CF-008

不完全情報展開型ゲームの求解における利得摂動に関する研究

眞坂 航宙 * 坂本 充生 † 阿部 拳之 † 蟻生 開人 † 岩崎 敦 *
Wataru Masaka Mitsuki Sakamoto Kenshi Abe Kaito Ariu Atsushi Iwasaki

1 はじめに

本研究では，不完全情報二人零和展開型ゲームの求解において有効である，利得に摂動を加える手法について吟味する．相手の情報が見えない意思決定問題は不完全情報ゲームとして定式化できる．さらに，プレイヤの意思決定に順番を持たせる展開型ゲームとして定式化することにより，先手や後手といった手番の概念を導入でき，現実の社会状況をより正確にモデル化することができる．このような状況では，各プレイヤが互いに最適な判断を下している状態である，均衡解という概念が重要になる．均衡解は各プレイヤが互いの最適な戦略を反映した結果として安定する状態を示し，この状態ではいかなるプレイヤも一方的な戦略変更による利益を上げることができなくなる．このため，均衡解の求解は市場競争，国際政治，交渉戦略など，実社会における多様な意思決定プロセスの解析において重要な役割を果たす．しかし，不完全情報下での均衡解の求解は困難な課題である．各プレイヤは観測できない情報を考慮しながら確率的な行動を選択する必要があるため，収束する過程で多くの確率的な変動が影響してしまい，学習が困難となるためである．

近年，展開型ゲームの均衡を近似的に計算する手法が発展し，AI が不完全情報ゲームにおいて人間を上回る事例が増えている．ポーカーでは，DeepStack [10] が一対一で初めてプロプレイヤに勝利し，その後 Libratus [3] が複数人のトッププロに対して圧倒的な強さを示した．また，戦略的要素の強いストラテゴにおいては，DeepMind が開発した *DeepNash* がトップレベルの人間プレイヤと同等以上の実力を示した [13].

本研究では累積損失を元に戦略を逐次更新するオンライン学習アルゴリズムである，Follow the Regularized Leader（FTRL）[9] に注目する．このアルゴリズムは正則化項を導入することで戦略の安定化を図り，様々なゲームへの適用が進んでいる．しかし，FTRL で得られる戦略は必ずしもナッシュ均衡 [11] に直接収束せず，均衡戦略を近似するためには逐次更新して得られる戦略の時間平均を求める必要がある．そこで，ゲームで得られる利得に適切な摂動を加えて戦略を均衡へ促す手法が近年研究されており [12]，このように均衡に直接収束する性質は終極反復収束と呼ばれる．

一方で，膨大な状態数を持つ展開型ゲームでは，ゲーム木の全探索が難しいため一部の履歴をサンプリングをし，期待利得を推定して計算量を削減するアプローチ [6] が取られるが，推定に伴う分散が学習を不安定化させる．さらに利得を摂動させると，ゲームの元の利得だけでなく，利得摂動についても，その期待値を推定する必要が生じる．この枠組みに適した利得の摂動方法は未だに明らかになっていない．

そこで本研究では摂動の推定時の分散を低減させる利得の摂動方法を提案する．Reverse Kullback-Leibler 距離（RKL）を使った摂動を用いることで，学習で必要な反実仮想価値がサンプリング下でも不偏であり，また摂動部分の推定量の分散が 0 になることを理論的に示す．計算機実験により，提案手法は Leduc poker というゲームの構造が非対称であるゲームにおいて特に有効であることを示し，効率的な学習を可能にする摂動手法を探る．

2 不完全情報二人零和展開型ゲーム

不完全情報二人零和展開型ゲームは $G := \langle N, H, Z, A, P, \pi_c, u, \mathcal{X} \rangle$ の組で定義できる．プレイヤを $i \in N = \{1, 2\}$ で表し，ゲーム内の確率的な要素や偶然の要因をチャンスプレイヤという仮想のプレイヤが行うものとみなし c で表す．チャンスプレイヤも含めた i 以外のプレイヤをまとめて $-i$ で表す．ゲームの進行は決定点の系列として捉えることができ，ある時点までのゲームの進行を表す概念として履歴 $h \in H$ が用いられる．履歴 h は初期状態からの行動の列で定義される．各履歴において，プレイヤ関数 $\tau : H \setminus Z \to N \cup \{c\}$ によって定まるプレイヤは次の行動 $a \in A(h)$ を選択する．ここで Z は終端履歴というゲームが終わった状態を表し，どのプレイヤも選択できる行動を持たない．利得関数を $u : H \times A \to \mathbb{R}$ で表し，展開型ゲームでは終端以外の履歴 $ha \neq z \in Z$ において $u_i(h, a) = 0$ となる．つまり終端でのみ各プレイヤは利得を得る．さらに，二人零和の仮定より $u_1(z) = -u_2(z)$ となる．

情報の不完全性からプレイヤには区別できない履歴の集合が存在し，これを情報集合 $x(h) \in X$ と呼ぶ．情報分割を $\mathcal{X} = \{X_i\}_{i \in N}$ とし，プレイヤ i は実際の履歴 h は観測できず，対応する情報集合 $x(h) \in X_i$ のみを観測する．各プレイヤは観測した情報集合ごとに戦略 $\pi_i(\cdot|x) \in \Delta(A(x))$ を持つ．戦略の組を戦略プロファイルと呼び，$\pi = (\pi_i, \pi_{-i})$ とする．チャンスプレイヤ c の行動の選択確率は事前に π_c で与えられる．

