CF-002

二人零和ゲームにおける 突然変異駆動型 Follow-The-Regularized-Leader の終極反復収束

豊島 健太郎 * Kentaro Toyoshima
阿部 拳之 † Kenshi Abe
坂本 充生 * Mitsuki Sakamoto
岩崎 敦 * Atsushi Iwasaki

概要

本研究では, 二人零和ゲームにおける Follow the Regularized Leader (FTRL) ダイナミクスに，突然変異のアイデアを取り入れた，Mutant FTRL (M-FTRL) を提案・解析する．FTRL とその様々な改良版は毎期の戦略を時間で平均することで，ナッシュ均衡に収束することが知られているが，それらのほとんどがナッシュ均衡点の周回軌道に収束してしまい，ナッシュ均衡点に直接収束すること（終極反復収束, last-iterate convergence）を保証できない．これに対して，本研究で提案する M-FTRL は各期で行動を取る確率を摂動させることで，終極反復収束を実現する．具体的には，M-FTRL の連続時間ダイナミクスが突然変異付きレプリケータダイナミクス (Replicator-Mutator Dyanmics, RMD) と等価になることを示し，M-FTRL が近似ナッシュ均衡である RMD の定常点に収束することを示した．さらに，近似でないナッシュ均衡への終極反復収束を保証するダイナミクスを構成することに成功した．

1 はじめに

本研究では，二人零和ゲームにおける均衡を学習アルゴリズムで計算する問題に注目する．2 人零和ゲームの均衡を見つけるには，$\min_x \max_y f(x,y)$ で表されるミニマックス最適化（または鞍点最適化）を解かなければならない．マルチエージェント強化学習 [6] や Generative Adversarial Networks (GANs) [11] などの成功からミニマックス最適化の近似解を効率的に計算するアルゴリズムの開発に大きな関心が寄せられている [5, 9].

後悔最小化学習において，戦略の組を二人零和ゲームのナッシュ均衡に収束させることを目指した研究は多い [3, 24, 8]．しかし、よく知られている Follow the Regularized Leader (FTRL) のような後悔最小化学習でも，その戦略を時間平均しないと循環して収束しない [20, 2]．FTRL は正則化項にエントロピーを設定したとき，進化ゲームでよく知られているレプリケータダイナミクスと力学的に同相になる．図 1a に Biased Rock-Paper-Scissors というグーで勝つとうれしいじゃんけんのゲーム (表 1) のダイナミクスを示す．ここでは赤い点で示したナッシュ均衡を中心にした周回軌道を取ることがみてとれる．楽観的 FTRL (Oppotunistic FTRL, O-FTRL) は数少ない例外であり，更新されていく戦略の軌跡がサイクルを形成することなく直接均衡に収束する [9, 10, 19, 23, 17]．このような収束特性のことをの終極反復収束 (last-iterate convergence) と呼ぶ．O-FTRL はこの優れた性質をもつが，一般に用いられるエントロピー正則化項を用いた場合，具体的な収束レートを出すためには与えられたゲームの均衡が一意に定まらなければならないと言った仮定を置く必要がある [10, 23].

表 1: Biased Rock-Paper-Scissors game

	R	P	S
R	0	−0.1	0.3
P	0.1	0	−0.1
S	−0.3	0.1	0

これに対して本研究では，FTRL に，突然変異のアイデアを取り入れた，Mutant FTRL (M-FTRL) を提案・解析する．M-FTRL は各期で行動を取る確率を摂動させることで，終極反復収束を実現する．具体的には，M-FTRL の連続時間ダイナミクスが，その正則化項をエントロピーにしたとき，突然変異付きレプリケータダイナミクス (Replicator-Mutator Dyanmics, RMD) と等価になることを示し，M-FTRL が近似ナッシュ均衡である RMD のの定常点に収束することを示した．直感的には図 1 に示した通り，均衡点を中心とした周回軌道に陥らずに，近似ナッシュ均衡である RMD の定常点に収束する．さらに，M-FTRL の基準戦略 (reference strategy) 項を適宜更新することで， 近似でないナッシュ均衡への終極反復収束を保証することに成功した．

2 問題設定

二人標準形零和ゲームでは，プレイヤ $i \in \{1,2\}$ (相手プレイヤは $-i$) は有限行動集合 A から行動 a_i を混合戦略 $\pi_i \in \Delta(A_i)$ に従い決定する．その行動の組を $a = (a_1, a_2) \in A = (A_1 \times A_2)$, 戦略の組を $\pi = (\pi_1, \pi_2)$ とする．プレイヤ i の利得は $u_i \in [-u_{max}, u_{max}]^{A_1 \times A_2}$ で与えられ，$u_1(a) = -u_2(a)$ を満たす．π が与えられたとき，プレイヤ i の期待利得を $v_i^\pi = \mathbb{E}_{a\sim\pi}[u_i(a_1, a_2)]$ とする．さらに，行動 a_i を取った時の期待利得を $q_i^\pi(a_i) = \mathbb{E}_{a_{-i}\sim\pi_{-i}}[u_i(a_i, a_{-i})]$ とする．ナッシュ均衡とは，片方のプレイ

* 電気通信大学大学院情報理工学研究科
† 株式会社サイバーエージェント

ヤがその戦略から逸脱しても，利得を改善することのできない戦略の組である．二人零和標準形ゲームで, ナッシュ均衡 $\pi^* = (\pi_1^*, \pi_2^*)$ は次の条件を満たす: $\forall \pi_1 \in \Delta(A_1), \forall \pi_2 \in \Delta(A_2)$,

$$v_1^{\pi_1^*,\pi_2} \geq v_1^{\pi_1^*,\pi_2^*} \geq v_1^{\pi_1,\pi_2^*}.$$

また, 以下の不等式を満たす戦略の組を ϵ ナッシュ均衡 (π_1, π_2) と呼ぶ:

$$\max_{\tilde{\pi}_1 \in \Delta(A_1)} v_1^{\tilde{\pi}_1,\pi_2} + \max_{\tilde{\pi}_2 \in \Delta(A_2)} v_2^{\pi_1,\tilde{\pi}_2} \leq \epsilon.$$

二人零和ゲームで，戦略の組 π がどれだけナッシュ均衡 π^* に近いかを図る指標が Exploitability である [15, 1]. Exploitability は $\mathrm{exploit}(\pi) := \max_{\tilde{\pi}_1 \in \Delta(A_1)} v_1^{\tilde{\pi}_1,\pi_2} + \max_{\tilde{\pi}_2 \in \Delta(A_2)} v_2^{\pi_1,\tilde{\pi}_2}$ と定義される [15, 14, 18, 22, 1]. 定義より，ナッシュ均衡 π^* の Exploitability は 0 である．

本研究では，時刻 $t = 1, .., T$ においてプレイヤが繰り返してゲームすることを考える．ある時刻 t でプレイヤ i は過去の観測をもとに (混合) 戦略 $\pi_i^t \in \Delta(A_i)$ を決める．次に，プレイヤ i は利得を観測するが，そこには様々な設定が考えられている．本研究では，完全フィードバックと部分フィードバックを扱う．完全フィードバックの下では，時刻 t の終わりにプレイヤ i は各行動に対する期待利得 $(q_i^{\pi^t}(a_i))_{a_i \in A_i}$ を観測する．一方，部分フィードバックの下では, 各プレイヤ i は π_i^t に従って，行動 a_i^t を選択し，実現利得 $u_i(a_1^t, a_2^t)$ のみを観測する．

FTRL はゲームの学習で広く用いられる学習アルゴリズムである．FTRL は時刻 t の戦略 π_i^t を次のように決定する:

$$\pi_i^t = \arg\max_{p \in \Delta(A_i)} \left\{ \left\langle y_i^t, p \right\rangle - \psi_i(p) \right\},$$
$$y_i^t(a_i) = \sum_{s=1}^{t-1} q_i^{\pi^s}(a_i).$$

ここで正則化項 $\psi_i : \Delta(A_i) \to \mathbb{R}$ は連続微分可能で狭義凸な関数である．

$\Delta(A_i)$ の内点を $\Delta^\circ(A_i) := \{p \in \Delta(A_i) \mid \forall a_i \in A_i,\ p(a_i) > 0\}$ と定義する．また, 連続微分可能な狭義凸関数 ψ において, *Bregman divergence* を $D_\psi(x, x') = \psi(x) - \psi(x') - \langle \nabla\psi(x'), x - x' \rangle$ と定義する．*Kullback-Leibler divergence* は，エントロピー正則化項 $\psi(x) = \sum_i x_i \ln x_i$ を持つ Bregman divergence で，$\mathrm{KL}(x, x') = \sum_i x_i \ln \frac{x_i}{x_i'}$ で示される．さらに，Bregman divergences の和と Kullback-Leibler divergences の和をそれぞれ $D_\psi(\pi, \pi') = \sum_{i=1}^2 D_{\psi_i}(\pi_i, \pi_i')$, $\mathrm{KL}(\pi, \pi') = \sum_{i=1}^2 \mathrm{KL}(\pi_i, \pi_i')$ と定義する．

