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

$$\hat{q}_t^\pi(a_i) = u_{\max} - \frac{u_{\max} - u_i(a_t^1, a_t^2)}{\pi_t^i(a_t^i)} \mathbb{I}[a_i = a_t^i].$$

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 $\pi^0$ is generated uniformly at random in $\prod_{i=1}^n \Delta^0(A_i)$ for each instance. We conduct experiments on BRPS under full-information feedback. Figure 1 shows the average exploitability of $\pi^t$ 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 $\pi^t$ 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).
Therefore, the time derivative of \( \pi \) is given as follows:

\[
\frac{d}{dt} \pi(t) = \sum_{i=1}^{2} \left( \max_{p \in \Delta(A_i)} \left\{ \langle z_i^t, p \rangle - \psi_i(p) \right\} - \langle z_i^t, \pi_i \rangle + \psi_i(\pi_i) \right).
\]

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 \( \pi^\mu_i \in \Delta^\circ(A_i) \).

D PROOFS

D.1 PROOF OF THEOREM 5.1

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

\[
\pi^t_i(a_i) = \frac{\exp(z_i^t(a_i))}{\sum_{a_i' \in A_i} \exp(z_i^t(a_i'))}.
\]

Therefore, the time derivative of \( \pi^t_i(a_i) \) is given as follows:

\[
\frac{d}{dt} \pi^t_i(a_i) = \frac{\frac{d}{dt} \exp(z_i^t(a_i))}{\sum_{a_i' \in A_i} \exp(z_i^t(a_i'))} - \frac{\exp(z_i^t(a_i))\frac{d}{dt} \left( \sum_{a_i' \in A_i} \exp(z_i^t(a_i')) \right)}{\left( \sum_{a_i' \in A_i} \exp(z_i^t(a_i')) \right)^2} \\
= \frac{\exp(z_i^t(a_i)) \frac{d}{dt} z_i^t(a_i)}{\sum_{a_i' \in A_i} \exp(z_i^t(a_i'))} - \frac{\exp(z_i^t(a_i))\left( \sum_{a_i' \in A_i} \exp(z_i^t(a_i')) \frac{d}{dt} z_i^t(a_i') \right)}{\left( \sum_{a_i' \in A_i} \exp(z_i^t(a_i')) \right)^2} \\
= \pi^t_i(a_i) \frac{d}{dt} z_i^t(a_i) - \pi^t_i(a_i) \sum_{a_i' \in A_i} \pi^t_i(a_i') \frac{d}{dt} z_i^t(a_i').
\]

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

\[
\frac{d}{dt} z_i^t(a_i) = q_i^t(a_i) + \frac{\mu}{\pi^t_i(a_i)} (c_i(a_i) - \pi^t_i(a_i)).
\]

By combining these equalities, we get:

\[
\frac{d}{dt} \pi^t_i(a_i) = \pi^t_i(a_i) \left( q_i^t(a_i) + \frac{\mu}{\pi^t_i(a_i)} (c_i(a_i) - \pi^t_i(a_i)) - \sum_{a_i' \in A_i} \pi^t_i(a_i') \left( q_i^t(a_i') + \frac{\mu}{\pi^t_i(a_i')} (c_i(a_i') - \pi^t_i(a_i')) \right) \right) \\
= \pi^t_i(a_i) \left( q_i^t(a_i) - v_i^t \right) + \mu (c_i(a_i) - \pi^t_i(a_i)) - \mu \pi^t_i(a_i) \sum_{a_i' \in A_i} (c_i(a_i') - \pi^t_i(a_i')) \\
= \pi^t_i(a_i) \left( q_i^t(a_i) - v_i^t \right) + \mu (c_i(a_i) - \pi^t_i(a_i)).
\]

\(\square\)
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:

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

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

\[
\frac{d}{dt} D_\psi(\pi, \pi^t) = \sum_{i=1}^2 \left( \frac{d}{dt} z_i^t, \pi^t_i - \pi_i \right) 
= \sum_{i=1}^2 \sum_{a_i \in A_i} \left( q_i^\pi(a_i) + \mu \frac{c_i(a_i)}{\pi^t_i(a_i)} (c_i(a_i) - \pi^t_i(a_i)) \right) \left( \pi^t_i(a_i) - \pi_i(a_i) \right) 
= \sum_{i=1}^2 \left( q_i^\pi(a_i) + \mu \frac{c_i(a_i)}{\pi^t_i(a_i)} (c_i(a_i) - \pi^t_i(a_i)) \right) \left( \pi^t_i(a_i) - \pi_i(a_i) \right) 
= \sum_{i=1}^2 \left( q_i^\pi(a_i) + \mu \frac{c_i(a_i)}{\pi^t_i(a_i)} (c_i(a_i) - \pi^t_i(a_i)) \right) \left( \pi^t_i(a_i) - \pi_i(a_i) \right) 
= \sum_{i=1}^2 \left( q_i^\pi(a_i) - q_i^\pi_{\pi_i} \pi^t_i + \mu \sum_{a_i \in A_i} \left( \pi^t_i(a_i) - \pi_i(a_i) \right) \frac{c_i(a_i)}{\pi^t_i(a_i)} \right) 
= -2 \sum_{i=1}^2 q_i^\pi_{\pi_i, \pi^t_i} - 2 \mu \sum_{i=1}^2 \sum_{a_i \in A_i} \frac{c_i(a_i) \pi^t_i(a_i)}{\pi^t_i(a_i)} 
= 2 \sum_{i=1}^2 q_i^\pi_{\pi_i, \pi^t_i} + 2 \mu \sum_{i=1}^2 \sum_{a_i \in A_i} \frac{c_i(a_i) \pi^t_i(a_i)}{\pi^t_i(a_i)},
\]

where the sixth equality follows from \( \sum_{i=1}^2 q_i^\pi = 0 \) and \( \mu \sum_{a_i \in A_i} \frac{c_i(a_i) \pi^t_i(a_i)}{\pi^t_i(a_i)} = \mu \sum_{a_i \in A_i} c_i(a_i) = \mu \), and the last equality follows from \( q_i^\pi_{\pi_i, \pi^t_i} = -q_i^\pi_{\pi_i, \pi^t_i} \) and \( q_i^\pi_{\pi_i, \pi^t_i} = -q_i^\pi_{\pi_i, \pi^t_i} \) 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 \( a_i \in A_i \): \[
\pi^t_i(a_i) \left( q_i^{\pi_i}(a_i) - v_i^{\pi_i} \right) + \mu \left( c_i(a_i) - \pi^t_i(a_i) \right) = 0.
\]

Then, we get:

\[
q_i^{\pi_i}(a_i) = v_i^{\pi_i} - \frac{\mu}{\pi^t_i(a_i)} \left( c_i(a_i) - \pi^t_i(a_i) \right).
\]
Note that from Lemma [2], $\frac{1}{\pi_i(a_i)}$ is well-defined. Then, for any $\pi'_i \in \Delta(A_i)$ we have:

$$
v'_i \frac{\pi'_i, \pi''_i}{\pi_i(a_i)} = \sum_{a_i \in A_i} \pi'_i(a_i) q'_i(a_i) = v'_i + \frac{\mu}{\pi'_i(a_i)} \sum_{a_i \in A_i} c_i(a_i) \left( \pi'_i(a_i) - \pi''_i(a_i) \right)
$$

$$
= v'_i + \frac{\mu}{\pi'_i(a_i)} \sum_{a_i \in A_i} c_i(a_i) \pi'_i(a_i)
$$

\begin{proof}[Proof of Theorem 5.2]

First, we prove the first part of the theorem. By setting $\pi = \pi''$ in Lemma 5.5 and $\pi' = \pi'$ in Lemma 5.6 we have:

$$
\frac{d}{dt} D_\phi(\pi''', \pi') = \sum_{i=1}^{2} v'_i \frac{\pi'_i, \pi'''_i}{\pi_i(a_i)} + 2\mu \sum_{i=1}^{2} c_i(a_i) \frac{\pi''_i(a_i)}{\pi'_i(a_i)}
$$

$$
= \sum_{i=1}^{2} v'_i + 4\mu \sum_{i=1}^{2} c_i(a_i) \left( \frac{\pi''_i(a_i)}{\pi'_i(a_i)} + \frac{\pi''_i(a_i)}{\pi'_i(a_i)} \right)
$$

$$
= 4\mu \sum_{i=1}^{2} c_i(a_i) \left( \frac{\pi''_i(a_i)}{\pi'_i(a_i)} + \frac{\pi''_i(a_i)}{\pi'_i(a_i)} \right)
$$

$$
= -\mu \sum_{i=1}^{2} c_i(a_i) \left( \frac{\pi''_i(a_i)}{\pi'_i(a_i)} - \frac{\pi''_i(a_i)}{\pi'_i(a_i)} \right)^2,
$$