展開型ゲームは木構造で表すことができる．具体例として

* 電気通信大学, The University of Electro-Communications
† サイバーエージェント, CyberAgent

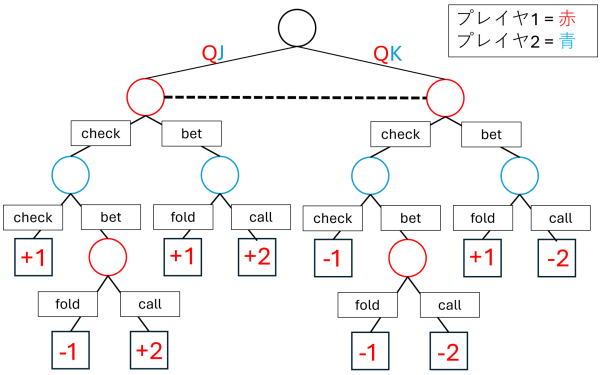

図 1: 展開型ゲームの例: Kuhn poker の一部

Kuhn poker [4] について説明する．Kuhn poker は 3 枚のカード（J，Q，K）のみを使用する単純化されたポーカーで，参加費として 1 コインを支払い，各プレイヤはベット，コールする際にさらに 1 コインをかけることができる．図 1 はプレイヤ 1 に Q が配られる状況についてのゲーム木で，利得はプレイヤ 1 にとっての利得を表している．二人零和ゲームでは二人の利得の合計が 0 となるため，プレイヤ 2 が受け取る利得は図に表されている利得の-1 倍となる．プレイヤ 1 はプレイヤ 2 の手札が J，K どちらか知らないままに意思決定をする必要がある．例えば，プレイヤ 1 に Q が配られた場合の最初の意思決定時には，区別できない $h_1 = \mathrm{QJ}$ と $h_2 = \mathrm{QK}$ という二つの履歴が存在し，これらが同じ情報集合 x に属し，$h_1, h_2 \in x$ と表される．

戦略 π の下での履歴 h への到達確率 $\rho^\pi(h)$ は次で表される：

$$\rho^\pi(h) = \prod_{(h'a') \sqsubseteq h} \pi_{\tau(h')}(a'|x(h')).$$

到達確率 $\rho^\pi(h)$ は全ての履歴 $h \in H$ で $\rho_i^\pi(h)\rho_{-i}^\pi(h)$ とプレイヤごとの到達確率の積に分解できる．ある履歴 h から h' への到達確率は次で与えられる．

$$\rho^\pi(h, h') = \begin{cases} \frac{\rho^\pi(h')}{\rho^\pi(h)} & \text{if } \rho^\pi(h) > 0 \text{ and } h \sqsubseteq h', \\ 0 & \text{otherwise.} \end{cases}.$$

戦略プロファイル π が与えられたとき，プレイヤ i の期待利得は次で定義される：

$$u_i(\pi) = \sum_{h \in H \setminus Z} \sum_{a \in A(h)} \rho^\pi(ha) u_i(h, a).$$

プレイヤ i が履歴 h で行動 a を選択したときの行動価値を，

$$q_i^\pi(h, a) = \sum_{h'a' \sqsupseteq ha} \rho^\pi(ha, h'a') u_i(h', a'),$$

と定義する．情報集合 x における反実仮想価値は $v_i^\pi(x, a)$ は，x に含まれる履歴 $h \in x$ において，相手の到達確率 $\rho_i^\pi(h)$ を重みとして行動価値を合計したものである：

$$v_i^\pi(x, a) = \sum_{h \in x} \rho_{-i}^\pi(h) q_i^\pi(h, a). \tag{1}$$

全プレイヤについて，相手の戦略を固定した場合に自身の戦略を変更する誘引を持たない，つまり期待利得が増加しない状態を均衡と呼び，展開型ゲームにおける均衡戦略 $\pi^* = (\pi_1^*, \pi_2^*)$ を，プレイヤ i の戦略空間は Σ_i として，

$$\forall \pi_1 \in \Sigma_1, \forall \pi_2 \in \Sigma_2, \quad u_1(\pi_1^*, \pi_2) \geq u_1(\pi_1^*, \pi_2^*) \geq u_1(\pi_1, \pi_2^*),$$

と定義する．この均衡戦略への近さを示す指標として，

$$\mathrm{exploit}(\pi) := \max_{\tilde{\pi}_1 \in \Sigma_1} u_1(\tilde{\pi}_1, \pi_2) + \max_{\tilde{\pi}_2 \in \Sigma_2} u_2(\pi_1, \tilde{\pi}_2), \tag{2}$$

で定義される Exploitability と呼ばれる値を用いる．Exploitability が 0 に近いほど，戦略 π は均衡に近いと言える．

3 Follow the Regularized Leader

3.1 学習手順

均衡戦略を近似する方法として，オンライン学習で逐次的に戦略を更新する方法がある．具体的には，以下の過程を T 回反復する．

1. 各反復 $t \geq 1$ で，各プレイヤは今までに観測した反実仮想価値 $\tilde{v}_i$ に基づいて戦略 π_i^t を決定する．
2. 各プレイヤは反実仮想価値 $v_i^{\pi^t}$ を計算する．

展開型ゲームでは戦略を更新するために FTRL [9] などのオンライン学習アルゴリズムを用いる．FTRL は時刻 $t+1$ の戦略 π_i^{t+1} を次のように決定する：

$$\pi_i^{t+1}(\cdot|x) = \operatorname*{arg\,max}_{\pi \in \Delta(A(x))} \left\{ \eta \left\langle \sum_{s=1}^{t} v_i^{\pi^s}(x, \cdot), \pi \right\rangle - \psi_i(\pi) \right\}.$$

ここで，η は学習率で $\psi_i : \Delta(A(x)) \to \mathbb{R}$ は強凸正則化関数とする．本研究ではエントロピー正則化 $\psi_i(\pi(\cdot|x)) = \sum_{a \in A(x)} \pi(a|x) \ln \pi(a|x)$ を用いる．FTRL で戦略を更新しても，戦略 π^t は均衡に直接収束しないことが知られており，ナッシュ均衡解を近似するためには π^t の時間平均 $\bar{\pi}^t$ を求めなければならない [9]．ある情報集合 x における戦略の時間平均は一般に次のように与えられる：

$$\bar{\pi}^t(x) = \frac{\sum_{s=1}^{t} \rho_i^{\pi^s}(x) \pi^s(x)}{\sum_{s=1}^{t} \rho_i^{\pi^s}(x)}. \tag{3}$$

これは，単純に時間平均を取るのではなく，各時刻 s での到達確率 $\rho_i^s(x)$ で重み付けをして平均を取っている．

3.2 反実仮想価値の推定

展開型ゲームにおける反実仮想価値を正確に計算するには，ゲーム木を完全に探索して全ての終端履歴を評価する必要がある．しかし，状態数が膨大になる場合は計算量が指数的に増大し，現実的ではない．そこで，各反復で終端履歴 Z の一部分をサンプリングし，反実仮想価値を推定して学習に用いる手法が用いられる．

本節ではまず一般的なサンプリングによる推定の枠組みを導入し，次に本研究で用いた Outcome Sampling を紹介する．

3.2.1 サンプリングの枠組みの導入

終端履歴集合 Z の部分集合の集合を $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_k\}$，$Y_j \subseteq Z$ と定義し，さらに $\bigcup_{j=1}^{k} Y_j = Z$ とする．これにより，$\mathcal{Y}$ は Z 全体をカバーする．アルゴリズムの各反復において，各終端履歴の集合 Y_j を確率 p_j に従ってサンプリングし，得られたサンプル Y_j から反実仮想価値を推定することになる．

このサンプリングの枠組みは CFR の文脈で発展したものであり，モンテカルロ木探索を利用した MCCFR（Monte-Carlo CFR）[6] として提案されたものである．MCCFR は CFR の不偏推定を行いつつ，CFR が持つ収束性などの望ましい性質を保持することが示されている [6].