3 Mutant Follow the Regularized Leader

M-FTRL は完全フィードバックの下で，戦略 x^t は次のように決定する：

$$\pi_i^t = \arg\max_{p \in \Delta(A_i)} \left\{ \eta \left\langle \sum_{s=1}^{t-1} q_i^{\mu,t}, p \right\rangle - \psi_i(p) \right\}, \tag{1}$$

$$q_i^{\mu,s}(a_i) = q_i^{\pi^s}(a_i) + \frac{\mu}{\pi_i^s(a_i)} \left(c_i(a_i) - \pi_i^s(a_i) \right). \tag{2}$$

ここで $\eta \in \mathbb{R}_{>0}$ は学習率, $\mu \in \mathbb{R}_{>0}$ は突然変異圧, $c_i \in \Delta(A_i) \cap \mathbb{R}_{>0}^{|A_i|}$ は参照戦略とする．

図 1a は，Biased Rock-Paper-Scissors における FTRL と等価な RD の学習推移であり，均衡解の周回軌道に陥ってしまっていることがわかる．しかし，FTRL でも時間平均をとることで均衡解に収束する [12]. 図 1b から図 1d に示すように，戦略 π_i^t を 式 1 に従って更新すると，ダイナミクスはゲームの均衡とは違う場所に収束する．この定常点は 2μ ナッシュ均衡であり，その定常点は (c_1, c_2) がナッシュ均衡でない限り，ゲームのナッシュ均衡とは異なる．そのため，学習戦略をゲームのナッシュ均衡に収束させるためには，適応的に参照戦略を更新する必要がある．具体的には，時刻 $N(\leq T)$ 毎に c_i を π_i^t に上書きする [21].

部分フィードバックの下では，各プレイヤ i は $q_i^{\mu,t} = \left(q_i^{\pi^t}(a_i) + \frac{\mu}{\pi_i^t(a_i)} \left(c_i(a_i) - \pi_i^t(a_i) \right) \right)_{a_i \in A_i}$ を実現利得 $u_i(a^t)$ から推定する必要がある．そこで重点サンプリング [13] を用いて $\hat{q}_i^{\mu,t}(\cdot)$ を推定する:

$$\hat{q}_i^{\mu,t}(a_i) = \frac{u_i(a_1^t, a_2^t)}{\pi_i^t(a_i^t)} \mathbb{1}[a_i = a_i^t] + \frac{\mu}{\pi_i^t(a_i)} \left(c_i(a_i) - \pi_i^t(a_i) \right). \tag{3}$$

ここで，$\hat{q}_i^{\mu,t}$ は $q_i^{\mu,t}$ の不偏推定量である. 部分フィードバックの下では, 式 1 において，$q_i^{\mu,t}$ の代わりに $\hat{q}_i^{\mu,t}$ を用いる．

4 理論解析

本節では，RMD の定常点との理論的関係性を明らかにする．解析のために，離散時間の M-FTRL の代わりに，連続時間における M-FTRL のダイナミクスを考える:

$$\pi_i^t = \arg\max_{p \in \Delta(A_i)} \left\{ \left\langle z_i^t, p \right\rangle - \psi_i(p) \right\}, \tag{4}$$

$$z_i^t(a_i) = \int_0^t \left(q_i^{\pi^s}(a_i) + \frac{\mu}{\pi_i^s(a_i)} \left(c_i(a_i) - \pi_i^s(a_i) \right) \right) ds. \tag{5}$$

まず，このダイナミクスが RMD [4] の一般化であることを示す．次の定理は正則化関数をエントロピー正則化，つまり $\psi_i(p) = \sum_{a_i \in A_i} p(a_i) \ln p(a_i)$ に限定したとき，M-FTRL のダイナミクスから RMD が導かれることを示す.

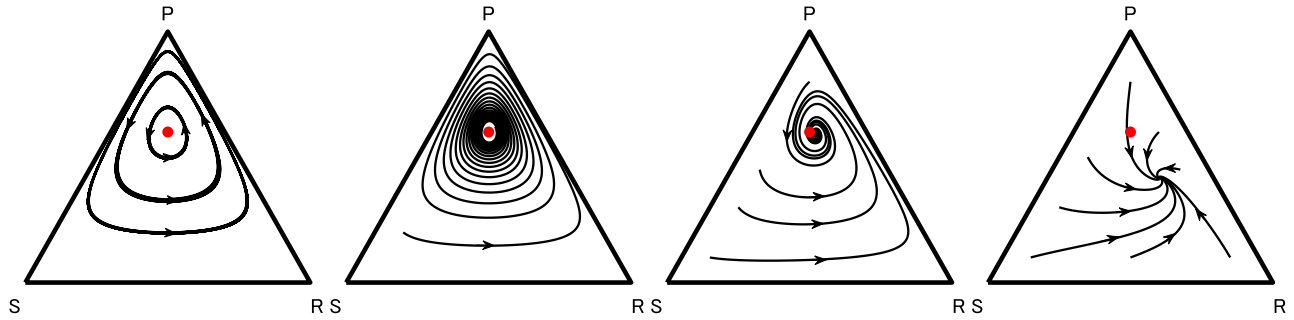
定理 1. エントロピー正則化項 $\psi_i(p) = \sum_{a_i \in A_i} p(a_i) \ln p(a_i)$ とする M-FTRL のダイナミクスは

$$\frac{d}{dt}\pi_i^t(a_i) = \pi_i^t(a_i) \left(q_i^{\pi^t}(a_i) - \langle \pi_i^t, q_i^{\pi^t} \rangle \right) + \mu \left(c_i(a_i) - \pi_i^t(a_i) \right). \tag{RMD}$$

と等しくなる．

証明. π_i^t をソフトマックス関数で表すと，

$$\pi_i^t(a_i) = \frac{\exp\left(z_i^t(a_i) \right)}{\sum_{a_i' \in A_i} \exp\left(z_i^t(a_i') \right)}.$$



(a) RD (b) RMD ($\mu = 0.01$) (c) RMD ($\mu = 0.1$) (d) RMD ($\mu = 1.0$)

図 1: Biased Rock-Paper-Scissors における RD と RMD の学習のダイナミクス (赤点はナッシュ均衡点).

$\pi_i^t(a_i)$ を時間 t で微分すると,

$$\begin{aligned}\frac{d}{dt}\pi_i^t(a_i) =& \frac{\frac{d}{dt}\exp\left(z_i^t(a_i)\right)}{\sum_{a_i'\in A_i}\exp\left(z_i^t(a_i')\right)} - \frac{\exp\left(z_i^t(a_i)\right)\frac{d}{dt}\left(\sum_{a_i'\in A_i}\exp\left(z_i^t(a_i')\right)\right)}{\left(\sum_{a_i'\in A_i}\exp\left(z_i^t(a_i')\right)\right)^2}\\ =&\pi_i^t(a_i)\frac{d}{dt}z_i^t(a_i) - \pi_i^t(a_i)\frac{\left(\sum_{a_i'\in A_i}\exp\left(z_i^t(a_i')\right)\frac{d}{dt}z_i^t(a_i')\right)}{\sum_{a_i'\in A_i}\exp\left(z_i^t(a_i')\right)}\\ =&\pi_i^t(a_i)\frac{d}{dt}z_i^t(a_i) - \pi_i^t(a_i)\sum_{a_i'\in A_i}\pi_i^t(a_i')\frac{d}{dt}z_i^t(a_i'). \qquad (6)\end{aligned}$$

式 5 の時間微分 $\frac{d}{dt}z_i^t(a_i)$ を式 6 に代入し,

$$\begin{aligned}\frac{d}{dt}\pi_i^t(a_i) =&\pi_i^t(a_i)\left(q_i^{\pi^t}(a_i) + \frac{\mu}{\pi_i^t(a_i)}\left(c_i(a_i) - \pi_i^t(a_i)\right)\right)\\ &- \pi_i^t(a_i)\sum_{a_i'\in A_i}\pi_i^t(a_i')\left(q_i^{\pi^t}(a_i') + \frac{\mu}{\pi_i^t(a_i')}\left(c_i(a_i') - \pi_i^t(a_i')\right)\right)\\ =&\pi_i^t(a)\left(q_i^{\pi^t}(a_i) - \langle\pi_i^t, q_i^{\pi^t}\rangle\right) + \mu\left(c_i(a_i) - \pi_i^t(a_i)\right).\end{aligned}$$

を得る．この式変形には，任意の $\pi_i \in \Delta(A_i)$ について，$\sum_{a_i'\in A_i}\pi_i(a_i') = 1$ であることを用いている．以上より，M-FTRL のダイナミクスは RMD と等しいことが示された． □

