A の公開鍵で暗号化されているため, B はそれらから何も情報を得ることはできない. A は step 4(b) で受け取るメッセージを除き何も受信しない. よってプロトコル全体はセキュアである.

4.2 プライベートな greedy 行動選択

前節で示したプライベートな更新の結果, B は暗号化された更新 Q 値を獲得する. この $Q(s, a)$ は以下の手順によってランダムシェア $Q^A(s, a), Q^B(s, a)$ に変換できる.

1. B: $c(s, a) \leftarrow e^A(Q(s, a)) \cdot e^A(-Q^B(s, a))$ を計算し, A へ送信. ただし $Q^B(s, a) \in_r \mathbb{Z}_N$.
2. A: $Q^A(s, a) \leftarrow d^A(c(s, a))$ を計算.

B が状態 s_t を観測したとき, 以下のプロトコルを実行することで, greedy な行動をプライベートに選択することができる.

1. A and B: 全ての a について $Q(s_t, a) \equiv Q^A(s_t, a) + Q^B(s_t, a) \mod N$ なる Q 値のランダムシェアを計算
2. A and B: $a^* \leftarrow \arg\max_a(Q^A(s_t, a) + Q^B(s_t, a))$ をランダムシェアの比較により計算

このプロトコルは,(1) (ϵ-)greedy 行動選択, (2) Q 学習における更新式の max 操作, (3) 最終的に得た Q 値からの政策の抽出, において直接利用可能である。このプロトコルのセキュリティはほぼ自明なため説明は省略する. またこれらのプロトコルの $m(> 2)$ エージェントへの拡張は容易である.

4.3 PPRL のセキュリティ

SARSA 学習による Q 値のプライベートな更新とランダム行動選択を繰り返し実行することによって, プライベートな SARSA 学習が構成される. そのセキュリティは以下の定理で示される.

Theorem 1. ランダム行動選択とプライペートな Q 値更新による *SARSA* 学習は *Statement 1* においてセキュアである.

またプライベートな (ϵ-)greedy 行動選択とプライベートな Q 値の更新を交互に実行することによって, プライベートな SARSA 学習・Q 学習を構成することができる. しかしながら, これらのアルゴリズムは Statement 1 を満足しない. なぜならば Statement 1 はエージェントが学習途中の greedy 行動を

得ることを許容しないからである. そのため, 以下の緩和された問題を考える.

Statement 2. 時間分割モデルにおける i 番目のエージェントの入力を H^i とする. *PPRL* の実行後, 全エージェントは学習途中の *greedy* 行動, 政策 π^c と等価な政策 π および, π から i 番目のエージェントが推測可能な情報を獲得するが, それ以外には何も得ない。

この Statement に基づき, セキュリティに関する同様の定理が示される.

Theorem 2. *(ϵ-)greedy* 行動選択とプライベートな Q 値更新による *SARSA* 学習*/Q* 学習は *Statement 2* においてセキュアである.

これらの証明は Lemma 1 に基づき SFE の安全性を証明する標準的な手続きに従って示すことが可能である.

5. Experiments

PPRL の効率性を示すための計算機実験を行った. 実験は 1.2 GHz CPU, 2GB RAM の計算機上の LINUX および java 1.5.0 を用いた. 暗号は [3], 鍵長は 1024-bit とした. SFE の実装には Fairplay [6] を用いた. 計算時間においては, 暗号および の SFE の計算時間が支配的であるため, 実験結果に通信時間は含まれていない.

時間分割モデルにおける一次元ランダムウォーク問題を考える. 状態空間を $S = \{s_1, ..., s_n\}(n = 40)$ とし, 行動空間は $A = \{a_1, a_2\}$ とした. 初期状態とゴール状態は s_1 and s_n である. 行動 a_1 が $s_p(p \neq n)$ で選択されたとき, エージェントは s_{p+1} に移動する. 行動 a_2 が状態 $s_p(p \neq 1)$ で選択された場合, エージェントは s_{p-1} に移動し, $p = 1$ ではエージェントは移動しない. 行動 a_1 が状態 s_{n-1} で選択された場合のみ報酬 $r = 1$ が与えられエピソードが打ち切られる. それ以外では $r = 0$ が与えられる.