Lanctot ら [6] は，モンテカルロ的に反実仮想価値を推定する手法として，External Sampling と Outcome Sampling の二種類のサンプリング手法を定義した．External Sampling は，対戦相手（およびチャンスプレイヤ）のみの選択をサンプリングし，プレイヤ自身の各行動に対して部分木を再帰的に全探索することで推定を行う．一方，Outcome Sampling は最も極端なサンプリング手法であり，一つの終端履歴のみをサンプリングする．つまり全ての $Y_j \in \mathcal{Y}$ がただ一つの終端履歴 z からなることを意味する ($|Y_j| = 1, \forall Y_j \in \mathcal{Y}$)．このため，External Sampling に比べて計算量は大幅に削減されるが，推定誤差は大きくなるという特徴がある．また，Outcome Sampling は唯一のモデルフリーな MCCFR であり，強化学習の標準的反復（環境からの経験のみで学習）と完全に整合する手法である．

3.2.2 サンプリング下における推定

3.1節では各プレイヤは木を全探索できたため，正確な反実仮想価値 $v_i^{\pi^t}$ を求められたが，サンプリング下においてはサンプリングした終端履歴の部分集合 Y_j から反実仮想価値を推定する必要がある．そこで，Y_j に含まれる履歴を含有する情報集合 x における，式 (1) の反実仮想価値の不偏推定量を考える．Y_j に含まれる履歴の集合を $H_j = \{h \in H \mid h \sqsubseteq z \wedge z \in Y_j\}$ とし，履歴 h をサンプリングする確率を $p(h) = \sum_{j:H_j \ni h} p_j$ とする．この時，$\frac{p(h')}{p(h)} = p(h, h')$ のようにして h から h' までの到達確率を，サンプリングする確率 p_j を用いて表すことができる．以上より，反実仮想価値の不偏推定量は次式で求められる：

$$\tilde{v}_i^{\pi}(x, a) = \sum_{h \in x} \frac{\rho_{-i}^{\pi}(h)}{p(h)} \tilde{q}_i(h, a), \tag{4}$$

$$\tilde{q}_i^{\pi}(h, a) = \sum_{h'a' \sqsupseteq ha \wedge h'a' \in H_j} \frac{\rho^{\pi}(ha, h'a')}{p(h, h'a')} u_i(h', a'). \tag{5}$$

3.2.3 本研究で用いたサンプリング手法

本研究では，各反復で終端履歴 $z \in Z$ を一つだけモンテカルロサンプリングし推定を行う Outcome Sampling [6] を採用し用いる．

終端履歴 z のサンプリングは，反実仮想価値を計算するプレイヤ i のサンプリング戦略，

$$\pi_i'(\cdot|x) = (1 - \epsilon)\pi_i^t(\cdot|x) + \frac{\epsilon}{|A(x)|}, \tag{6}$$

に従って行動を選択することで行われ，自分以外のプレイヤ $-i$ は現在の戦略 π_{-i}^t と事前に与えられた π_c に従って行動を選択する．ここで，$0 \leq \epsilon \leq 1$ は探索と活用のバランスを調整するためのパラメータであり，この式によるサンプリング法は ϵ-greedy と呼ばれる．なお，McAleer らの先行研究によって [8]，サンプリング戦略を反復間で固定したまま用いると，ϵ-greedy でサンプリングを行うよりも性能が向上することが示されている．また，固定サンプリング戦略として一様分布を用いる（すなわち式 (6) において $\epsilon = 1$ とする）のが有効であることも示されている．本研究ではこれらの先行研究に従い，一様分布の固定サンプリング戦略を採用する．

4 提案手法

本研究では，まず FTRL に利得の摂動 d を加えて収束を促す手法 **Perturbed FTRL（PFTRL）** を導入する．なお，Outcome Sampling 下でのアルゴリズムは **OS-PFTRL** と呼ぶことにする．

まず摂動を定義するために参照戦略 σ を導入する．参照戦略とは，ある情報集合 x について，訪問された回数が $T_\sigma = T$ に達する毎に定期的に現在の戦略に置き換えて更新を行う．こうすることで現在更新している戦略を均衡解に近づけることができる [1].

摂動 d は参照戦略 σ との距離関数で定義され，摂動を加えると学習が安定することがわかっている [12]．サンプリング下の PFTRL における，摂動を加えた**摂動反実仮想価値**の推定量は次で与えられる：

$$\tilde{v}_i^{\pi,\sigma}(x, a) = \sum_{h \in x} \frac{\rho_{-i}^{\pi}(h)}{p(h)} \left(\tilde{q}_i^{\pi}(h, a) + \mu \tilde{\delta}_i^{\pi,\sigma}(h, a) \right), \tag{7}$$

$$\tilde{\delta}_i^{\pi,\sigma}(h, a) = \mathbb{1}_{h \in H_j} d_i^{\pi,\sigma}(h, a) + \sum_{h' \sqsupseteq ha \wedge h' \in H_j} \sum_{a' \in A(h')} \frac{\rho^{\pi}(ha, h'a')}{p(h, h')} d_i^{\pi,\sigma}(h', a'). \tag{8}$$

