Mutation-Driven Follow the Regularized Leader for Last-Iterate Convergence in Zero-Sum Games (Supplementary Material)

Kenshi Abe[1] **Mitsuki Sakamoto**[2] **Atsushi Iwasaki**[2]

[1]CyberAgent, Inc.
[2]University of Electro-Communications

A UNBIASED ESTIMATOR FOR FTRL AND O-FTRL UNDER BANDIT FEEDBACK

For FTRL and O-FTRL under bandit feedback, we use the following unbiased estimator of $q_i^{\pi^t}$ which is proposed by [Lattimore and Szepesvári, 2020]:

$$\hat{q}_i^{\pi^t}(a_i) = u_{\max} - \frac{u_{\max} - u_i(a_1^t, a_2^t)}{\pi_i^t(a_i^t)} \mathbb{1}[a_i = a_i^t].$$

This estimator takes values in $(-\infty, u_{\max}]$ while the standard importance-weighted estimator takes values in $(-\infty, \infty)$.

B SENSITIVITY ANALYSIS ON MUTATION PARAMETERS

In this section, we investigate the performance of M-FTRL with a fixed reference strategy with varying $\mu \in \{10^{-3}, 5 \times 10^{-3}, 10^{-2}, 10^{-1}, 1\}$. We set the reference strategy to $c_i = \left(\frac{1}{|A_i|}\right)_{a_i \in A_i}$, and set the learning rate to $\eta = 10^{-1}$. The initial strategy profile π^0 is generated uniformly at random in $\prod_{i=1}^2 \Delta^\circ(A_i)$ for each instance. We conduct experiments on BRPS under full-information feedback. Figure 1 shows the average exploitability of π^t for 100 instances. This result highlights the trade-off between the convergence rate and exploitability as shown in Theorem 5.4.

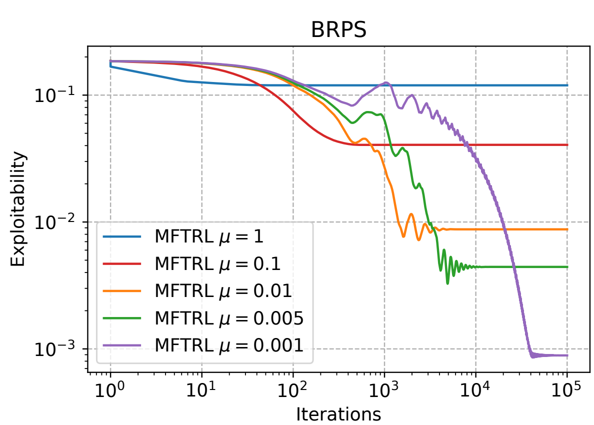

Figure 1: Exploitability of π^t for M-FTRL with a fixed reference strategy in BRPS under full-information feedback.

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

C ADDITIONAL LEMMAS

Lemma C.1. *For any $\pi \in \prod_{i=1}^2 \Delta(A_i)$, π^t updated by M-FTRL satisfies that:*

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

Lemma C.2. *Let $\pi^\mu \in \prod_{i=1}^2 \Delta(A_i)$ be a stationary point of (RMD). For a player $i \in \{1, 2\}$, if $c_i \in \Delta^\circ(A_i)$ and $\mu > 0$, then we also have $\pi_i^\mu \in \Delta^\circ(A_i)$.*

D PROOFS

D.1 PROOF OF THEOREM 5.1

Proof of Theorem 5.1. By the method of Lagrange multiplier, we have:

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

Therefore, the time derivative of $\pi_i^t(a_i)$ is given as follows:

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

From the definition of $z_i^t(a_i)$, we have:

$$\frac{d}{dt} z_i^t(a_i) = q_i^{\pi^t}(a_i) + \frac{\mu}{\pi_i^t(a_i)} \left(c_i(a_i) - \pi_i^t(a_i)\right).$$

By combining these equalities, we get:

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

□

D.2 PROOF OF LEMMA 5.5

Proof of Lemma 5.5. Let us define $\psi_i^*(z_i) = \max_{p \in \Delta(A_i)} \{\langle z_i, p\rangle - \psi_i(p)\}$. Then, from Lemma C.1, the time derivative of $D_\psi(\pi, \pi^t)$ is given as:

