Supplemental Materials for *Dynamic Interruption Policies for Reinforcement Learning Game Playing Using Multi-Sampling Multi-Armed Bandits*

A. Appendix

A.1. Proofs of the following two equations introduced in Section 2.2.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\lambda_{[1:\infty]}}[C_{I_t,t}] &= \mathop{\mathbb{E}}_{\lambda}\Big[\sum_J \Big(\prod_{j=1}^{J-1}\epsilon_j\Big)\cdot c_{I_t,t,J}\Big] \\
&= \mathop{\mathbb{E}}_{c}\Big[\sum_J \Big(\prod_{j=1}^{J-1}(1-c_{I_t,t,j})\Big)c_{I_t,t,J}\Big] \\
&= 1,
\end{aligned} \tag{1}$$

$$\begin{aligned}
\mathop{\mathbb{E}}_{\lambda,I_t}\Big[\frac{R_{I_t,t}\mathbb{1}_{I_t=i}}{p_{i,t}}\Big] &= \sum_{I_t=1}^{K} p_{I_t,t}\mathbb{E}\Big[\frac{R_{I_t,t}\mathbb{1}_{I_t=i}}{p_{i,t}}\Big] = \mathbb{E}[R_{i,t}] \\
&= \mathop{\mathbb{E}}_{\lambda}\Big[\sum_J \Big(\prod_{j=1}^{J-1}\epsilon_j\Big)\cdot\Big(r_{i,t,J}\Big)\Big] \\
&= \mathop{\mathbb{E}}_{r,c}\Big[\sum_J \Big(\prod_{j=1}^{J-1}(1-c_{i,t,j})\Big)r_{i,t,J}\Big] \\
&= \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]},
\end{aligned} \tag{2}$$

A.2. Proof of Theorem 1.

We first proof a lemma described as follows,

Lemma 1: Let

$$\beta = \sqrt{\frac{\ln(K)}{nK}}, \tilde{R}_{i,t} = \frac{R_{I_t,t}\mathbb{1}_{I_t=i} + 3\beta}{p_{i,t}}, n \geq 100,$$

we can obtain the following inequality,

$$\mathop{\mathbb{E}}_{\lambda,I_t}\Big[\max_i \Big(\beta\sum_{t=1}^{n}\frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]}C_{I_t,t} - \beta\sum_{t=1}^{n}\tilde{R}_{i,t}\Big)\Big] \leq \ln(K) + 3. \tag{3}$$

Proof of Lemma 1: Let $\beta = 0.33\sqrt{\frac{\ln(K)}{nK}}$, $B = 2\ln(n+1)$. It is easy to verify that,

$$\mathbb{1}_{C_{I_t,t}>B} \leq \frac{\exp(\frac{C_{I_t,t}}{2}) - 1}{\exp(\frac{B}{2}) - 1} = \frac{\exp(\frac{C_{I_t,t}}{2}) - 1}{n}. \tag{4}$$

Using (4) and the inequality $\exp(x) \leq 1 + x + x^2$ for $x \leq 1$, we have,

$$\begin{aligned}
\mathbb{P}(C_{I_t,t} > B) &= \sum_J \left(\prod_{j=1}^{J-1} (1 - c_{I_t,t,j}) \right) c_{I_t,t,J} \cdot \mathbb{1}_{C_{I_t,t} > B} \\
&\leq \frac{1}{n} \sum_J \left(\prod_{j=1}^{J-1} (1 - c_{I_t,t,j}) \right) c_{I_t,t,J} \cdot \left(\exp\Big(\sum_{j=1}^{J} \frac{c_{I_t,t,j}}{2} \Big) - 1 \right) \\
&= \frac{1}{n} \left(\frac{\mathbb{E}[c_{I_t,t} \exp(\frac{c_{I_t,t}}{2})]}{1 - \mathbb{E}[(1 - c_{I_t,t}) \exp(\frac{c_{I_t,t}}{2})]} - 1 \right) \\
&\leq \frac{1}{n} \left(\frac{\mathbb{E}[c_{I_t,t} + \frac{c_{I_t,t}^2}{2} + \frac{c_{I_t,t}^3}{4}]}{\mathbb{E}[\frac{c_{I_t,t}}{2} + \frac{c_{I_t,t}^2}{4} + \frac{c_{I_t,t}^3}{4}]} - 1 \right) \\
&< \frac{1}{n},
\end{aligned}$$

where the last inequality comes from the fact $\frac{\mathbb{E}[c_{I_t,t} + \frac{c_{I_t,t}^2}{2} + \frac{c_{I_t,t}^3}{4}]}{\mathbb{E}[\frac{c_{I_t,t}}{2} + \frac{c_{I_t,t}^2}{4} + \frac{c_{I_t,t}^3}{4}]} < 2$.