エージェント A が 15000 ステップの知覚を得て, その後エージェント B が 15000 ステップの知覚を得るものとする. 比較手法は CRL(全知覚を統合して 1 エージェントが学習), IDRL(エージェント間の情報交換無しで学習), PPRL(提案法), および SFE(全操作を SFE で実装) である. 学習手法は, 全設定ともにランダムあるいは ϵ-greedy 行動選択と SARSA 学習の組み合わせである. ゴール状態到達に要するステップ数 (30 試行平均), ゴール状態到達に成功した試行数と計算に要した時間, 各設定のプライバシー保護の状況を表 1 に示す. CRL は最適政策を獲得するが, プライバシは全く保護されない. IDRL は通信を行わないため, プライバシーは完全に保護されるが, 知覚を共有しないため学習精度が劣る. PPRL, SFE は CRL 同様学習の最適性が保障されており CRL と同様の学習精度を達成するが,SFE の計算コストは極めて高い. PPRL の計算コストも比較的高いものの, IDRL と同じプライバシー保護レベルを達成しつつ, 現実的な計算時間内で最適政策の獲得に成功していることがわかる.

表 1: ランダムウォーク問題における比較

	comp. (sec)	accuracy avg.	accuracy #goal	privacy loss
CRL/rnd.	0.901	40.0	30/30	disclosed all
IDRL/rnd.	0.457	247	8/30	Stmt. 1
PPRL/rnd.	4.71×10^3	40.0	30/30	Stmt. 1
SFE/rnd.	$> 7.0 \times 10^6$	40.0	30/30	Stmt. 1
CRL/ϵ-grd.	0.946	40.0	30/30	disclosed all
IDRL/ϵ-grd.	0.481	—	0/30	Stmt. 2
PPRL/ϵ-grd.	3.36×10^4	40.0	30/30	Stmt. 2
SFE/ϵ-grd.	$> 7.0 \times 10^6$	40.0	30/30	Stmt. 2

6. Conclusion

本稿では時間分割モデルにおける分散秘密情報源からの強化学習法を提案した. またプライバシを保護した SARSA 学習および Q 学習において, 最適性とプライバシ保護が両立できることを示した. この研究では最大化すべき報酬関数がエージェント間で一致する設定を扱ったが, エージェントによって追及する利益が異なるようなヘテロジニアスな環境におけるプライバシを保護した強化学習への拡張が今後の課題である.

参考文献

[1] N. Abe, N. Verma, C. Apte, and R. Schroko. Cross channel optimized marketing by reinforcement learning. In *Proc. of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 767–772. ACM Press New York, NY, USA, 2004.

[2] R. Cogill, M. Rotkowitz, B. Van Roy, and S. Lall. An Approximate Dynamic Programming Approach to Decentralized Control of Stochastic Systems. In *Lecture Notes in Control and Information Sciences*, volume 329, pages 243–256. Springer, 2006.

[3] I. Dåmgard and M. Jurik. A Generalisation, a Simplification and Some Applications of Paillier's Probabilistic Public-Key System. In *Proc. of Public Key Cryptography 2001*. Springer, 2001.

[4] O. Goldreich. *Foundations of Cryptography: Volume 2, Basic Applications*. Cambridge University Press, 2004.

[5] Y. Lindell and B. Pinkas. Privacy Preserving Data Mining. *Journal of Cryptology*, 15(3):177–206, 2002.

[6] D. Malkhi, N. Nisan, B. Pinkas, and Y. Sella. Fairplay: a secure two-party computation system. In *Proc. of the 13th USENIX Security Symposium*, pages 287–302, 2004.

[7] C. C. Moallemi and B. Van Roy. Distributed optimization in adaptive networks. In *NIPS 16*, volume 16. MIT Press, 2004.

[8] J. Sakuma and S. Kobayashi. Large-scale k-means clustering with User-centric Privacy-Preservation. In *Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. Springer, to appear, 2008.

[9] J. Sakuma, S. Kobayashi, and R. Wright. Privacy-preserving reinforcement learning. In *Proc. of International Conference on Machine Learning*, to appear, 2008.

[10] J. Schneider, W.K. Wong, A. Moore, and M. Riedmiller. Distributed value functions. In *Proc. of International Conference on Machine Learning*, pages 371–378, 1999.

[11] C.J.C.H. Watkins. *Learning from Delayed Rewards*. Cambridge University, 1989.

[12] A. C.-C. Yao. How to generate and exchange secrets. In *Proc. of the 27th IEEE Symposium on Foundations of Computer Science*, pages 162–167, 1986.

[13] H. Yu, X. Jiang, and J. Vaidya. Privacy-preserving SVM using nonlinear kernels on horizontally partitioned data. In *Proc. of the 2006 ACM Symposium on Applied Computing*, pages 603–610. ACM Press New York, NY, USA, 2006.

[14] S. Zhang and F. Makedon. Privacy preserving learning in negotiation. In *Proc. of the 2005 ACM Symposium on Applied Computing*, pages 821–825. ACM Press New York, NY, USA, 2005.