次に，RMD の定常点と任意の時刻 t における M-FTRL の戦略 との関係を明らかにする．なお，文献 [4] の補題 3.3 より，任意の $\mu \in \mathbb{R}_{>0}$ に対して，RMD の定常点 $\pi^\mu \in \prod_{i=1}^2\left(\Delta(A_i)\cap\mathbb{R}_{>0}^{|A_i|}\right)$ が存在する．これにより，π^μ と π^t の Bregman divergence $D_\psi(x,x') = \psi(x) - \psi(x') - \langle\nabla\psi(x'), x - x'\rangle$ の時間微分を得る．

定理 2. 突然変異圧を μ とし，プレイヤ $i \in \{1,2\}$ に関して RMD の定常点を $\pi^\mu = (\pi_1^\mu, \pi_2^\mu) \in \Pi_{i\in\{1,2\}}\Delta(A_i)$ とする．さらに M-FTRL にしたがって戦略を更新するときの t 期における戦略の分布 $\pi^t = (\pi_1^t, \pi_2^t)$ とする．このとき，2 つの戦略分布 π^μ と π^t の間の Bregman divergence は

$$\frac{d}{dt}D_\psi(\pi^\mu,\pi^t) = -\mu\sum_{i=1}^2\sum_{a_i\in A_i}c_i(a_i)\left(\sqrt{\frac{\pi_i^t(a_i)}{\pi_i^\mu(a_i)}} - \sqrt{\frac{\pi_i^\mu(a_i)}{\pi_i^t(a_i)}}\right)^2. \qquad (7)$$

さらに，正則化項 $\psi_i(p)$ がエントロピーの和 $\sum_{a_i\in A_i}p(a_i)\ln p(a_i)$ に従うとすると， π^t は

$$\frac{d}{dt}\mathrm{KL}(\pi^\mu,\pi^t) \le -\mu\xi\mathrm{KL}(\pi^\mu,\pi^t). \qquad (8)$$

を満たす．ただし， $\xi = \min_{i\in\{1,2\},a_i\in A_i}\frac{c_i(a_i)}{\pi_i^\mu(a_i)}$ とする．

定理 2 を証明する前に 2 つの補題について述べる．これらの証明は付録に示した．

補題 1. すべての $\pi \in \Delta(A_1)\times\Delta(A_2)$ について，M-FTRL によって更新される π^t は以下の式を満たす: $\frac{d}{dt}D_\psi(\pi,\pi^t) =$

$$\sum_{i=1}^2\sum_{a\in A}\pi_i^t(a_i)\pi_{-i}(a_{-i})u_i(a) + 2\mu - \mu\sum_{i=1}^2\sum_{a_i\in A_i}c_i(a_i)\frac{\pi_i(a_i)}{\pi_i^t(a_i)}. \qquad (9)$$

次の補題では，任意の $\pi_i' \in \Delta(A_i)$ について，期待利得 v^{π^μ} と v^{π_i',π_{-i}^μ} の間の関係を導出する．

補題 2. すべての $i \in \{1,2\}$ について， RMD の定常点を $\pi_i^\mu \in \Delta(A_i)$ とすると，任意の $\pi_i' \in \Delta(A_i)$ で: $\sum_{a\in A}\pi_i'(a_i)\pi_{-i}^\mu(a_{-i})u_i(a) =$

$$\sum_{a\in A}\pi_i^\mu(a_i)\pi_{-i}^\mu(a_{-i})u_i(a) + \mu - \mu\sum_{a_i\in A_i}c_i(a_i)\frac{\pi_i'(a_i)}{\pi_i^\mu(a_i)} \qquad (10)$$

が成り立つ.

この補題は $\pi_i^\mu \in \Delta(A_i)$ が RMD の定常点となっている，つまり， $\pi_i^\mu(a_i)\left(q_i^{\pi^\mu}(a_i) - \langle\pi_i^\mu, q_i^{\pi^\mu}\rangle\right) + \mu\left(c_i(a_i) - \pi_i^\mu(a_i)\right) = 0$ を満たすことを意味している．では，定理 2 を証明していく．

証明. 補題 1 および 2 より， 2 つの戦略分布 π^μ と π^t の間の Bregman divergence の時間微分は， $\frac{d}{dt}D_\psi(\pi^\mu,\pi^t) =$

$$\sum_{i=1}^2\sum_{a\in A}\pi_i^t(a_i)\pi_{-i}^\mu(a_{-i})u_i(a) + 2\mu - \mu\sum_{i=1}^2\sum_{a_i\in A_i}c_i(a_i)\frac{\pi_i^\mu(a_i)}{\pi_i^t(a_i)} \qquad (11)$$

$$= \sum_{i=1}^2\sum_{a\in A}\pi_i^\mu(a_i)\pi_{-i}^\mu(a_{-i})u_i(a) + 4\mu - \mu\sum_{i=1}^2\sum_{a_i\in A_i}c_i(a_i)\left(\frac{\pi_i^t(a_i)}{\pi_i^\mu(a_i)} + \frac{\pi_i^\mu(a_i)}{\pi_i^t(a_i)}\right) \qquad (12)$$

$$= -\mu\sum_{i=1}^2\sum_{a_i\in A_i}c_i(a_i)\left(\sqrt{\frac{\pi_i^t(a_i)}{\pi_i^\mu(a_i)}} - \sqrt{\frac{\pi_i^\mu(a_i)}{\pi_i^t(a_i)}}\right)^2 \qquad (13)$$

と変形できる．ここで式 12 から式 13 を得るには，

$$\sum_{i=1}^{2}\sum_{a\in A}\pi_i^{\mu}(a_i)\pi_{-i}^{\mu}(a_{-i})u_i(a)=\sum_{i=1}^{2}v_i^{\pi^{\mu}}=0$$

を用いる．以上より，定理 2 の式 7 が証明された．

次に，定理 2 の式 8 を証明する．$\xi_i=\min_{a_i\in A_i}\frac{c_i(a_i)}{\pi_i^{\mu}(a_i)}$ と定義する．式 7 より，

$$\begin{aligned}\frac{d}{dt}D_{\psi}(\pi^{\mu},\pi^{t})&=-\mu\sum_{i=1}^{2}\sum_{a_i\in A_i}c_i(a_i)\left(\frac{\pi_i^{t}(a_i)}{\pi_i^{\mu}(a_i)}+\frac{\pi_i^{\mu}(a_i)}{\pi_i^{t}(a_i)}-2\right)\\&\le-\mu\sum_{i=1}^{2}\xi_i\sum_{a_i\in A_i}\frac{(\pi_i^{t}(a_i)-\pi_i^{\mu}(a_i))^2}{\pi_i^{t}(a_i)}\\&\le-\mu\sum_{i=1}^{2}\xi_i\sum_{a_i\in A_i}\pi_i^{\mu}(a_i)\ln\left(\frac{\pi_i^{\mu}(a_i)}{\pi_i^{t}(a_i)}\right)\end{aligned}$$

2 つ目の式から 3 つ目の式は 凹関数 $\ln(\cdot)$ の性質とイェンセンの不等式を用いた．

また，$\psi_i(p)=\sum_{a_i\in A_i}p(a_i)\ln p(a_i)$ ならば，$\mathrm{KL}(\pi_i^{\mu},\pi_i^{t})=\sum_{a_i\in A_i}\pi_i^{\mu}(a_i)\ln\left(\frac{\pi_i^{\mu}(a_i)}{\pi_i^{t}(a_i)}\right)$ であることから，

$$\frac{d}{dt}\mathrm{KL}(\pi^{\mu},\pi^{t})\le-\mu\xi\mathrm{KL}(\pi^{\mu},\pi^{t}).\tag{14}$$

以上より，定理 2 の式 8 が証明された． □

定理 2 で示したのは，もし定常点 π^{μ} と t 期における戦略 π^{t} が等しければ，それらの Bregman divergence の時間微分がゼロになり，等しくなければ必ず負になることである．よって，リアプノフ安定論 [16] より，定常点 π^{μ} と戦略 π^{t} の Bregman divergence は 0 に漸近的に収束，すなわち π^{t} は π^{μ} に漸近的に収束する．なお，定理 2 は RMD のすべての定常点に対して成り立つので，ある μ と $(c_i)_{i=1}^2$ に対して，定常点が一意に定まることがわかる．

次の系では，正則化項をエントロピーにしたときの収束速度を示す．

系 1. 正則化項をエントロピー $\psi_i(p)=\sum_{a_i\in A_i}p(a_i)\ln p(a_i)$ と仮定する．このとき，M-FTRL は RMD の定常点に指数速度で収束する．

$$\mathrm{KL}(\pi^{\mu},\pi^{t})\le\mathrm{KL}(\pi^{\mu},\pi^{0})\exp\left(-\mu\xi t\right).$$