Moreover,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\lambda} \left[\frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} C_{I_t,t} \right]^2 &= \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} \mathop{\mathbb{E}}_{c} \left[\sum_J \left(\prod_{j=1}^{J-1} (1 - c_{I_t,t,j}) \right) c_{I_t,t,J} \cdot \Big(\sum_{j=1}^{J} c_{I_t,t,j} \Big)^2 \right] \\
&= \Big(\frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} \Big)^2 \times \Big(2 - \frac{\mathbb{E}[c_{I_t,t}^2]}{\mathbb{E}[c_{I_t,t}]} \Big).
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\lambda, I_t} \left[\frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} C_{I_t,t} \cdot \frac{R_{I_t,t} \mathbb{1}_{I_t = i}}{p_{i,t}} \right] = \Big(\frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} \Big)^2 \times \Big(2 - \frac{\mathbb{E}[c_{i,t}^2]}{\mathbb{E}[c_{i,t}]} \Big), \\
&\mathop{\mathbb{E}}_{\lambda, I_t} \left[\frac{R_{I_t,t} \mathbb{1}_{I_t = i}}{p_{i,t}} \right]^2 = \frac{1}{p_{i,t}} \left(2 \Big(\frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} \Big)^2 + \frac{\mathbb{E}[r_{i,t}]^2}{\mathbb{E}[c_{i,t}]} - 2 \Big(\frac{\mathbb{E}[r_{i,t}] \mathbb{E}[r_{i,t} c_{i,t}]}{\mathbb{E}[c_{i,t}]^2} \Big) \right).
\end{aligned}$$

And thus,

$$\mathop{\mathbb{E}}_{\lambda, I_t} \left[\beta \frac{E[r_{i,t}]}{E[c_{i,t}]} C_{I_t,t} - \beta \frac{R_{I_t,t} \mathbb{1}_{I_t = i}}{p_{i,t}} \right]^2 \leq \frac{3\beta^2}{p_{i,t}}.$$

When $C_{I_t,t} \leq B$, we have $\beta \frac{E[r_{i,t}]}{E[c_{i,t}]} C_{I_t,t} \leq 1$. Then,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\lambda,I_t} \exp\Big(\beta \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} C_{I_t,t} - \beta \tilde{R}_{I_t,t}\Big) \\
&\quad = \mathbb{P}(C_{I_t,t} \leq B) \cdot \mathop{\mathbb{E}}_{\lambda,I_t} \Big[\exp\Big(\beta \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} C_{I_t,t} - \beta \tilde{R}_{I_t,t}\Big) \Big| C_{I_t,t} \leq B \Big] \\
&\quad + \mathbb{P}(C_{I_t,t} > B) \cdot \mathop{\mathbb{E}}_{\lambda,I_t} \Big[\exp\Big(\beta \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} C_{I_t,t} - \beta \tilde{R}_{I_t,t}\Big) \Big| C_{I_t,t} > B \Big] \\
&\quad \leq \mathbb{P}(C_{I_t,t} \leq B) \cdot \mathop{\mathbb{E}}_{\lambda,I_t} \Big[1 + \mathop{\mathbb{E}}_{\lambda,I_t} \big[\beta \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} C_{I_t,t} - \beta \frac{R_{I_t,t} \mathbb{1}_{I_t=i}}{p_{i,t}}\big] + \mathop{\mathbb{E}}_{\lambda,I_t} \big[\beta \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} C_{I_t,t} - \beta \frac{R_{I_t,t} \mathbb{1}_{I_t=i}}{p_{i,t}}\big]^2 \Big| C_{I_t,t} \leq B \Big] \\
&\quad \times \exp\Big(-\frac{3\beta^2}{p_{i,t}}\Big) + \mathbb{P}(C_{I_t,t} > B) \cdot \Big[\exp\Big(\beta \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} (B+1)\Big) \Big| C_{I_t,t} > B \Big] \\
&\quad \leq \Big(1 + \mathop{\mathbb{E}}_{\lambda,I_t} \Big[\beta \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} C_{I_t,t} - \beta \frac{R_{I_t,t} \mathbb{1}_{I_t=i}}{p_{i,t}}\Big] + \mathop{\mathbb{E}}_{\lambda,I_t} \Big[\beta \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} C_{I_t,t} - \beta \frac{R_{I_t,t} \mathbb{1}_{I_t=i}}{p_{i,t}}\Big]^2 \Big) \\
&\quad \times \exp\Big(-\frac{3\beta^2}{p_{i,t}}\Big) + \frac{\exp(\beta(B+1))}{n} \\
&\quad \leq \Big(1 + \frac{3\beta^2}{p_{i,t}}\Big) \times \exp\Big(-\frac{3\beta^2}{p_{i,t}}\Big) + \frac{2}{n} \\
&\quad \leq 1 + \frac{2}{n}.
\end{aligned}$$