ここで，$\tilde{\delta}_i^{\pi,\sigma}$ を**累積摂動**と呼び，先の状態の摂動も含めた累積摂動の期待値を表す．μ は摂動強度と呼ばれる摂動の影響の強さを調整するパラメータである．式 7から摂動反実仮想価値 $\tilde{v}_i^{\pi,\sigma}(x, a)$ の推定量は，行動価値 $\tilde{q}_i^{\pi}(h, a)$ の推定量と累積摂動の期待値 $\tilde{\delta}_i^{\pi,\sigma}$ の推定量に分けて考えることができる．摂動反実仮想価値を基に戦略を更新していく流れは付録のアルゴリズム 1に示す．Outcome Sampling 下における摂動反実仮想価値の推定の流れは付録のアルゴリズム 2に示す．

摂動 d には様々な距離関数を用いることができ，現在の戦略 π と参照戦略 σ との Kullback-Leibler 距離（KL），

$$d_i^{\pi,\sigma}(h, a) = \mathbb{1}[i = \tau(h)] \log \frac{\sigma_i(a|x(h))}{\pi_i(a|x(h))},$$

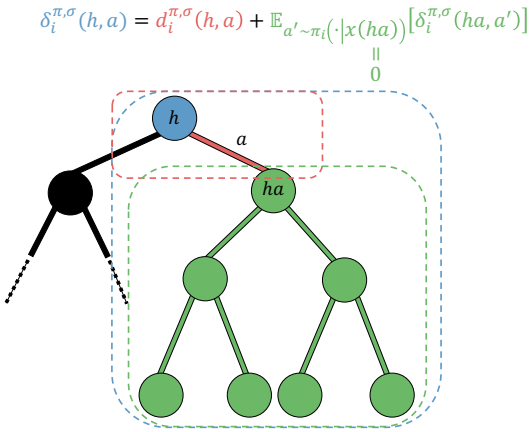

図 2: RKL による利得摂動の図．緑の部分が先の状態の累積摂動を表し，期待値が必ず 0 となる．

を用いた場合，*DeepNash* に使用されている既存手法の Reward Transformed FTRL と等しくなる [12]．本研究では，新たに Reverse KL 距離（RKL），

$$d_i^{\pi,\sigma}(h,a) = \frac{\mathbb{1}[i=\tau(h)]}{\pi_i(a|x(h))}(\sigma_i(a|x(h)) - \pi_i(a|x(h))),$$

を用いた摂動方法を提案する．RKL による摂動はサンプリング下において，次の 2 つの望ましい性質があることを証明した．なお，証明は付録に示した．

定理 4.1. 任意の $i \in N, x \in X_i, a \in A(x)$ において，*RKL* による摂動反実仮想価値 $\tilde{v}_i^{\pi,\sigma}$ は不偏推定量である：

$$\mathbb{E}_{j\sim p_j}[\tilde{v}_i^{\pi,\sigma}(x,a)] = v_i^{\pi,\sigma}(x,a)$$

定理 4.2. Y_j がサンプリングされたとき，任意の $h \in H_j, a \in A(h)$ において，*RKL* による累積摂動の推定量の分散は 0 である：

$$\mathrm{Var}_{j\sim p_j}[\tilde{\delta}_i^{\pi,\sigma}(h,a) \mid h \in H_j] = 0.$$

この定理は，RKL による累積摂動の期待値の推定量 $\tilde{\delta}_i^{\pi,\sigma}(h,a)$ をサンプリング下でも正確に求めることが出来ることを意味している．これは図 2のように，任意の h, π において累積摂動の期待値が 0 になる，という RKL の独自な特徴によるものである．よって現在の情報集合と行動から決まる摂動さえ考慮すれば良いことを意味する：

$$\mathbb{E}_{a\sim\pi_i(\cdot|x(h))}[\delta_i^{\pi,\sigma}(h,a)] = 0. \tag{9}$$

5 計算機実験

5.1 ベンチマークによる性能比較

提案手法の有効性を検証するため，Kuhn poker，Leduc poker，Goofspiel，Liar's dice という 4 つの不完全情報ゲームにおけるベンチマークを用いて計算機実験を行った．各ゲームにおける情報集合の数は，Kuhn poker：12，Leduc poker：936，

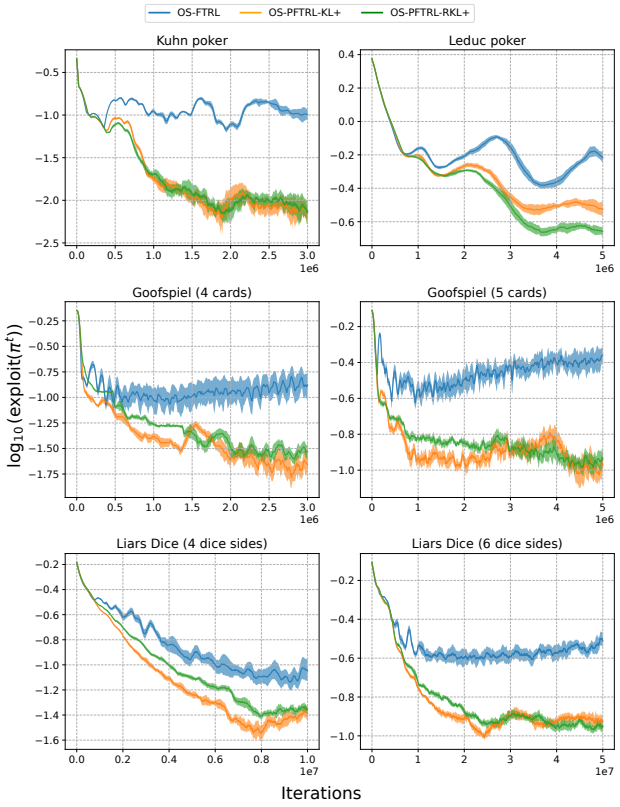

図 3: Outcome Sampling 下での現在の戦略の $\mathrm{exploit}(\pi^t)$

Goofspeil（4 枚のカード）：162，Goofspiel（5 枚のカード）：2,124，Liar's dice（4 面サイコロ）：1,024，Liar's dice（6 面サイコロ）：24,576，である．本研究では OpenSpiel [5] という強化学習フレームワークを用いて実験を行った．各ゲームにおけるアルゴリズムの性能は 2 章の式 (2) で定義した，均衡からの乖離度を示す指標である Exploitability を用いて評価する．

図 3と図 4に実験結果を示す．横軸は反復回数 (Iterations)，縦軸は Exploitability の対数（底は 10）を表す．比較対象としたアルゴリズムは次の 3 つである．