証明. 定理 2 の式 8 から，$\frac{d}{dt}\mathrm{KL}(\pi^{\mu},\pi^{t})\le-\mu\xi\mathrm{KL}(\pi^{\mu},\pi^{t})$. 両辺を $\mathrm{KL}(\pi^{\mu},\pi^{t})$ で割ることにより，$\frac{1}{\mathrm{KL}(\pi^{\mu},\pi^{t})}\frac{d}{dt}\mathrm{KL}(\pi^{\mu},\pi^{t})\le-\mu\xi$. 両辺を t で積分して指数をとると，

$$\mathrm{KL}(\pi^{\mu},\pi^{t})\le\mathrm{KL}(\pi^{\mu},\pi^{0})\exp\left(-\mu\xi t\right).$$

ただし，積分定数 C は $\mathrm{KL}(\pi^{\mu},\pi^{0})=\exp(C)$ を満たす． □

この系 1 および文献 [4] の補題 3.5 より，π^{t} の Exploitability の上界を導出する．

定理 3. 正則化項をエントロピー $\psi_i(p)=\sum_{a_i\in A_i}p(a_i)\ln p(a_i)$ と仮定する．このとき，Exploitability の上界は

$$\mathrm{exploit}(\pi^{t})\le2\mu+2u_{\max}\sqrt{(\ln 2)\mathrm{KL}(\pi^{\mu},\pi^{0})}\exp\left(-\frac{\mu\xi}{2}t\right)$$

となる．

定理 3 は，π^{t} の Exploitability が 2μ ナッシュ均衡に指数速度で収束することを示している．

証明. Exploitability の定義より，$\mathrm{exploit}(\pi^{t})=$

$$\begin{aligned}\sum_{i=1}^{2}\max_{\tilde{\pi}_i\in\Delta(A_i)}v_i^{\tilde{\pi}_i,\pi_{-i}^{t}}&\le\sum_{i=1}^{2}\left(\max_{\tilde{\pi}_i\in\Delta(A_i)}v_i^{\tilde{\pi}_i,\pi_{-i}^{\mu}}+\max_{\tilde{\pi}_i\in\Delta(A_i)}\left(v_i^{\tilde{\pi}_i,\pi_{-i}^{t}}-v_i^{\tilde{\pi}_i,\pi_{-i}^{\mu}}\right)\right)\\&\le\sum_{i=1}^{2}\left(\max_{\tilde{\pi}_i\in\Delta(A_i)}v_i^{\tilde{\pi}_i,\pi_{-i}^{\mu}}+\|\pi_i^{\mu}-\pi_i^{t}\|_1\max_{\tilde{\pi}_{-i}\in\Delta(A_{-i})}\|q_i^{\pi_i^{t},\tilde{\pi}_{-i}}\|_{\infty}\right)\\&\le\sum_{i=1}^{2}\left(\max_{\tilde{\pi}_i\in\Delta(A_i)}v_i^{\tilde{\pi}_i,\pi_{-i}^{\mu}}+u_{\max}\sqrt{2(\ln 2)\mathrm{KL}(\pi_i^{\mu},\pi_i^{t})}\right).\end{aligned}\tag{15}$$

最後の不等式は文献 [7] の Lemma 11.6.1 を用いて変形した．

文献 [4] の Lemma 3.5 より，RMD の定常点 π^{μ} は，すべての $i\in\{1,2\}$ と $a_i\in A_i$ について，$q_i^{\pi^{\mu}}(a_i)-\langle\pi_i^{t},q_i^{\pi^{\mu}}\rangle\le\mu$ を満たす．したがって，$\max_{\tilde{\pi}_i\in\Delta(A_i)}v_i^{\tilde{\pi}_i,\pi_{-i}^{\mu}}$ の和は以下の式でおさえることができる: $\sum_{i=1}^{2}\max_{\tilde{\pi}_i\in\Delta(A_i)}v_i^{\tilde{\pi}_i,\pi_{-i}^{\mu}}=$

$$\sum_{i=1}^{2}\left(\max_{\tilde{\pi}_i\in\Delta(A_i)}v_i^{\tilde{\pi}_i,\pi_{-i}^{\mu}}-v_i^{\pi^{\mu},\pi_{-i}^{\mu}}\right)=\sum_{i=1}^{2}\left(\max_{a_i\in A_i}q_i^{\pi^{\mu}}(a_i)-\langle\pi_i^{\mu},q_i^{\pi^{\mu}}\rangle\right)\le2\mu.\tag{16}$$

式 15 および式 16 より:

$$\begin{aligned}\mathrm{exploit}(\pi^{t})&\le2\mu+u_{\max}\sum_{i=1}^{2}\sqrt{2(\ln 2)\mathrm{KL}(\pi_i^{\mu},\pi_i^{t})}\\&\le2\mu+2u_{\max}\sqrt{(\ln 2)\mathrm{KL}(\pi^{\mu},\pi^{0})}\exp\left(-\frac{\mu\xi}{2}t\right).\end{aligned}$$

1 つ目の式から 2 つ目の式は，すべての $a,b>0$ について，$\sqrt{a}+\sqrt{b}\le\sqrt{2(a+b)}$ であることを用いた．また，2 つ目の式から 3 つ目の式では，系 1 の結果を用いた．以上のことから，定理 3 が証明された． □

ここまで M-FTRL のダイナミクスが RMD の定常点へ収束することを証明した．次の定理 4 では，M-FTRL のダイナミクスがナッシュ均衡へ収束することを示す．このために，任意の $i\in\{1,2\}$ において，π_i^{c} を参照戦略を c としたときの (RMD) の定常点とする．また，参照戦略 c から定常点 π^{c} への写像 $F(c)$ を $\prod_{i=1}^{2}\Delta^{\circ}(A_i)\to\prod_{i=1}^{2}\Delta^{\circ}(A_i)$ とする．

定理 4. 任意の $c^0\in\prod_{i=1}^{2}\Delta^{\circ}(A_i)$ で始まる定常点の数列を $c^k=F(c^{k-1})$ $(k\ge1)$ で表す．このとき，$\{c^k\}_{k\ge0}$ は与えられたゲームの（近似でない）ナッシュ均衡 π^{*} に収束する．

定理 4 を証明する前に 2 つの補題について述べる．これらの証明は付録に示した．

補題 3. 任意の $i \in \{1,2\}$ において，参照戦略を c とした (RMD) の定常点を π_i^c とする．また，Π^* をオリジナルゲームのナッシュ均衡の集合とする．このとき，$c \notin \Pi^*$ ならば:

$$\min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, \pi^c) < \min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, c).$$

さらに，$c \in \Pi^*$ ならば，$\pi^c = c \in \Pi^*$ が成り立つ．

補題 4. 任意の $i \in \{1,2\}$ において，参照戦略を c とした (RMD) の定常点を π_i^c とする．また，$F(c) : \prod_{i=1}^2 \Delta^\circ(A_i) \to \prod_{i=1}^2 \Delta^\circ(A_i)$ を参照戦略 c から定常点 π^c への写像とする．ただし，$F(\cdot)$ は $\prod_{i=1}^2 \Delta^\circ(A_i)$ 上の連続関数である．

証明. 補題 3 より，$\pi^* \in \Pi^*$ と c^k の間の Kullback-Leibler divergence は k が大きくなるにつれて単調減少する．また，$\min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, c^k) \geq 0$ であるから，$\min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, c^k)$ はある実数 $b \geq 0$ に収束する．ここで，$b = 0$ であることを証明する．$b > 0$ と仮定する．ある $b' > 0$ をおき，$\Omega_{b'} := \{c \in \prod_{i=1}^2 \Delta^\circ(A_i) \mid \min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, c) \leq b'\}$ と $\bar{\Omega}_{b'} := \{c \in \prod_{i=1}^2 \Delta^\circ(A_i) \mid \min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, c) < b'\}$ を定義する．すべての $b' > 0$ について，$\Omega_{b'}$ 有界な閉集合であり，$\bar{\Omega}_{b'}$ 有界な開集合である．ここで，$B = \min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, c^0)$ を定義する．$b > 0$ であるから，すべての $k \geq 0$ について，$c^k \in \Omega_B \setminus \bar{\Omega}_b$ である．ゆえに，Ω_B は有界な閉集合，$\bar{\Omega}_b$ は有界な開集合，$\Omega_B \setminus \bar{\Omega}_b$ は有界な閉集合である．したがって，$\Omega_B \setminus \bar{\Omega}_b$ はコンパクトである．

ここで，補題 4 より，$\min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, F(c)) - \min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, c)$ は連続関数である．コンパクト集合上の連続関数は最大値を持つ，すなわち $\Delta := \max_{c \in \Omega_B \setminus \bar{\Omega}_b} \{\min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, F(c)) - \min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, c)\}$ が存在する．