Thus, we have,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\lambda,I_t} \Big[\exp\Big(\beta \sum_{t=1}^{n} \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} C_{I_t,t} - \beta \sum_{t=1}^{n} \tilde{R}_{I_t,t}\Big) \Big] \\
&\leq (1 + \frac{2}{n})^n \\
&\leq e^2.
\end{aligned}$$

Moreover, Markov's inequality implies $\mathbb{P}(X > \ln(\delta^{-1})) \leq \delta \mathbb{E} e^X$ and thus, with probability at least $1 - \delta e^2$,

$$\beta \sum_{t=1}^{n} \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} C_{I_t,t} - \beta \sum_{t=1}^{n} \tilde{R}_{I_t,t} \leq \ln(\delta^{-1}).$$

As a consequence, with probability at most $K\delta e^2$,

$$\max_i \Big(\beta \sum_{t=1}^{n} \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} C_{I_t,t} - \beta \sum_{t=1}^{n} \tilde{R}_{I_t,t}\Big) > \ln(\delta^{-1}).$$

This equals to say, with probability at most δ,

$$\max_i \Big(\beta \sum_{t=1}^{n} \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} C_{I_t,t} - \beta \sum_{t=1}^{n} \tilde{R}_{I_t,t}\Big) - \ln K - 2 > \ln(\delta^{-1}).$$

Using following equation:

$$\mathbb{E}[W] \leq \int_0^1 \frac{1}{\delta} \mathbb{P}(W > \ln(\frac{1}{\delta})) d\delta.$$

In particular, taking

$$W = \max_i \Big(\beta \sum_{t=1}^{n} \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} C_{I_t,t} - \beta \sum_{t=1}^{n} \tilde{R}_{I_t,t}\Big) - \ln K - 2,$$

yields $\mathbb{E}[W] \leq 1$, which is equivalent to inequality (3).

Proof of Theorem 1: One can immediately see that

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}(n)] &= \mathop{\mathbb{E}}_{\lambda,I_t} \max_i \Big[\gamma \sum_{t=1}^{n} \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} C_{I_t,t} + (1-\gamma)\Big(\sum_{t=1}^{n} \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} C_{I_t,t} - \sum_{t=1}^{n} \tilde{R}_{i,t} + \sum_{t=1}^{n} \tilde{R}_{i,t}\Big) - \sum_{t=1}^{n} R_{I_t,t}\Big] \\
&\leq \gamma \mathop{\mathbb{E}}_{\lambda,I_t} \sum_{t=1}^{n} C_{I_t,t} + (1-\gamma) \max_i \Big(\sum_{t=1}^{n} \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]} C_{I_t,t} - \sum_{t=1}^{n} \tilde{R}_{i,t}\Big) + \mathop{\mathbb{E}}_{\lambda,I_t} \max_i \Big[(1-\gamma)\sum_{t=1}^{n} \tilde{R}_{i,t} - \sum_{t=1}^{n} R_{I_t,t}\Big] \\
&\leq \gamma n + \frac{(1-\gamma)(3+\ln K)}{\beta} + \mathop{\mathbb{E}}_{\lambda,I_t} \max_i \Big[(1-\gamma)\sum_{t=1}^{n} \tilde{R}_{i,t} - \sum_{t=1}^{n} R_{I_t,t}\Big].
\end{aligned}$$

Where the last inequality comes from Lemma 1.

Let $u = (\frac{1}{K}, \ldots, \frac{1}{K})$ be the uniform distribution over the arms, $w_t = \frac{p_t - u\gamma}{1-\gamma}$ be the distribution induced by Exp3.P.MS at time t without the mixing. Since,

$$w_{i,1} = 0, \quad w_{i,t} = \frac{\exp(\eta \sum_{s=1}^{t-1} \tilde{R}_{i,s})}{\sum_{k=1}^{K} \exp(\eta \sum_{s=1}^{t-1} \tilde{R}_{k,s})},$$

we have:

$$\begin{aligned}
\sum_{t=1}^{n} \ln \mathop{\mathbb{E}}_{i\sim w_t} \exp(\eta \tilde{R}_{i,t})) &= \sum_{t=1}^{n} \ln \sum_{i=1}^{k} \frac{\exp(\eta \sum_{\tau=1}^{t-1} \tilde{R}_{i,\tau}) \exp(\eta \tilde{R}_{i,t})}{\sum_{k=1}^{K} \exp(\eta \sum_{\tau=1}^{t-1} \tilde{R}_{k,\tau})} \\
&= \ln \prod_{t=1}^{n} \frac{\sum_{k=1}^{K} \exp(\eta \sum_{\tau=1}^{t} \tilde{R}_{k,\tau})}{\sum_{k=1}^{K} \exp(\eta \sum_{\tau=1}^{t-1} \tilde{R}_{k,\tau})} \\
&= \ln(\sum_{k=1}^{K} \exp(\eta \sum_{t=1}^{n} \tilde{R}_{k,t})) - \ln(K) \qquad (5) \\
&\geq \max_k \ln(\exp(\eta \sum_{t=1}^{n} \tilde{R}_{k,t})) - \ln(K) \\
&= \max_k \eta \sum_{t=1}^{n} \tilde{R}_{k,t} - \ln(K).
\end{aligned}$$

