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 $q^\pi_t$ which is proposed by [Lattimore and Szepesvári 2020]:

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

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 $\pi^0$ 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.](image-url)
C ADDITIONAL LEMMAS

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

$$D(\pi, \pi^t) = \frac{2}{\prod_{i=1}^{2} (\max_{p \in \Delta(A_i)} \{ \langle z_i^t, p \rangle - \psi_i(p) \} - \langle z_i^t, \pi_i \rangle + \psi_i(\pi_i))}.$$

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

D PROOFS

D.1 PROOF OF THEOREM 5.1

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

$$\pi_i^t(a_i) = \frac{\exp(z_i^t(a_i))}{\sum_{a_i \in A_i} \exp(z_i^t(a_i))},$$

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

$$\frac{d}{dt} \pi_i^t(a_i) = \frac{\frac{d}{dt} \exp(z_i^t(a_i)) \sum_{a_i \in A_i} \exp(z_i^t(a_i))}{\sum_{a_i \in A_i} \exp(z_i^t(a_i))} - \frac{\exp(z_i^t(a_i)) \frac{d}{dt} \left( \sum_{a_i \in A_i} \exp(z_i^t(a_i)) \right)}{\left( \sum_{a_i \in A_i} \exp(z_i^t(a_i)) \right)^2}$$

$$= \frac{\exp(z_i^t(a_i)) \frac{d}{dt} \sum_{a_i \in A_i} \exp(z_i^t(a_i))}{\sum_{a_i \in A_i} \exp(z_i^t(a_i))} - \frac{\exp(z_i^t(a_i)) \sum_{a_i \in A_i} \exp(z_i^t(a_i)) \frac{d}{dt} z_i^t(a_i)}{\left( \sum_{a_i \in A_i} \exp(z_i^t(a_i)) \right)^2}$$

$$= \pi_i^t(a_i) \frac{d}{dt} z_i^t(a_i) - \pi_i^t(a_i) \sum_{a_i \in A_i} \pi_i^t(a_i) \frac{d}{dt} z_i^t(a_i).$$

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

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

By combining these equalities, we get:

$$\frac{d}{dt} \pi_i^t(a_i) = \pi_i^t(a_i) \left( q_i^t(a_i) + \frac{\mu}{\pi_i^t(a_i)} (c_i(a_i) - \pi_i^t(a_i)) - \sum_{a_i \in A_i} \pi_i^t(a_i) \left( q_i^t(a_i) + \frac{\mu}{\pi_i^t(a_i)} (c_i(a_i) - \pi_i^t(a_i)) \right) \right)$$

$$= \pi_i^t(a) \left( q_i^t(a_i) - v_i^t \right) + \mu (c_i(a_i) - \pi_i^t(a_i)) - \mu \pi_i^t(a) \sum_{a_i \in A_i} (c_i(a_i) - \pi_i^t(a_i))$$

$$= \pi_i^t(a) \left( q_i^t(a_i) - v_i^t \right) + \mu (c_i(a_i) - \pi_i^t(a_i)).$$
D.2 PROOF OF LEMMA 5.5

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

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

$$
= \sum_{i=1}^2 \frac{d}{dt} \left( \psi_v^*(z_i) - \langle z_i^t, \pi_i \rangle \right)
$$

$$
= \sum_{i=1}^2 \left( \left\langle \frac{d}{dt} z_i^t, \nabla \psi_v^*(z_i) \right\rangle - \left\langle \frac{d}{dt} z_i^t, \pi_i \right\rangle \right)
$$

$$
= \sum_{i=1}^2 \left( \left\langle \frac{d}{dt} z_i^t, \nabla \psi_v^*(z_i) \right\rangle - \psi_v^*(z_i) \right).
$$

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

$$
\frac{d}{dt} D_\psi(\pi, \pi^t) = \sum_{i=1}^2 \left( \left\langle \frac{d}{dt} z_i^t, \pi_i^t - \pi_i \right\rangle \right)
$$

$$
= \sum_{i=1}^2 \sum_{a_i \in A_i} \left( q_i^{\pi^t}(a_i) + \frac{\mu}{\pi_i^{\pi^t}(a_i)} (c_i(a_i) - \pi_i^t(a_i)) \right) (\pi_i^t(a_i) - \pi_i(a_i))
$$

$$
= \sum_{i=1}^2 \sum_{a_i \in A_i} \left( \pi_i^t(a_i) - \pi_i(a_i) \right) \left( q_i^{\pi^t}(a_i) + \mu \frac{c_i(a_i)}{\pi_i^{\pi^t}(a_i)} \right)
$$

$$
= \sum_{i=1}^2 \left( v_i^{\pi^t} - v_i^{\pi^t, \pi_i} + \mu \sum_{a_i \in A_i} \pi_i^t(a_i) - \pi_i(a_i) \right) c_i(a_i) \pi_i^{\pi^t}(a_i)
$$

$$
= -2 \mu v_i^{\pi^t, \pi_i} + 2 \mu \sum_{i=1}^2 \sum_{a_i \in A_i} c_i(a_i) \pi_i^t(a_i) \pi_i^{\pi^t}(a_i)
$$

$$
= 2 \mu v_i^{\pi^t, \pi_i} + 2 \mu \sum_{i=1}^2 \sum_{a_i \in A_i} c_i(a_i) \pi_i^{\pi^t}(a_i),
$$

where the sixth equality follows from $\sum_{i=1}^2 v_i^{\pi^t} = 0$ and $\mu \sum_{a \in A} \pi_i^t(a) c_i(a) \pi_i^{\pi^t}(a) = \mu \sum_{a \in A} c_i(a) = \mu$, and the last equality follows from $v_1^{\pi^t, \pi_2} = -v_2^{\pi^t, \pi_2}$ and $v_2^{\pi^t, \pi_1} = -v_1^{\pi^t, \pi_1}$ 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 $a_i \in A_i$:

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

Then, we get:

$$
q_i^{\pi^t}(a_i) = v_i^{\pi^t} - \frac{\mu}{\pi_i^{\pi^t}(a_i)} (c_i(a_i) - \pi_i^t(a_i)).
$$
where the third equality follows from Lemma C.2. Then, for any \( \pi' \in \Delta(A_i) \) we have:

\[
v_i^{\pi_i', \pi''} = \sum_{a_i \in A_i} \pi_i'(a_i) q_i^{\pi''}(a_i)
= v_i^{\pi''} - \mu \sum_{a_i \in A_i} \frac{\pi_i'(a_i)}{\pi_i''(a_i)} (c_i(a_i) - \pi_i''(a_i))
= v_i^{\pi''} + \mu \sum_{a_i \in A_i} c_i(a_i) \frac{\pi_i'(a_i)}{\pi_i''(a_i)}.
\]

\[\square\]

### D.4 PROOF OF THEOREM 5.2

**Proof of Theorem 5.2** First, we prove the first part of the theorem. By setting \( \pi = \pi'' \) in Lemma 5.5 and \( \pi' = \pi' \) in Lemma 5.6 we have:

\[
\frac{d}{dt} D_\psi(\pi''', \pi') = \sum_{i=1}^{2} \frac{\partial}{\partial \pi_i}(v_i^{\pi_i', \pi''} + 2\mu - \mu \sum_{a_i \in A_i} c_i(a_i) \frac{\pi_i'(a_i)}{\pi_i''(a_i)})
= \sum_{i=1}^{2} v_i^{\pi''} + 4\mu - \mu \sum_{a_i \in A_i} c_i(a_i) \left( \frac{\pi_i'(a_i)}{\pi_i''(a_i)} + \frac{\pi_i''(a_i)}{\pi_i'(a_i)} \right)
= 4\mu - \mu \sum_{a_i \in A_i} c_i(a_i) \left( \frac{\pi_i'(a_i)}{\pi_i''(a_i)} + \frac{\pi_i''(a_i)}{\pi_i'(a_i)} \right)
= - \mu \sum_{a_i \in A_i} c_i(a_i) \left( \frac{\pi_i'(a_i)}{\pi_i''(a_i)} - \frac{\pi_i''(a_i)}{\pi_i'(a_i)} \right)^2,
\]

where the third equality follows from \( \sum_{i=1}^{2} v_i^{\pi''} = 0 \) 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:

\[
\frac{d}{dt} D_\psi(\pi'', \pi') = - \mu \sum_{a_i \in A_i} c_i(a_i) \left( \frac{\pi_i'(a_i)}{\pi_i''(a_i)} + \frac{\pi_i''(a_i)}{\pi_i'(a_i)} - 2 \right)
\leq - \mu \sum_{a_i \in A_i} c_i(a_i) \left( \min_{\pi_i''(a_i)} \pi_i'(a_i) \sum_{a_i \in A_i} \pi_i''(a_i) \left( \frac{\pi_i'(a_i)}{\pi_i''(a_i)} + \frac{\pi_i''(a_i)}{\pi_i'(a_i)} - 2 \right) \right)
= - \mu \sum_{a_i \in A_i} c_i(a_i) \left( \min_{\pi_i''(a_i)} \pi_i'(a_i) \sum_{a_i \in A_i} \pi_i''(a_i) \left( \pi_i'(a_i) - \pi_i''(a_i) \right)^2 \pi_i''(a_i) \right)
\leq - \mu \sum_{a_i \in A_i} c_i(a_i) \left( \min_{\pi_i''(a_i)} \pi_i'(a_i) \sum_{a_i \in A_i} \pi_i''(a_i) \left( \pi_i'(a_i) - \pi_i''(a_i) \right)^2 \pi_i''(a_i) \right)
= - \mu \sum_{a_i \in A_i} c_i(a_i) \left( \min_{\pi_i''(a_i)} \pi_i'(a_i) \sum_{a_i \in A_i} \pi_i''(a_i) \pi_i''(a_i) \pi_i'(a_i) \right)
\leq - \mu \sum_{a_i \in A_i} c_i(a_i) \left( \min_{\pi_i''(a_i)} \pi_i'(a_i) \sum_{a_i \in A_i} \pi_i''(a_i) \pi_i''(a_i) \pi_i'(a_i) \right)
= - \mu \left( \min_{i \in \{1,2\}, a_i \in A_i} \frac{c_i(a_i)}{\pi_i''(a_i)} \right) \sum_{i=1}^{2} KL(\pi_i', \pi_i^*),
\]

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

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

\[ \square \]

### E PROOFS OF ADDITIONAL LEMMAS

#### E.1 PROOF OF LEMMA C.1

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

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

(2)

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

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

Therefore, we have:

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

(3)

By combining (2) and (3):

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

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

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

\[ \square \]

#### 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:

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

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

\[ \square \]

References