$b > 0$ であるから，すべての $\pi^* \in \Pi^*$ について，$\pi^* \notin \Omega_B \setminus \bar{\Omega}_b$ である，したがって，補題 3 より，$\Delta < 0$．これらのことから，次の式が成り立つ．$\min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, c^k) =$

$$\begin{aligned} &\min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, c^0) + \sum_{l=0}^{k-1} \left(\min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, c^{l+1}) - \min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, c^l) \right) \\ \leq & B + \sum_{l=0}^{k-1} \Delta = B + k\Delta. \end{aligned}$$

これは，$k > \frac{B}{-\Delta}$ のとき，$\min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, c^k) < 0$ となるが，$\min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, c) \geq 0$ であることに矛盾する．したがって，$\min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, c^k)$ の列は 0 に収束し，c^k は Π^* 内のある要素に収束する． □

5 おわりに

本研究では，二人零和ゲームにおける学習アルゴリズムとして，FTRL に突然変異項を導入した M-FTRL を提案し，そのナッシュ均衡への終極反復収束を証明した．M-FTRL のダイナミクスは正則化項にエントロピーを用いたとき RMD と等価になり，その定常点に収束する．さらに，近似でないナッシュ均衡へ直接収束するようアルゴリズムを構成することに成功し，その収束速度も示した．終極反復収束を保証する学習ダイナミクスはほとんど知られておらず，本研究はその非常に稀な一例を発見した．今後，突然変異のアイデアを利用したアルゴリズムの拡張が進んでいくことが期待される．今後の課題としては，M-FTRL を拡張し，展開型ゲームやマルコフゲームといった複雑なゲームでの理論解析や計算機実験などが挙げられる．

参考文献

[1] K. Abe and Y. Kaneko. Off-policy exploitability-evaluation in two-player zero-sum markov games. In *AAMAS*, 2021.

[2] J. P. Bailey and G. Piliouras. Multiplicative weights update in zero-sum games. In *Proceedings of the ACM Conference on Economics and Computation (EC)*, pp. 321–338, 2018.

[3] B. Banerjee and J. Peng. Efficient no-regret multiagent learning. In *AAAI*, pp. 41–46, 2005.

[4] J. Bauer, M. Broom, and E. Alonso. The stabilization of equilibria in evolutionary game dynamics through mutation: mutation limits in evolutionary games. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2231):20190355, 2019.

[5] A. Blum and Y. Monsour. Learning, regret minimization, and equilibria. In *Algorithmic game theory*, pp. 79–101. Cambridge University Press, 2007.

[6] L. Busoniu, R. Babuska, and B. De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

[7] T. M. Cover and J. A. Thomas. *Elements of information theory 2nd Edition*. Wiley-Interscience, 2006.

[8] C. Daskalakis, A. Deckelbaum, and A. Kim. Near-optimal no-regret algorithms for zero-sum games. In *Symposium on Discrete Algorithms*, pp. 235–254, 2011.

[9] C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training gans with optimism. In *ICLR*, 2018.

[10] C. Daskalakis and I. Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. *Innovations in Theoretical Computer Science*, 2019.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[12] J. Hofbauer, S. Sorin, and Y. Viossat. Time average replicator and best-reply dynamics. *Mathematics of Operations Research*, 34(2):263–269, 2009.

[13] S. Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *COLT*, pp. 2552–2583, 2021.

[14] M. Johanson, N. Bard, N. Burch, and M. Bowling. Find-

ing optimal abstract strategies in extensive-form games. In *AAAI*, pp. 1371–1379, 2012.

[15] M. Johanson, K. Waugh, M. Bowling, and M. Zinkevich. Accelerating best response calculation in large extensive games. In *IJCAI*, 2011.

[16] H. K. Khalil. *Nonlinear Control, Global Edition*. Pearson Education, 2015.

[17] Q. Lei, S. G. Nagarajan, I. Panageas, et al. Last iterate convergence in no-regret learning: constrained min-max optimization for convex-concave landscapes. In *AISTATS*, pp. 1441–1449, 2021.

[18] E. Lockhart, M. Lanctot, J. Pérolat, J.-B. Lespiau, D. Morrill, F. Timbers, and K. Tuyls. Computing approximate equilibria in sequential adversarial games by exploitability descent. *arXiv preprint arXiv:1903.05614*, 2019.

[19] P. Mertikopoulos, B. Lecouat, H. Zenati, C.-S. Foo, V. Chandrasekhar, and G. Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *ICLR*, 2019.

[20] P. Mertikopoulos, C. Papadimitriou, and G. Piliouras. Cycles in adversarial regularized learning. In *Symposium on Discrete Algorithms*, pp. 2703–2717, 2018.

[21] J. Perolat, R. Munos, J.-B. Lespiau, S. Omidshafiei, M. Rowland, P. Ortega, N. Burch, T. Anthony, D. Balduzzi, B. De Vylder, et al. From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In *ICML*, pp. 8525–8535, 2021.

[22] F. Timbers, E. Lockhart, M. Lanctot, M. Schmid, J. Schrittwieser, T. Hubert, and M. Bowling. Approximate exploitability: Learning a best response in large games. *arXiv preprint arXiv:2004.09677*, 2020.

[23] C.-Y. Wei, C.-W. Lee, M. Zhang, and H. Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *ICLR*, 2021.

[24] M. Zinkevich, M. Johanson, M. Bowling, and C. Piccione. Regret minimization in games with incomplete information. In *NeurIPS*, pp. 1729–1736, 2007.

付録 A　補題 1 の証明

証明. すべての $\pi \in \Delta(A_1) \times \Delta(A_2)$ について，

$$D_\psi(\pi, \pi^t) = \sum_{i=1}^{2} \left(\psi_i(\pi_i) - \psi_i(\pi_i^t) - \left\langle \nabla\psi_i(\pi_i^t), \pi_i - \pi_i^t \right\rangle \right) \quad (17)$$

である．ψ_i の定義と，最適化問題 $\pi_i^t = \arg\max_{p \in \Delta(A_i)} \left\{ \left\langle z_i^t, p \right\rangle - \psi_i(p) \right\}$ の第一必要条件から，ある $\lambda \in \mathbb{R}$ が存在し，$z_i^t - \nabla\psi_i(\pi_i^t) = \lambda \mathbf{1}$. したがって，$\left\langle z_i^t, \pi_i - \pi_i^t \right\rangle = \left\langle \lambda\mathbf{1} + \nabla\psi_i(\pi_i^t), \pi_i - \pi_i^t \right\rangle =$

$$\langle \lambda\mathbf{1}, \pi_i \rangle - \left\langle \lambda\mathbf{1}, \pi_i^t \right\rangle + \left\langle \nabla\psi_i(\pi_i^t), \pi_i - \pi_i^t \right\rangle = \left\langle \nabla\psi_i(\pi_i^t), \pi_i - \pi_i^t \right\rangle \quad (18)$$

最後の式は任意の $\pi_i \in \Delta(A_1)$ について，$\sum \pi_i(a_i) = 1$ であることを用いている．式 17 に 式 18 を代入することで:

$$\begin{aligned} D_\psi(\pi, \pi^t) &= \sum_{i=1}^{2} \left(\psi_i(\pi_i) - \psi_i(\pi_i^t) - \left\langle z_i^t, \pi_i - \pi_i^t \right\rangle \right) \\ &= \sum_{i=1}^{2} \left(\max_{p \in \Delta(A_i)} \left\{ \left\langle z_i^t, p \right\rangle - \psi_i(p) \right\} - \langle z_i^t, \pi_i \rangle + \psi_i(\pi_i) \right). \end{aligned} \quad (19)$$

$\psi_i^*(z_i) = \max_{p \in \Delta(A_i)} \{\langle z_i, p\rangle - \psi_i(p)\}$ と定義すると，$D_\psi(\pi, \pi^t)$ の時間微分は，以下のように書ける: $\frac{d}{dt} D_\psi(\pi, \pi^t) =$

$$\begin{aligned} &\sum_{i=1}^{2} \frac{d}{dt} \left(\max_{p \in \Delta(A_i)} \left\{ \left\langle z_i^t, p \right\rangle - \psi_i(p) \right\} - \langle z_i^t, \pi_i \rangle + \psi_i(\pi_i) \right) \\ &= \sum_{i=1}^{2} \left(\left\langle \frac{d}{dt} z_i^t, \nabla\psi_i^*(z_i^t) \right\rangle - \left\langle \frac{d}{dt} z_i^t, \pi_i \right\rangle \right) = \sum_{i=1}^{2} \left\langle \frac{d}{dt} z_i^t, \nabla\psi_i^*(z_i^t) - \pi_i \right\rangle. \end{aligned}$$