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

From the maximizing argument of [Shalev-Shwartz, 2011], we have $\nabla \psi_i^*(z_i) = \underset{p \in \Delta(A_i)}{\arg\max} \{\langle z_i, p\rangle - \psi_i(p)\}$ and then $\nabla \psi_i^*(z_i^t) = \pi_i^t$. Furthermore, from the definition of $z_i^t(a_i)$, we have $\frac{d}{dt} z_i^t(a_i) = q_i^{\pi^t}(a_i) + \frac{\mu}{\pi_i^t(a_i)} \left(c_i(a_i) - \pi_i^t(a_i)\right)$. Then,

$$\begin{aligned}
\frac{d}{dt} D_\psi(\pi, \pi^t) &= \sum_{i=1}^2 \left\langle \frac{d}{dt} z_i^t, \pi_i^t - \pi_i \right\rangle \\
&= \sum_{i=1}^2 \sum_{a_i \in A_i} \left(q_i^{\pi^t}(a_i) + \frac{\mu}{\pi_i^t(a_i)} \left(c_i(a_i) - \pi_i^t(a_i)\right) \right) \left(\pi_i^t(a_i) - \pi_i(a_i)\right) \\
&= \sum_{i=1}^2 \sum_{a_i \in A_i} \left(\pi_i^t(a_i) - \pi_i(a_i)\right) \left(q_i^{\pi^t}(a_i) + \mu \left(\frac{c_i(a_i)}{\pi_i^t(a_i)} - 1 \right) \right) \\
&= \sum_{i=1}^2 \sum_{a_i \in A_i} \left(\pi_i^t(a_i) - \pi_i(a_i)\right) \left(q_i^{\pi^t}(a_i) + \mu \frac{c_i(a_i)}{\pi_i^t(a_i)} \right) \\
&= \sum_{i=1}^2 \left(v_i^{\pi^t} - v_i^{\pi_i, \pi_{-i}^t} + \mu \sum_{a_i \in A_i} \left(\pi_i^t(a_i) - \pi_i(a_i)\right) \frac{c_i(a_i)}{\pi_i^t(a_i)} \right) \\
&= -\sum_{i=1}^2 v_i^{\pi_i, \pi_{-i}^t} + 2\mu - \mu \sum_{i=1}^2 \sum_{a_i \in A_i} c_i(a_i) \frac{\pi_i(a_i)}{\pi_i^t(a_i)} \\
&= \sum_{i=1}^2 v_i^{\pi_i^t, \pi_{-i}} + 2\mu - \mu \sum_{i=1}^2 \sum_{a_i \in A_i} c_i(a_i) \frac{\pi_i(a_i)}{\pi_i^t(a_i)},
\end{aligned}$$

where the sixth equality follows from $\sum_{i=1}^2 v_i^{\pi^t} = 0$ and $\mu \sum_{a \in A} \pi_i^t(a_i) \frac{c_i(a_i)}{\pi_i^t(a_i)} = \mu \sum_{a \in A} c_i(a_i) = \mu$, and the last equality follows from $v_1^{\pi_1, \pi_2^t} = -v_2^{\pi_1, \pi_2^t}$ and $v_2^{\pi_1^t, \pi_2} = -v_1^{\pi_1^t, \pi_2}$ by the definition of two-player zero-sum games. □

D.3 PROOF OF LEMMA 5.6

Proof of Lemma 5.6. By using the ordinary differential equation (RMD), we have for all $i \in \{1, 2\}$ and $a_i \in A_i$:

$$\pi_i^\mu(a_i) \left(q_i^{\pi^\mu}(a_i) - v_i^{\pi^\mu} \right) + \mu \left(c_i(a_i) - \pi_i^\mu(a_i)\right) = 0.$$

Then, we get:

$$q_i^{\pi^\mu}(a_i) = v_i^{\pi^\mu} - \frac{\mu}{\pi_i^\mu(a_i)} \left(c_i(a_i) - \pi_i^\mu(a_i)\right).$$

Note that from Lemma C.2, $\frac{1}{\pi_i^\mu(a_i)}$ is well-defined. Then, for any $\pi_i' \in \Delta(A_i)$ we have:

$$\begin{aligned}
v_i^{\pi_i', \pi_{-i}^\mu} &= \sum_{a_i \in A_i} \pi_i'(a_i) q_i^{\pi^\mu}(a_i) \\
&= v_i^{\pi^\mu} - \mu \sum_{a_i \in A_i} \frac{\pi_i'(a_i)}{\pi_i^\mu(a_i)} (c_i(a_i) - \pi_i^\mu(a_i)) \\
&= v_i^{\pi^\mu} + \mu - \mu \sum_{a_i \in A_i} c_i(a_i) \frac{\pi_i'(a_i)}{\pi_i^\mu(a_i)}.
\end{aligned}$$