where the third equality follows from $\sum_{i=1}^{2} v''_i = 0$ 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:

$$
\frac{d}{dt} D_\phi(\pi'', \pi') = -\mu \sum_{i=1}^{2} c_i(a_i) \left( \frac{\pi''_i(a_i)}{\pi'_i(a_i)} + \frac{\pi''_i(a_i)}{\pi'_i(a_i)} - 2 \right)
$$

$$
\leq -\mu \sum_{i=1}^{2} \left( \min_{a_i \in A_i} \frac{c_i(a_i)}{\pi''_i(a_i)} \right) \sum_{a_i \in A_i} \pi''_i(a_i) \left( \frac{\pi''_i(a_i)}{\pi'_i(a_i)} + \frac{\pi''_i(a_i)}{\pi'_i(a_i)} - 2 \right)
$$

$$
= -\mu \sum_{i=1}^{2} \left( \min_{a_i \in A_i} \frac{c_i(a_i)}{\pi''_i(a_i)} \right) \sum_{a_i \in A_i} \left( \pi''_i(a_i) - \pi''_i(a_i) \right)^2 \pi'_i(a_i)
$$

$$
\leq -\mu \sum_{i=1}^{2} \left( \min_{a_i \in A_i} \frac{c_i(a_i)}{\pi''_i(a_i)} \right) \ln \left( 1 + \sum_{a_i \in A_i} \left( \pi''_i(a_i) - \pi''_i(a_i) \right)^2 \pi'_i(a_i) \right)
$$

$$
= -\mu \sum_{i=1}^{2} \left( \min_{a_i \in A_i} \frac{c_i(a_i)}{\pi''_i(a_i)} \right) \ln \left( \sum_{a_i \in A_i} \pi''_i(a_i) \pi''_i(a_i) \pi'_i(a_i) \right)
$$

$$
\leq -\mu \sum_{i=1}^{2} \left( \min_{a_i \in A_i} \frac{c_i(a_i)}{\pi''_i(a_i)} \right) \sum_{a_i \in A_i} \pi''_i(a_i) \ln \left( \frac{\pi''_i(a_i)}{\pi'_i(a_i)} \right)
$$

$$
= -\mu \sum_{i=1}^{2} \left( \min_{a_i \in A_i} \frac{c_i(a_i)}{\pi''_i(a_i)} \right) \text{KL}(\pi''_i, \pi'_i) \leq -\mu \left( \min_{i \in \{1, 2\}, a_i \in A_i} \frac{c_i(a_i)}{\pi''_i(a_i)} \right) \sum_{i=1}^{2} \text{KL}(\pi''_i, \pi'_i),
$$

(1)

\end{proof}
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 \( \psi_i(p) = \sum_{a_i \in A_i} p(a_i) \ln p(a_i) \), \( D_\psi(\pi_i^\mu, \pi_i^t) = \text{KL}(\pi_i^\mu, \pi_i^t) \). Thus, we have \( D_\psi(\pi_i^\mu, \pi_i^t) = \sum_{i=1}^{2} \text{KL}(\pi_i^\mu, \pi_i^t) \). From this fact and (1), we have:

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

\[\blacktriangle\]

**E PROOFS OF ADDITIONAL LEMMAS**

**E.1 PROOF OF LEMMA C.1**

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

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

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

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

Therefore, we have:

\[
\langle z_i^t, \pi_i - \pi_i^t \rangle = \langle \lambda \mathbf{1} + \nabla \psi_i(\pi_i^t), \pi_i - \pi_i^t \rangle = \langle \nabla \psi_i(\pi_i^t), \pi_i - \pi_i^t \rangle.
\]

By combining (2) and (3):

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

\[
= \sum_{i=1}^{2} \left( \langle z_i^t, \pi_i^t \rangle - \psi_i(\pi_i^t) - \langle z_i^t, \pi_i \rangle + \psi_i(\pi_i) \right)
\]

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

\[\blacktriangle\]

**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:

\[
\frac{d}{dt} \pi_i^\mu(a_i) = \pi_i^\mu(a_i) \left( q_i^\mu(a_i) - \psi_i^\mu \right) + \mu \left( c_i(a_i) - \pi_i^\mu(a_i) \right) = \mu c_i(a_i) > 0.
\]

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

\[\blacktriangle\]

**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.