- OS-FTRL: 利得摂動を用いないベースライン．
- OS-PFTRL-KL+: 既存手法．利得摂動に KL 距離を使用．
- OS-PFTRL-RKL+: 提案手法．利得摂動に RKL 距離を使用．

末尾の + は参照戦略更新を用いていることを表している．実験では全てのゲームで，学習率を $\eta = 0.0001$，摂動強度を $\mu = 0.1$ とし，サンプリング戦略には前述の通り一様戦略を用いた．また，参照戦略更新の間隔は $T_\sigma = 100,000$ と設定している．反復回数はゲームごとに変更している．これは 3 つのアルゴリズム全てで現在の戦略 π^t が収束している点を用いているためである．図 3は各アルゴリズムによって求められた t 期の戦略 π^t の Exploitability の推移を表し，図 4は各アルゴリズムで求められた戦略を式 3で時間平均を取った平均戦略 $\bar{\pi}^t$

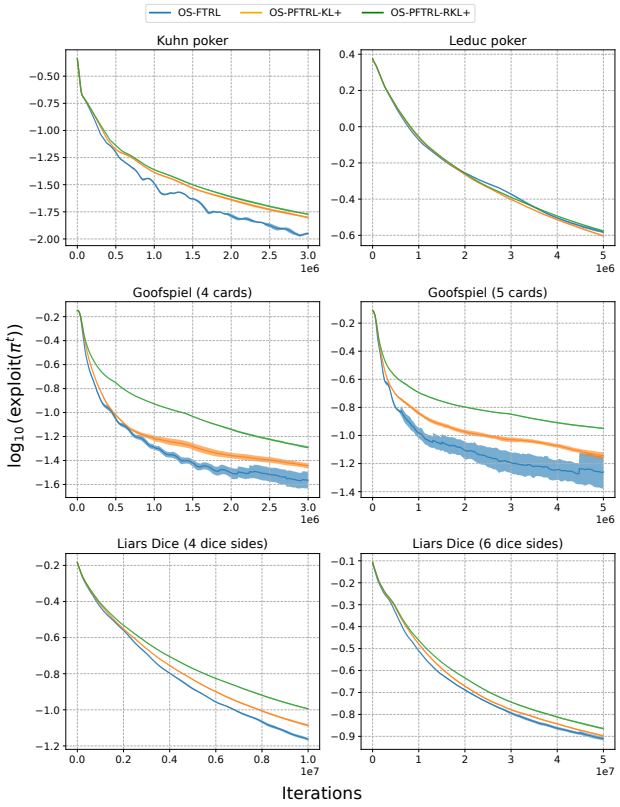

図 4: Outcome Sampling 下での平均戦略の $\mathrm{exploit}(\bar{\pi}^t)$

における Exploitability の推移を表している．また，サンプリングによる性能のばらつきを抑えて比較するため，各アルゴリズムを異なるシード値で 10 回ずつ実行した．各プロットの中心に見える実線はこれら 10 回の試行における Exploitability の平均値を示し，その周囲の半透明帯は平均値± 2 ×標準誤差（n = 10）を表している．

図 3の結果から，利得摂動を加えた PFTRL (OS-PFTRL-KL+，OS-PFTRL-RKL+) は，摂動なしの OS-FTRL と比較して，全てのゲームにおいてより均衡に近い Exploitability を達成していることがわかる．これは，利得摂動が学習の効率を高め，戦略を均衡により近く収束させる効果があることを示している．また，提案手法である RKL+ と既存手法である KL+ を比較すると，特に Leduc poker において，提案手法がより低い Exploitability に到達した．Leduc poker はゲーム途中で降りることが可能なことにより非対称なゲーム構造を持ち，これが RKL 距離を用いた摂動が有効に働いた原因だと考えられる．一方，他のゲームにおいては両手法に顕著な性能差は見られなかった．これらの結果は，提案する RKL 距離を用いた利得摂動が，不完全情報展開型ゲームの求解において，特に特定のゲーム構造において有効な手法となり得ることを示唆している．

図 4の結果から，t 期の戦略の場合とは対照的に，OS-FTRL が一般に KL+ および RKL+ を上回り，ほとんどのゲームで最も低い Exploitability を達成している．これは，摂動を導入することが平均戦略における Exploitability を改善することはなく，むしろ悪化させてしまうことを示唆している．また，平均戦略の結果を KL と RKL で比較すると，すべてのゲームにおいて KL が RKL を上回る性能を示している．これらの結果から，RKL による摂動と戦略の時間平均化は，特に相性の悪い組み合わせであると考えられる．

5.2 KL と RKL のより詳細な比較

5.1節の結果からは RKL を摂動関数に導入したことの効果がわかりにくい．そこで，摂動の強さを調節し，参照戦略更新の影響を取り除くことで，摂動関数として用いる距離関数 (KL と RKL) の影響の違いを純粋に比較検証した．

実験の手順は次の通りである．まず完全ゲーム木探索下で PFTRL-RKL (末尾に + がついていないため, 参照戦略更新を行なっていないことに注意) を摂動強度 μ を $\mu \in \{0.05, 0, 1, 0.2\}$ で固定して実行し，収束時の Exploitability（図5のオレンジ破線）を求めた．次に同等の Exploitability が得られる PFTRL-KL の摂動強度を調べた．これにより KL と RKL で同じ Exploitability を達成する摂動強度の組が分かったので，これを用いて Outcome Sampling 下で実験を行った．その結果が図5であり，Outcome Sampling 下で，OS-PFTRL-RKL（$\mu \in \{0.05, 0, 1, 0.2\}$）を用いた場合と，それぞれに対応する KL 摂動（$\mu \in \{0.10147, 0, 17, 0.2946\}$）を用いた OS-PFTRL-KL の Exploitability 推移を示している．横軸は反復回数，縦軸は対数スケールの Exploitability である．

- RKL ($\mu = 0.05$) vs. KL ($\mu = 0.10147$)：図5の左図を参照．両手法ともほぼ同等の Exploitability に収束し，性能差は小さい．
- RKL ($\mu = 0.1$) vs. KL ($\mu = 0.17$)：図5の中央図を参照．両者とも約–1.419 付近で収束するが，OS-PFTRL-RKL が若干優位である．
- RKL ($\mu = 0.2$) vs. KL ($\mu = 0.2946$)：図5の右図を参照．高い摂動強度において RKL が KL を上回り，50 万反復以降により低い Exploitability を達成している．