□

D.4 PROOF OF THEOREM 5.2

Proof of Theorem 5.2. First, we prove the first part of the theorem. By setting $\pi = \pi^\mu$ in Lemma 5.5 and $\pi' = \pi^t$ in Lemma 5.6, we have:

$$\begin{aligned}
\frac{d}{dt} D_\psi(\pi^\mu, \pi^t) =& \sum_{i=1}^2 v_i^{\pi_i^t, \pi_{-i}^\mu} + 2\mu - \mu \sum_{i=1}^2 \sum_{a_i \in A_i} c_i(a_i) \frac{\pi_i^\mu(a_i)}{\pi_i^t(a_i)} \\
=& \sum_{i=1}^2 v_i^{\pi^\mu} + 4\mu - \mu \sum_{i=1}^2 \sum_{a_i \in A_i} c_i(a_i) \left(\frac{\pi_i^t(a_i)}{\pi_i^\mu(a_i)} + \frac{\pi_i^\mu(a_i)}{\pi_i^t(a_i)} \right) \\
=& 4\mu - \mu \sum_{i=1}^2 \sum_{a_i \in A_i} c_i(a_i) \left(\frac{\pi_i^t(a_i)}{\pi_i^\mu(a_i)} + \frac{\pi_i^\mu(a_i)}{\pi_i^t(a_i)} \right) \\
=& 4\mu - \mu \sum_{i=1}^2 \sum_{a_i \in A_i} c_i(a_i) \left(\left(\sqrt{\frac{\pi_i^t(a_i)}{\pi_i^\mu(a_i)}} - \sqrt{\frac{\pi_i^\mu(a_i)}{\pi_i^t(a_i)}} \right)^2 + 2 \right) \\
=& -\mu \sum_{i=1}^2 \sum_{a_i \in A_i} c_i(a_i) \left(\sqrt{\frac{\pi_i^t(a_i)}{\pi_i^\mu(a_i)}} - \sqrt{\frac{\pi_i^\mu(a_i)}{\pi_i^t(a_i)}} \right)^2,
\end{aligned}$$

where the third equality follows from $\sum_{i=1}^2 v_i^{\pi^\mu} = 0$ by the definition of zero-sum games.

Next, we prove the second part of the theorem. From the first part of the theorem, we have:

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

where the second inequality follows from $x \geq \ln(1+x)$ for all $x > 0$, and the third inequality follows from the concavity of the $\ln(\cdot)$ function and Jensen's inequality for concave functions. On the other hand, when $\psi_i(p) = \sum_{a_i \in A_i} p(a_i) \ln p(a_i)$, $D_{\psi_i}(\pi_i^\mu, \pi_i^t) = \mathrm{KL}(\pi_i^\mu, \pi_i^t)$. Thus, we have $D_\psi(\pi^\mu, \pi^t) = \sum_{i=1}^2 \mathrm{KL}(\pi_i^\mu, \pi_i^t)$. From this fact and (1), we have:

$$\frac{d}{dt}\mathrm{KL}(\pi^\mu, \pi^t) \leq -\mu \left(\min_{i \in \{1,2\}, a_i \in A_i} \frac{c_i(a_i)}{\pi_i^\mu(a_i)} \right) \mathrm{KL}(\pi^\mu, \pi^t).$$

□

E PROOFS OF ADDITIONAL LEMMAS

E.1 PROOF OF LEMMA C.1

Proof of Lemma C.1. First, for any $\pi \in \prod_{i=1}^2 \Delta(A_i)$,

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

From the assumptions on ψ_i and the first-order necessary conditions for the optimization problem of $\underset{p \in \Delta(A_i)}{\arg\max} \{\langle z_i^t, p\rangle - \psi_i(p)\}$, for $\pi_i^t = \underset{p \in \Delta(A_i)}{\arg\max} \{\langle z_i^t, p\rangle - \psi_i(p)\}$, there exists $\lambda \in \mathbb{R}$ such that

$$z_i^t - \nabla \psi_i(\pi_i^t) = \lambda \mathbf{1}.$$

Therefore, we have:

$$\left\langle z_i^t, \pi_i - \pi_i^t \right\rangle = \left\langle \lambda \mathbf{1} + \nabla \psi_i(\pi_i^t), \pi_i - \pi_i^t \right\rangle = \left\langle \nabla \psi_i(\pi_i^t), \pi_i - \pi_i^t \right\rangle. \tag{3}$$

By combining (2) and (3):

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

□

E.2 PROOF OF LEMMA C.2

Proof of Lemma C.2. We assume that there exists $i \in \{1,2\}$ and $a_i \in A_i$ such that $\pi_i^\mu(a_i) = 0$. Then, for such i and a_i, we have:

$$\frac{d}{dt}\pi_i^\mu(a_i) = \pi_i^\mu(a_i) \left(q_i^{\pi^\mu}(a_i) - v_i^{\pi^\mu} \right) + \mu \left(c_i(a_i) - \pi_i^\mu(a_i) \right) = \mu c_i(a_i) > 0.$$

This contradicts that $\frac{d}{dt}\pi_i^\mu(a_i) = 0$ since π^μ is a stationary point. Therefore, for all $i \in \{1,2\}$ and $a_i \in A_i$, we have $\pi_i^\mu(a_i) > 0$. □

References

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2): 107–194, 2011.