$\nabla\psi_i^*(z_i)$ の定義と式 4 より，$\nabla\psi_i^*(z_i) = \arg\max_{p \in \Delta(A_i)} \{\langle z_i, p\rangle - \psi_i(p)\}$ および，$\nabla\psi_i^*(z_i^t) = \pi_i^t$ である．よって，$\frac{d}{dt} D_\psi(\pi, \pi^t)$

$$\begin{aligned} &= \sum_{i=1}^{2} \sum_{a_i \in A_i} \left(q_i^{\pi^t}(a_i) + \frac{\mu}{\pi_i^t(a_i)} \left(c_i(a_i) - \pi_i^t(a_i) \right) \right) \left(\pi_i^t(a_i) - \pi_i(a_i) \right) \\ &= \sum_{i=1}^{2} \sum_{a_i \in A_i} \left(\pi_i^t(a_i) - \pi_i(a_i) \right) \sum_{a_{-i} \in A_{-i}} \pi_{-i}^t(a_{-i}) \left(u_i(a_i, a_{-i}) + \mu \frac{c_i(a_i)}{\pi_i^t(a_i)} \right), \end{aligned}$$

この式変式には，$\frac{d}{dt} z_i^t(a_i) = q_i^{\pi^t}(a_i) + \frac{\mu}{\pi_i^t(a_i)} \left(c_i(a_i) - \pi_i^t(a_i) \right)$ であることを用いた．また，最後の式は，すべての $\pi_i \in \Delta(A_i)$ について，$\sum_{a_i \in A_i} \pi_i(a_i) \sum_{a_{-i} \in A_{-i}} \mu\pi_{-i}^t(a_{-i}) = \mu$ であることを用いた．

さらに，$\frac{d}{dt} D_\psi(\pi, \pi^t) =$

$$\begin{aligned} &\sum_{i=1}^{2} \sum_{a_i \in A_i} \left(\pi_i^t(a_i) - \pi_i(a_i) \right) \sum_{a_{-i} \in A_{-i}} \pi_{-i}^t(a_{-i}) \left(u_i(a_i, a_{-i}) + \mu \frac{c_i(a_i)}{\pi_i^t(a_i)} \right) \\ &= \sum_{i=1}^{2} \sum_{a \in A} \pi_i^t(a_i) \pi_{-i}(a_{-i}) u_i(a) + 2\mu - \mu \sum_{i=1}^{2} \sum_{a_i \in A_i} c_i(a_i) \frac{\pi_i(a_i)}{\pi_i^t(a_i)}. \end{aligned}$$

この式変形には，$\sum_{i=1}^{2} \sum_{a \in A} \pi_i^t(a_i) \pi_{-i}^t(a_{-i}) u_i(a_i, a_{-i}) = \sum_{i=1} v_i^{\pi^t} = 0$ であることと，$\sum_{a \in A} \pi_i^t(a_i) \pi_{-i}^t(a_{-i}) \mu \frac{c_i(a_i)}{\pi_i^t(a_i)} = \mu \sum_{a_i \in A_i} c_i(a_i) \frac{\pi_i(a_i)}{\pi_i^t(a_i)} \sum_{a_{-i} \in A_{-i}} \pi_{-i}^t(a_{-i}) = \mu \sum_{a_i \in A_i} c_i(a_i) \frac{\pi_i(a_i)}{\pi_i^t(a_i)}$ であることを用いた. □

付録 B　補題 2 の証明

証明. すべての $i \in \{1, 2\}$，$a_i \in A_i$ について，定常点 $\pi_i^\mu(a_i)$ における RMD のダイナミクスは，以下のように書ける:

$$\begin{aligned} &\pi_i^\mu(a_i) \left(q_i^{\pi^\mu}(a_i) - \langle \pi_i^\mu, q_i^{\pi^\mu} \rangle \right) + \mu \left(c_i(a_i) - \pi_i^\mu(a_i) \right) \\ &= \pi_i^\mu(a_i) \sum_{a_{-i} \in A_{-i}} \pi_{-i}^\mu(a_{-i}) u_i(a_i, a_{-i}) \\ &\quad - \pi_i^\mu(a_i) \sum_{a_i' \in A_i} \sum_{a_{-i} \in A_{-i}} \pi_i^\mu(a_i') \pi_{-i}^\mu(a_{-i}) u_i(a_i', a_{-i}) + \mu \left(c_i(a_i) - \pi_i^\mu(a_i) \right) = 0. \end{aligned}$$

よって，$\pi_i^\mu(a_i)\sum_{a_{-i}\in A_{-i}}\pi_{-i}^\mu(a_{-i})u_i(a_i,a_{-i}) =$

$$\pi_i^\mu(a_i)\sum_{a_i'\in A_i}\sum_{a_{-i}\in A_{-i}}\pi_i^\mu(a_i')\pi_{-i}^\mu(a_{-i})u_i(a_i',a_{-i}) - \mu\left(c_i(a_i)-\pi_i^\mu(a_i)\right). \tag{20}$$

ここで，式 20 の両辺を $\pi_i^\mu(a_i)$ で割るために，$\pi_i^\mu(a_i)>0$ を示す．$i\in\{1,2\}$，$a_i\in A_i$ において，$\pi_i^\mu(a_i)=0$ となる定常点 $\pi_i^\mu(a_i)$ が存在すると仮定する．そのような i と a_i について:

$$\begin{aligned}\frac{d}{dt}\pi_i^\mu(a_i) &= \pi_i^\mu(a_i)\left(q_i^{\pi^\mu}(a_i)-\langle\pi_i^\mu,q_i^{\pi^\mu}\rangle\right)+\mu\left(c_i(a_i)-\pi_i^\mu(a_i)\right)\\ &= \mu c_i(a_i)>0.\end{aligned}$$

が成り立つ．これは，$\frac{d}{dt}\pi_i^\mu(a_i)=0$ であることに矛盾する．したがって，すべての $i\in\{1,2\}$ および $a_i\in A_i$ において，$\pi_i^\mu(a_i)>0$. 以上より，$\sum_{a_{-i}\in A_{-i}}\pi_{-i}^\mu(a_{-i})u_i(a_i,a_{-i}) =$

$$\sum_{a_i'\in A_i}\sum_{a_{-i}\in A_{-i}}\pi_i^\mu(a_i')\pi_{-i}^\mu(a_{-i})u_i(a_i',a_{-i}) - \frac{\mu}{\pi_i^\mu(a_i)}\left(c_i(a_i)-\pi_i^\mu(a_i)\right).$$

すべての $\pi_i'\in\Delta(A_i)$ について，$\sum_{a\in A}\pi_i'(a_i)\pi_{-i}^\mu(a_{-i})u_i(a_i,a_{-i}) =$

$$\begin{aligned}&\sum_{a\in A}\pi_i^\mu(a_i)\pi_{-i}^\mu(a_{-i})u_i(a_i,a_{-i}) - \mu\sum_{a_i\in A_i}\frac{\pi_i'(a_i)}{\pi_i^\mu(a_i)}\left(c_i(a_i)-\pi_i^\mu(a_i)\right)\\ &=\sum_{a\in A}\pi_i^\mu(a_i)\pi_{-i}^\mu(a_{-i})u_i(a_i,a_{-i})+\mu-\mu\sum_{a_i\in A_i}c_i(a_i)\frac{\pi_i'(a_i)}{\pi_i^\mu(a_i)}\end{aligned}$$

□

付録 C　補題 3 の証明

証明. まず，任意の $i\in\{1,2\}$ において，参照戦略を c としたときの (RMD) の定常点を π^c を (RMD) とする．このとき，$c=\pi^c$ ならば，c は与えられたゲームのナッシュ均衡となることを示す．RMD の定義より，すべての $i\in\{1,2\}$ と $a_i\in A_i$ について，定常点 π^c は

$$\pi_i^c(a_i)\left(q_i^{\pi^c}(a_i)-\langle\pi_i^c,q_i^{\pi^c}\rangle\right)+\mu\left(c_i(a_i)-\pi_i^c(a_i)\right)=0,$$

を満たす．このとき，$c=\pi^c$ であれば，$c_i(a_i)\left(q_i^c(a_i)-\langle c_i,q_i^c\rangle\right)=0$ を得る．参照戦略の定義より，すべての $i\in\{1,2\}$ について $c_i(a_i)>0$ が成立し，$v_i^c=\max_{a_i\in A_i}q_i^c(a_i)$ となる．このとき，プレイヤ i の戦略 c_i は，相手の戦略 c_{-i} の最適反応になっているため，c はナッシュ均衡となる．

以上の議論より，$c\notin\Pi^*$ と $c\neq\pi^c$ が常に成り立つ．ここで，$\tilde\pi=\arg\min_{\pi^*\in\Pi^*}\mathrm{KL}(\pi^*,c)$ とする．付録 E にある補題 5 より，$c\neq\pi^c$ ならば，

$$\min_{\pi^*\in\Pi^*}\mathrm{KL}(\pi^*,c)=\mathrm{KL}(\tilde\pi,c)>\mathrm{KL}(\tilde\pi,\pi^c)\geq\min_{\pi^*\in\Pi^*}\mathrm{KL}(\pi^*,\pi^c).$$