この結果は，摂動強度 μ が高いほど RKL が KL と比較してより効果的に Exploitability を低減できる可能性を示唆している．一般に，反実仮想価値の推定量の分散が大きくなるほど Exploitability は大きくなってしまうことが知られており，OS-PFTRL-RKL が OS-PFTRL-KL を上回ったのは定理4.2で示された RKL の零分散の性質の効果を示すものであると考えられる．

一方で図 3で Leduc poker 以外のゲームで RKL と KL の間に大きな差が見られなかったのは，参照戦略更新 (‘+’ 付きの手法) の効果によるものである．つまり，現在の戦略 π^t と参照戦略 σ の距離が参照戦略の更新ごとに小さくなることで [1]，結果的に累積摂動の推定誤差が小さくなっていき，距離関数ごとの分散の大小による影響が抑えられたためだと考えられる．

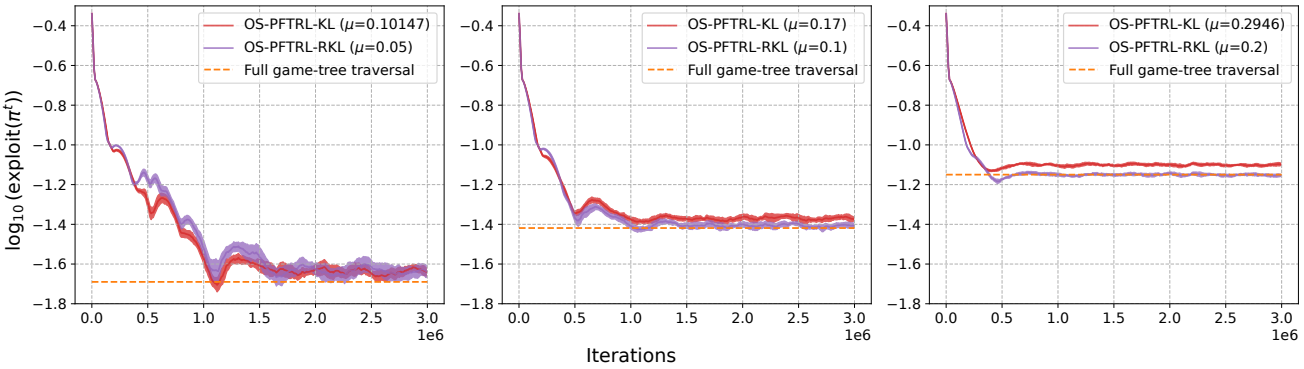

図 5: Exploitability difference of last-iterate between OS-PFTRL-RKL and -KL in Kuhn poker, with varying tuned perturbation strengths.

6 考察

本研究で触れられなかった最新のアルゴリズムと摂動の組み合わせについて考察する．Lee らが終局反復収束を達成するために提案した楽観的手法（Optimistic variants）[7] を，サンプリング下で有効に実行するのは容易ではない．Optimistic FTRL や Dilated OMWU[7] のように単純に“楽観性”を導入するだけでは，報酬（または勾配）の推定誤差やノイズによって予測ベクトルが不安定になってしまうため不十分である [2]．それでも，サンプリング下で楽観的手法を実現可能にするために，摂動の導入が有効かどうかを探ることは有望な今後の研究テーマである．

もうひとつの重要な研究方向としては，現在最先端を走るモダンな CFR の派生手法と摂動の組み合わせである．Zhang らの手法 [14] は CFR+ に摂動を組み込んでいるものと解釈ができ，完全ゲーム木探索の設定下における SOTA を達成している．しかし，サンプリング下では必ずしも PFTRL を上回るとは限らず，性能が不安定となることが実験で確かめられている．

7 おわりに

本研究では，不完全情報展開型ゲームのサンプリング下での学習において，利得摂動が FTRL アルゴリズムにどのような影響を与えるかを吟味した．利得摂動は終極反復収束の意味で一貫して性能を改善することを実験的に示した．さらに，我々が提案した PFTRL-RKL+ は，既存手法と比較して反実仮想価値の推定量の分散を小さくすることを理論的に示し，計算機実験により特定のゲームにおいては提案手法が既存手法を上回ることを示した．今後の課題として，PFTRL-RKL+ が終極反復収束することを理論的に示すことや，RKL が Leduc poker で特に優れた性能を発揮した原因を特定する実験を設計し明らかにすること，ニューラルネットワークを用いた深層学習時における摂動の影響を分析することなどが挙げられる．

参考文献

[1] Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Atsushi Iwasaki. Adaptively perturbed mirror descent for learning in games. In *ICML*, volume 235, pages 31–80, 2024.

[2] Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, Kentaro Toyoshima, and Atsushi Iwasaki. Last-iterate convergence with full- and noisy-information feedback in two-player zero-sum games. In *AISTATS*, pages 7999–8028, 2023.

[3] Noam Brown and Tuomas Sandholm. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

[4] Harold W. Kuhn. *A simplified two-person poker*, pages 97–104. Princeton University Press, 1951.

[5] Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. OpenSpiel: A framework for reinforcement learning in games. *arxiv preprint arXiv:1908.09453*, 2019.

[6] Marc Lanctot, Kevin Waugh, Martin Zinkevich, and Michael Bowling. Monte carlo sampling for regret minimization in extensive games. In *NeurIPS*, volume 22, pages 1078–1086, 2009.

[7] Chung-Wei Lee, Christian Kroer, and Haipeng Luo. Last-iterate convergence in extensive-form games. In *NeurIPS*, pages 14293–14305, 2021.

[8] Stephen Marcus McAleer, Gabriele Farina, Marc Lanctot, and Tuomas Sandholm. ESCHER: eschewing importance

sampling in games by computing a history value function to estimate regret. In *ICLR*, 2023.

[9] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms(SODA)*, pages 2703–2717, 2018.

[10] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

[11] John Nash. Non-cooperative games. *Annals of mathematics*, 54(2):286–295, 1951.

[12] Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, et al. From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In *ICML*, volume 139, pages 8525–8535, 2021.

[13] Julien Pérolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas W. Anthony, Stephen McAleer, Romuald Élie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Tobias Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, L. Sifre, Nathalie Beauguerlange, Rémi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378:990 – 996, 2022.

[14] Naifeng Zhang, Stephen McAleer, and Tuomas Sandholm. Faster game solving via hyperparameter schedules. *arXiv preprint arXiv:2404.09097*, 2024.

A 定理 4.1 の証明

証明. サンプリング下での摂動反実仮想価値の期待値が，真の摂動反実仮想価値と一致することを示す.

$$
\begin{aligned}
&\mathbb{E}_{j\sim p_j}[\tilde{v}_i^{\pi,\sigma}(x,a;j)] \\
&= \sum_j p_j \sum_{h\in x} \frac{\rho^\pi_{-i}(h)}{p(h)} \tilde{q}_i^{\pi,\sigma}(h,a) \\
&= \sum_j p_j \sum_{h\in x} \frac{\rho^\pi_{-i}(h)}{p(h)} \left\{ \tilde{q}_i^{\pi}(h,a) + \mu\tilde{\delta}_i^{\pi,\sigma}(h,a) \right\} \\
&= \sum_j p_j \sum_{h\in x} \frac{\rho^\pi_{-i}(h)}{p(h)} \left\{ \sum_{h'a'\sqsupseteq ha \wedge h'a'\in H_j} \frac{\rho^\pi(ha,h'a')}{p(h,h'a')} u_i(h',a') \right. \\
&\qquad + \mu \mathbb{1}_{h\in H_j} d_i^{\pi,\sigma}(h,a) \\
&\qquad \left. + \mu \sum_{h'\sqsupseteq ha \wedge h'\in H_j} \sum_{a'\in A(h')} \frac{\rho^\pi(ha,h'a')}{p(h,h')} d_i^{\pi,\sigma}(h',a') \right\} \\
&= \sum_j p_j \sum_{h\in x} \rho^\pi_{-i}(h) \left\{ \sum_{h'a'\sqsupseteq ha \wedge h'a'\in H_j} \frac{\rho^\pi(ha,h'a')}{p(h'a')} u_i(h',a') \right. \\
&\qquad + \mu \frac{\mathbb{1}_{h\in H_j}}{p(h)} d_i^{\pi,\sigma}(h,a) \\
&\qquad \left. + \mu \sum_{h'\sqsupseteq ha \wedge h'\in H_j} \sum_{a'\in A(h')} \frac{\rho^\pi(ha,h'a')}{p(h')} d_i^{\pi,\sigma}(h',a') \right\} \\
&= \sum_{h\in x} \rho^\pi_{-i}(h) \sum_{h'a'\sqsupseteq ha} \frac{\rho^\pi(ha,h'a')}{p(h'a')} u_i(h',a') \sum_{j:h'a'\in H_j} p_j \\
&+ \mu \sum_{h\in x} \frac{\rho^\pi_{-i}(h)}{p(h)} d_i^{\pi,\sigma}(h,a) \sum_{j:h\in H_j} p_j \\
&+ \mu \sum_{h\in x} \rho^\pi_{-i}(h) \\
&\qquad \sum_{h'\sqsupseteq ha} \sum_{a'\in A(h')} \frac{\rho^\pi(ha,h'a')}{p(h')} d_i^{\pi,\sigma}(h',a') \sum_{j:h'\in H_j} p_j \\
&= \sum_{h\in x} \rho^\pi_{-i}(h) \sum_{h'a'\sqsupseteq ha} \rho^\pi(ha,h'a') u_i(h',a') \\
&+ \mu \sum_{h\in x} \rho^\pi_{-i}(h) \\
&\left(d_i^{\pi,\sigma}(h,a) + \sum_{h'\sqsupseteq ha} \sum_{a'\in A(h')} \rho^\pi(ha,h'a') d_i^{\pi,\sigma}(h',a') \right) \\
&= \sum_{h\in x} \rho^\pi_{-i}(h) q_i^\pi(h,a) + \mu \sum_{h\in x} \rho^\pi_{-i}(h) \delta_i^{\pi,\sigma}(h,a) \\
&= \sum_{h\in x} \rho^\pi_{-i}(h) \left(q_i^\pi(h,a) + \mu\delta_i^{\pi,\sigma}(h,a) \right) \\
&= \sum_{h\in x} \rho^\pi_{-i}(h) q_i^{\pi,\sigma}(h,a) \\
&= v_i^{\pi,\sigma}(x,a).
\end{aligned}
$$

□

B 定理 4.2 の証明

証明. 任意の履歴 $h' \in H \setminus Z$ において，次の二つの等式が明らかに成り立つ：

$$
\sum_{a'\in A(h')} \sigma_{\tau(h')}(a'|x(h')) = 1, \qquad \sum_{a'\in A(h')} \pi_{\tau(h')}(a'|x(h')) = 1.
$$

これらが成り立つことから，RKL を摂動関数に用いた際に，累積期待摂動を次のように書き表すことができる：

$$
\begin{aligned}
&\delta_i^{\pi,\sigma}(h,a) \\
&= \sum_{h'a'\sqsupseteq ha} \rho^\pi(ha,h'a') d_i^{\pi,\sigma}(h',a')
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{h'a' \sqsupseteq ha} \rho^{\pi}(ha, h'a') \\
&\left(\frac{\mathbb{1}_{i=\tau(h')}}{\pi_{\tau(h')}(a'|x(h'))} \left(\sigma_{\tau(h')}(a'|x(h')) - \pi_{\tau(h')}(a'|x(h')) \right) \right) \\
&= \frac{1}{\pi_i(a|x(h))} \left(\sigma_i(a|x(h)) - \pi_i(a|x(h)) \right) \\
&+ \sum_{h' \sqsupseteq ha} \rho^{\pi}(ha, h') \\
&\mathbb{1}_{i=\tau(h')} \sum_{a' \in A(h')} \left(\sigma_{\tau(h')}(a'|x(h')) - \pi_{\tau(h')}(a'|x(h')) \right) \\
&= \frac{1}{\pi_i(a|x(h))} \left(\sigma_i(a|x(h)) - \pi_i(a|x(h)) \right).
\end{aligned}
$$