したがって，$c\notin\Pi^*$ ならば，$\min_{\pi^*\in\Pi^*}\mathrm{KL}(\pi^*,\pi^c)<\min_{\pi^*\in\Pi^*}\mathrm{KL}(\pi^*,c)$ が成り立つ．

次に，$c\in\Pi^*$ ならば，$\pi^c=c\in\Pi^*$ であることを示す．$c\in\Pi^*$ ならば，$\pi^c\neq c$ を仮定して矛盾を導く．このとき，補題 5 より，すべての $\pi^*\in\Pi^*$ について，$\mathrm{KL}(\pi^*,\pi^c)<\mathrm{KL}(\pi^*,c)$ が成り立つ．また，$c\in\Pi^*$ のときは，あるナッシュ均衡 $\tilde\pi^*$ が存在して，$\mathrm{KL}(\tilde\pi^*,c)=0$ である．よって，$\mathrm{KL}(\tilde\pi^*,\pi^c)<\mathrm{KL}(\tilde\pi^*,c)=0$ となるが，$\mathrm{KL}(\tilde\pi^*,\pi^c)\geq 0$ であることに矛盾する．したがって，$c\in\Pi^*$ ならば，$\pi^c=c$ である．□

付録 D　補題 4 の証明

証明. $c\in\prod_{i=1}^2\Delta^\circ(A_i)$ について，以下の M-FTRL を考える．

$$\begin{aligned}\pi_i^t &= \arg\max_{p\in\Delta(A_i)}\left\{\left\langle z_i^t,p\right\rangle-\psi_i(p)\right\},\\ z_i^t(a_i) &= \int_0^t\left(q_i^{\pi^s}(a_i)+\frac{\mu}{\pi_i^s(a_i)}\left(c_i(a_i)-\pi_i^s(a_i)\right)\right)ds.\end{aligned}$$

式 19 より，$c'\in\prod_{i=1}^2\Delta^\circ(A_i)$ があって，$\frac{d}{dt}D_\psi(\pi^{c'},\pi^t)$

$$\begin{aligned}&=\sum_{i=1}^2\sum_{a_i\in A_i}\left(q_i^{\pi^t}(a_i)+\frac{\mu}{\pi_i^t(a_i)}\left(c_i(a_i)-\pi_i^t(a_i)\right)\right)\left(\pi_i^t(a_i)-\pi_i^{c'}(a_i)\right)\\ &+\sum_{i=1}^2\sum_{a_i\in A_i}\left(\frac{\mu}{\pi_i^t(a_i)}\left(c_i(a_i)-\pi_i^t(a_i)\right)-\frac{\mu}{\pi_i^t(a_i)}\left(c_i'(a_i)-\pi_i^t(a_i)\right)\right)\left(\pi_i^t(a_i)-\pi_i^{c'}(a_i)\right).\end{aligned} \tag{21}$$

式 21 の第一項は以下のようにもかける．

$$\begin{aligned}&\sum_{i=1}^2\sum_{a_i\in A_i}\left(q_i^{\pi^t}(a_i)+\frac{\mu}{\pi_i^t(a_i)}\left(c_i'(a_i)-\pi_i^t(a_i)\right)\right)\left(\pi_i^t(a_i)-\pi_i^{c'}(a_i)\right)\\ &=\sum_{i=1}^2\sum_{a_i\in A_i}\left(\pi_i^t(a_i)-\pi_i^{c'}(a_i)\right)\left(q_i^{\pi^t}(a_i)+\mu\frac{c_i'(a_i)}{\pi_i^t(a_i)}\right)\\ &=\sum_{i=1}^2 v_i^{\pi_i^t,\pi_{-i}^{c'}}+2\mu-\mu\sum_{i=1}^2\sum_{a_i\in A_i}c_i'(a_i)\frac{\pi_i^{c'}(a_i)}{\pi_i^t(a_i)}.\end{aligned}$$

上記の式変形は $\sum_{i=1}^2 v_i^{\pi^t}=0$ と $\mu\sum_{a_i\in A_i}\pi_i^t(a_i)\frac{c_i'(a_i)}{\pi_i^t(a_i)}=\mu\sum_{a_i\in A_i}c_i'(a_i)=\mu$ を用いた．よって，補題 2 より，すべての $i\in\{1,2\}$ について，$v_i^{\pi_i^t,\pi_{-i}^{c'}}=v_i^{\pi^{c'}}+\mu-\mu\sum_{a_i\in A_i}c_i'(a_i)\frac{\pi_i^t(a_i)}{\pi_i^{c'}(a_i)}$．また，式 21 の第一項は，$\sum_{i=1}^2 v_i^{\pi^{c'}}=0$ であることを用いて，以下のようにかける．

$$\begin{aligned}&\sum_{i=1}^2\sum_{a_i\in A_i}\left(q_i^{\pi^t}(a_i)+\frac{\mu}{\pi_i^t(a_i)}\left(c_i'(a_i)-\pi_i^t(a_i)\right)\right)\left(\pi_i^t(a_i)-\pi_i^{c'}(a_i)\right)\\ &=\sum_{i=1}^2 v_i^{\pi^{c'}}+4\mu-\mu\sum_{i=1}^2\sum_{a_i\in A_i}c_i'(a_i)\left(\frac{\pi_i^{c'}(a_i)}{\pi_i^t(a_i)}+\frac{\pi_i^t(a_i)}{\pi_i^{c'}(a_i)}\right)\\ &=-\mu\sum_{i=1}^2\sum_{a_i\in A_i}c_i'(a_i)\left(\sqrt{\frac{\pi_i^t(a_i)}{\pi_i^{c'}(a_i)}}-\sqrt{\frac{\pi_i^{c'}(a_i)}{\pi_i^t(a_i)}}\right)^2.\end{aligned} \tag{22}$$

また，式 21 の第二項は，以下のように変形できる．

$$\begin{aligned}&\sum_{i=1}^2\sum_{a_i\in A_i}\left(\frac{\mu}{\pi_i^t(a_i)}\left(c_i(a_i)-\pi_i^t(a_i)\right)-\frac{\mu}{\pi_i^t(a_i)}\left(c_i'(a_i)-\pi_i^t(a_i)\right)\right)\left(\pi_i^t(a_i)-\pi_i^{c'}(a_i)\right)\\ &=\mu\sum_{i=1}^2\sum_{a_i\in A_i}\frac{1}{\pi_i^t(a_i)}\left(c_i(a_i)-c_i'(a_i)\right)\left(\pi_i^t(a_i)-\pi_i^{c'}(a_i)\right)\\ &\leq\mu\sum_{i=1}^2\sum_{a_i\in A_i}\frac{1}{\pi_i^t(a_i)}\left|c_i(a_i)-c_i'(a_i)\right|.\end{aligned} \tag{23}$$

式 21 に 式 22 と 式 23 を代入することで，

$$\frac{d}{dt}D_\psi(\pi^{c'},\pi^t) \le -\mu\sum_{i=1}^{2}\sum_{a_i\in A_i} c_i'(a_i)\left(\sqrt{\frac{\pi_i^t(a_i)}{\pi_i^{c'}(a_i)}}-\sqrt{\frac{\pi_i^{c'}(a_i)}{\pi_i^t(a_i)}}\right)^2 + \mu\sum_{i=1}^{2}\sum_{a_i\in A_i}\frac{1}{\pi_i^t(a_i)}\left|c_i(a_i)-c_i'(a_i)\right|.$$

開始点を $\pi^0=\pi^c$ とすると，すべての $t\ge 0$ について，$\pi^t=\pi^c$ が成り立つ．このとき，$\frac{d}{dt}D_\psi(\pi^{c'},\pi^t)=0$ であるから，

$$\sum_{i=1}^{2}\sum_{a_i\in A_i} c_i'(a_i)\left(\sqrt{\frac{\pi_i^c(a_i)}{\pi_i^{c'}(a_i)}}-\sqrt{\frac{\pi_i^{c'}(a_i)}{\pi_i^c(a_i)}}\right)^2 \le \sum_{i=1}^{2}\sum_{a_i\in A_i}\frac{1}{\pi_i^c(a_i)}\left|c_i(a_i)-c_i'(a_i)\right|.$$