一方で，サンプリング下における累積摂動の推定量の定義より，任意の履歴 $h \in H$ と行動 $a \in A(h)$ について次が成り立つ：

$$
\begin{aligned}
&\tilde{\delta}_i^{\pi,\sigma}(h, a; j) \\
&= \mathbb{1}[h \in H_j] d_i^{\pi,\sigma}(h, a) \\
&+ \sum_{h' \sqsupseteq ha \wedge h' \in H_j} \sum_{a' \in A(h')} \frac{\rho^{\pi}(ha, h'a')}{p(h, h')} d_i^{\pi,\sigma}(h', a') \\
&= \mathbb{1}[h \in H_j] \frac{1}{\pi_i(a|x(h))} \left(\sigma_i(a|x(h)) - \pi_i(a|x(h)) \right) \\
&+ \sum_{h' \sqsupseteq ha \wedge h' \in H_j} \sum_{a' \in A(h')} \frac{\rho^{\pi}(ha, h'a')}{p(h, h')} \\
&\left(\frac{\mathbb{1}_{i=\tau(h')}}{\pi_{\tau(h')}(a'|x(h'))} \left(\sigma_{\tau(h')}(a'|x(h')) - \pi_{\tau(h')}(a'|x(h')) \right) \right) \\
&= \mathbb{1}[h \in H_j] \frac{1}{\pi_i(a|x(h))} \left(\sigma_i(a|x(h)) - \pi_i(a|x(h)) \right) \\
&+ \sum_{h' \sqsupseteq ha \wedge h' \in H_j} \frac{\rho^{\pi}(ha, h')}{p(h, h')} \mathbb{1}_{i=\tau(h')} \\
&\sum_{a' \in A(h')} \left(\sigma_{\tau(h')}(a'|x(h')) - \pi_{\tau(h')}(a'|x(h')) \right) \\
&= \mathbb{1}_{h \in H_j} \frac{1}{\pi_i(a|x(h))} \left(\sigma_i(a|x(h)) - \pi_i(a|x(h)) \right).
\end{aligned}
$$

よって，Q_j がサンプリングされた時，任意の履歴 $h \in H_j$ と行動 $a \in A(h)$ について，

$$
\tilde{\delta}_i^{\pi,\sigma}(h, a; j) = \frac{1}{\pi_i(a|x(h))} \left(\sigma_i(a|x(h)) - \pi_i(a|x(h)) \right),
$$

となる．したがって，これらを組み合わせると任意の履歴 $h \in H$ と行動 $a \in A(h)$ について，

$$
\tilde{\delta}_i^{\pi,\sigma}(h, a; j) = \delta_i^{\pi,\sigma}(h, a),
$$

となり，これは，累積摂動の推定量の分散が 0 であることを示している：

$$
\mathrm{Var}_{j \sim p_j}[\tilde{\delta}_i^{\pi,\sigma}(h, a; j) \mid h \in H_j] = 0.
$$

□

C　アルゴリズム

Algorithm 1: OS-PFTRL-RKL+ and -KL+.

Input : Time horizon T, learning rate η, mutation parameter μ, update interval T_σ

$\pi_i^1(\cdot|x) \leftarrow \left(\frac{1}{|A(x)|} \right)_{a \in A(x)}$ for all $i \in N$ and $x \in X_i$
$\sigma \leftarrow \pi^1$
$\kappa_i[x] \leftarrow 0$ for all $i \in N$ and $x \in X_i$
for $t = 1, 2 \cdots, T$ **do**
  for $i \in N$ **do**
    Estimate the perturbed counterfactual values $v_i^{\pi^t,\sigma}$ by Algorithm 2 with inputs (π^t, i, μ, σ)
    Let $X_{i,\mathrm{vst}}^t$ be the player i's information sets visited at iteration t
    for $x \in X_{i,\mathrm{vst}}^t$ **do**
      Update the strategy by $\pi_i^{t+1}(\cdot|x) = \arg\max_{\pi \in \Delta(A(x))} \left\{ \eta \left\langle \sum_{s=1}^t \tilde{v}_i^{\pi^s,\sigma}(x, \cdot), \pi \right\rangle - \psi_i(\pi) \right\}$
      $\kappa_i[x] \leftarrow \kappa_i[x] + 1$;
      if $\kappa_i[x] = T_\sigma$ **then**
        $\sigma_i(\cdot|x) \leftarrow \pi_i^{t+1}(\cdot|x)$
        $\kappa_i[x] \leftarrow 0$
      end if
    end for
  end for
end for

Algorithm 2: VALUEESTIMATEOS(π, i, μ, σ) for Outcome sampling.

$\tilde{v}_i^{\pi,\sigma}(x, a) \leftarrow 0$ for all $x \in X_i$ and $a \in A(x)$
subroutine TRAVERSEOS (h, i, ρ_i)
  if $h \in Z$ **then**
    return 0
  else if $\tau(h) = c$ **then**
    Sample action $a \sim \pi_c(a|h)$
    return TRAVERSEOS $(ha, i, \rho_i) + u_i(h, a)$
  end if
  Let x be the information set containing h
  if $\tau(h) = i$ **then**
    $\pi'_{\tau(h)}(\cdot|x) \leftarrow (1 - \epsilon)\pi_{\tau(h)}(\cdot|x) + \frac{\epsilon}{|A(x)|}$
  else
    $\pi'_{\tau(h)}(\cdot|x) \leftarrow \pi_{\tau(h)}(\cdot|x)$
  end if
  Sample action $a \sim \pi'_{\tau(h)}(a|x)$
  if $\tau(h) = i$ **then**
    $q_i[h] \leftarrow 0$
    $q_i[h, a'] \leftarrow 0$ for all $a' \in A(x)$
    for $a' \in A(x)$ **do**
      if $a' = a$ **then**
        $q_i[h, a'] \leftarrow$ TRAVERSEOS $(ha', i, \pi'_i(a'|x) \cdot \rho_i) + u_i(h, a')$
      end if
      $q_i[h, a'] \leftarrow \frac{q_i[h,a']}{\pi'_i(a'|h)} + \mu d_i^{\pi,\sigma}(h, a')$
      $\tilde{v}_i^{\pi,\sigma}(x, a') \leftarrow \frac{q_i[h,a']}{\rho_i}$
      $q_i[h] \leftarrow q_i[h] + \pi_i(a'|x) \cdot q_i[h, a']$
    end for
    return $q_i[h]$
  else
    return TRAVERSEOS $(ha, i, \rho_i) + u_i(h, a)$
  end if
TRAVERSEOS$(\emptyset, i, 1)$
return $\tilde{v}_i^{\pi,\sigma}$