ここで，c_i は $\Delta(A_i)$ の内点であるから，ある $\nu_1>0$ が存在し，$\forall i,\forall a_i\in A_i,\ c_i(a)>\nu_1$ である．すべての $i\in\{1,2\}$ について，$c_i\in\Delta^\circ(A_i)$ かつ $\mu>0$ ならば，$\pi_i^\mu\in\Delta^\circ(A_i)$ であるから，π_i^c も $\Delta(A_i)$ の内点である．ゆえに，ある $\nu_2>0$ が存在し，$\forall i,\forall a_i\in A_i,\ \pi_i^c(a)>\nu_2$．次に，$\varepsilon>0$ について，$\delta=\frac{\varepsilon^2\nu_1\nu_2}{4\ln 2+\varepsilon^2\nu_2}\sqrt{\frac{1}{\sum_{i=1}^2|A_i|}}$ を定義する．$\|c'-c\|_2<\delta$ ならば，$\|c'-c\|_1\le\|c'-c\|_2\sqrt{\sum_{i=1}^2|A_i|}<\frac{\varepsilon^2\nu_1\nu_2}{4\ln 2+\varepsilon^2\nu_2}$ である．よって，$\forall i,\forall a_i\in A_i,\ c_i'(a_i)>\left(1-\frac{\varepsilon^2\nu_2}{4\ln 2+\varepsilon^2\nu_2}\right)\nu_1>0$ となる．以上より，$\|c'-c\|_2<\delta$ ならば，$\sum_{i=1}^2\sum_{a\in A_i}c_i'(a)\left(\sqrt{\frac{\pi_i^{c'}(a)}{\pi_i^c(a)}}-\sqrt{\frac{\pi_i^c(a_i)}{\pi_i^{c'}(a)}}\right)^2=$

$$\sum_{i=1}^{2}\sum_{a\in A_i}\frac{c_i'(a)}{\pi_i^{c'}(a)}\frac{(\pi_i^c(a)-\pi_i^{c'}(a))^2}{\pi_i^c(a)} \ge \left(1-\frac{\varepsilon^2\nu_2}{4\ln 2+\varepsilon^2\nu_2}\right)\nu_1\sum_{i=1}^{2}\sum_{a\in A_i}\frac{(\pi_i^c(a)-\pi_i^{c'}(a))^2}{\pi_i^c(a)} \ge \left(1-\frac{\varepsilon^2\nu_2}{4\ln 2+\varepsilon^2\nu_2}\right)\nu_1\sum_{i=1}^{2}\mathrm{KL}(\pi_i^{c'},\pi_i^c).$$

さらに，

$$\sum_{i=1}^{2}\mathrm{KL}(\pi_i^{c'},\pi_i^c)\ge\frac{1}{2\ln 2}\sum_{i=1}^{2}\|\pi_i^{c'}-\pi_i^c\|_1^2\ge\frac{1}{4\ln 2}\|\pi^{c'}-\pi^c\|_1^2.$$

この式変形には，[7] の Lemma 11.6.1 と，$a,b\ge 0$ について，$(a^2+b^2)\ge\frac{1}{2}(a+b)^2$ が成り立つことを用いた．よって，

$$\left(1-\frac{\varepsilon^2\nu_2}{4\ln 2+\varepsilon^2\nu_2}\right)\frac{\nu_1}{4\ln 2}\|\pi^{c'}-\pi^c\|_1^2 \le\sum_{i=1}^{2}\sum_{a_i\in A_i}\frac{1}{\pi_i^c(a_i)}\left|c_i(a_i)-c_i'(a_i)\right|<\frac{1}{\nu_2}\|c'-c\|_1.$$

したがって，$\|c'-c\|<\delta$ のとき，$\|\pi^{c'}-\pi^c\|_2\le\|\pi^{c'}-\pi^c\|_1$

$$<\sqrt{\frac{4\ln 2}{\left(1-\frac{\varepsilon^2\nu_2}{4\ln 2+\varepsilon^2\nu_2}\right)\nu_1\nu_2}}\|c'-c\|_1 =\sqrt{\frac{4\ln 2}{1-\frac{\varepsilon^2\nu_2}{4\ln 2+\varepsilon^2\nu_2}}\frac{\varepsilon^2}{4\ln 2+\varepsilon^2\nu_2}}=\varepsilon.$$

よって，すべての $\varepsilon>0$ について，ある $\delta>0$ が存在し，すべての $c'\in\prod_{i=1}^2\Delta^\circ(A_i)$ について，$\|c'-c\|_2<\delta$ ならば，$\|\pi^{c'}-\pi^c\|_2<\varepsilon$ が成り立つ．したがって，$F(\cdot)$ は $\prod_{i=1}^2\Delta^\circ(A_i)$ 上の連続関数である． □

付録 E　補題 5

補題 5. 任意の $i\in\{1,2\}$ において，π_i^c を，相対変異確率が c の (RMD) の定常点とする．$c\ne\pi^c$ のとき，オリジナルゲームのすべてのナッシュ均衡 π^* で，$\mathrm{KL}(\pi^*,\pi^c)-\mathrm{KL}(\pi^*,c)<0$, が成り立つ．

証明. Kullback-Leibler divergence の差，

$$\mathrm{KL}(\pi^*,\pi^c)-\mathrm{KL}(\pi^*,c)=\sum_{i=1}^{2}\sum_{a_i\in A_i}\pi_i^*(a_i)\ln\frac{c_i(a_i)}{\pi_i^c(a_i)}\le\sum_{i=1}^{2}\ln\left(\sum_{a_i\in A_i}\pi_i^*(a_i)\frac{c_i(a_i)}{\pi_i^c(a_i)}\right).$$

この不等式は，凹関数 $\ln(\cdot)$ の性質と，凹関数におけるイェンセンの不等式を用いた．さらに，$\ln(\cdot)$ は狭義の凹関数であるから，すべての $i\in\{1,2\}$ について，$\frac{c_i(a_1)}{\pi_i^c(a_1)}=\frac{c_i(a_2)}{\pi_i^c(a_2)}=\frac{c_i(a_{|A_i|})}{\pi_i^c(a_{|A_i|})}$ が成り立つ．$c\ne\pi^c$ ならば，ある $i\in\{1,2\}$ と $a_i,a_i'\in A_i$ が存在し，$\frac{c_i(a_i)}{\pi_i^c(a_i)}\ne\frac{c_i(a_i')}{\pi_i^c(a_i')}$ である．したがって，

$$\mathrm{KL}(\pi^*,\pi^c)-\mathrm{KL}(\pi^*,c)<\sum_{i=1}^{2}\ln\left(\sum_{a_i\in A_i}\pi_i^*(a_i)\frac{c_i(a_i)}{\pi_i^c(a_i)}\right). \tag{24}$$

ここで，(RMD) から，すべての $i\in\{1,2\}$ と $a_i\in A_i$ について，$\pi_i^c(a_i)\left(q_i^{\pi^c}(a_i)-\langle\pi_i^c,q_i^{\pi^c}\rangle\right)+\mu\left(c_i(a_i)-\pi_i^c(a_i)\right)=0$. 式を変形すると，

$$\frac{c_i(a_i)}{\pi_i^c(a_i)}=1-\frac{1}{\mu}\left(q_i^{\pi^c}(a_i)-\langle\pi_i^c,q_i^{\pi^c}\rangle\right). \tag{25}$$

式 25 を式 24 に代入すると，$\mathrm{KL}(\pi^*,\pi^c)-\mathrm{KL}(\pi^*,c)$

$$<\sum_{i=1}^{2}\ln\left(\sum_{a_i\in A_i}\pi_i^*(a_i)\left(1-\frac{1}{\mu}\left(q_i^{\pi^c}(a_i)-\langle\pi_i^c,q_i^{\pi^c}\rangle\right)\right)\right) =\sum_{i=1}^{2}\ln\left(1-\frac{1}{\mu}\left(v_i^{\pi_i^*,\pi_{-i}^c}-v_i^{\pi^c}\right)\right).$$

$\sum_{a_i\in A_i}\pi_i^*(a_i)\frac{c_i(a_i)}{\pi_i^c(a_i)}>0$ であるから，$1-\frac{1}{\mu}\left(v_i^{\pi_i^*,\pi_{-i}^c}-v_i^{\pi^c}\right)>0$. π^* はナッシュ均衡のであるから以下の式が成り立つ．

$$\sum_{i=1}^{2}\left(1-\frac{1}{\mu}\left(v_i^{\pi_i^*,\pi_{-i}^c}-v_i^{\pi^c}\right)\right)=2-\frac{1}{\mu}\sum_{i=1}^{2}v_i^{\pi_i^*,\pi_{-i}^c}\le 2.$$

よって，ある定数 $\alpha\in(0,2]$ と $x\in(0,\alpha)$ が存在し:

$$\sum_{i=1}^{2}\ln\left(1-\frac{1}{\mu}\left(v_i^{\pi_i^*,\pi_{-i}^c}-v_i^{\pi^c}\right)\right)=\ln(\alpha-x)+\ln(x).$$

したがって，$\mathrm{KL}(\pi^*,\pi^c)-\mathrm{KL}(\pi^*,c)$

$$<\max_{\alpha\in(0,2]}\max_{x\in(0,\alpha)}\{\ln(\alpha-x)+\ln(x)\} =\max_{\alpha\in(0,2]}\max_{x\in(0,\alpha)}\ln\left(-\left(x-\frac{\alpha}{2}\right)^2+\frac{\alpha^2}{4}\right)=\max_{\alpha\in(0,2]}\ln\left(\frac{\alpha^2}{4}\right)\le 0.$$

□