Additionally, it is easily to verify that,

$$-\sum_{t=1}^{n} R_{I_t,t} = -\sum_{t=1}^{n} \mathop{\mathbb{E}}_{i\sim p_t} \tilde{R}_{i,t} + 3\beta n K.$$

Then,

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}(n)] &\leq \gamma n + \frac{(1-\gamma)(3+\ln K)}{\beta} + \mathop{\mathbb{E}}_{\lambda,I_t} \max_i \Big[(1-\gamma)\sum_{t=1}^{n} \tilde{R}_{i,t} - \sum_{t=1}^{n} R_{I_t,t}\Big] \\
&\leq \gamma n + \frac{(1-\gamma)(3+\ln K)}{\beta} + \mathop{\mathbb{E}}_{\lambda,I_t} \Big[\frac{1-\gamma}{\eta} \sum_{t=1}^{n} \ln \mathop{\mathbb{E}}_{i\sim w_t} \exp\Big(\eta \tilde{R}_{i,t}\Big) \\
&\quad - \sum_{t=1}^{n} \mathop{\mathbb{E}}_{i\sim p_t} \tilde{R}_{I_t,t} + 3\beta n K\Big] + \frac{(1-\gamma)\ln K}{\eta} \\
&= \mathop{\mathbb{E}}_{\lambda,I_t} \Big[(1-\gamma)\Big(\frac{1}{\eta} \ln \mathop{\mathbb{E}}_{i\sim w_t} \exp(\eta \tilde{R}_{i,t}) - \mathop{\mathbb{E}}_{k\sim p_t} \eta \tilde{R}_{k,t} - \gamma \mathop{\mathbb{E}}_{i\sim \mu} \tilde{R}_{i,t}\Big] \\
&\quad + \gamma n + \frac{(1-\gamma)(3+\ln K)}{\beta} + 3\beta n K + \frac{(1-\gamma)\ln K}{\eta} \\
&< \mathop{\mathbb{E}}_{\lambda,I_t} \Big[\frac{1-\gamma}{\eta} (\ln \mathop{\mathbb{E}}_{i\sim w_t} \exp(\eta \tilde{R}_{i,t}) - \mathop{\mathbb{E}}_{k\sim p_t} \eta \tilde{R}_{k,t})\Big] \\
&\quad + \gamma n + \frac{(1-\gamma)(3+\ln K)}{\beta} + 3\beta n K + \frac{(1-\gamma)\ln K}{\eta}.
\end{aligned}$$

Using the inequalities $\ln x \leq x - 1$ and $\exp(x) \leq 1 + x + x^2$, for all $x \leq 1$, as well as the fact that $(1+3\beta)\eta K \leq \gamma$:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\lambda}\Big[\ln \mathop{\mathbb{E}}_{i\sim w_t} \exp(\eta\tilde{R}_{i,t}) - \mathop{\mathbb{E}}_{k\sim p_t} \eta\tilde{R}_{k,t}\Big] &\leq \mathop{\mathbb{E}}_{\lambda} \mathop{\mathbb{E}}_{i\sim w_t} [\exp(\eta\tilde{R}_{i,t}) - 1 - \eta\tilde{R}_{i,t}] \\
&= \mathop{\mathbb{E}}_{\lambda} \mathop{\mathbb{E}}_{i\sim w_t} \Big[\exp(\frac{\eta R_{i,t}\mathbb{1}_{I_t=i} + 3\eta\beta}{p_{i,t}}) - 1 - \eta\tilde{R}_{i,t}\Big] \\
&= w_{I_t,t}\frac{\mathbb{E}[c_{I_t,t}\exp(\frac{\eta r_{I_t,t}+3\eta\beta}{p_{I_t,t}})]}{1-\mathbb{E}[(1-c_{I_t,t})\exp(\frac{\eta r_{I_t,t}}{p_{I_t,t}})]} + \sum_{i\neq I_t} w_{i,t}\exp\Big(\frac{3\eta\beta}{p_{i,t}}\Big) \\
&\quad - 1 - \frac{w_{I_t,t}\eta\mathbb{E}[r_{It,t}]}{p_{I_t,t}\mathbb{E}[c_{It,t}]} - \mathop{\mathbb{E}}_{i\sim w_t}\frac{3\eta\beta}{p_{i,t}} \\
&= w_{I_t,t}\Big(\frac{\mathbb{E}[c_{I_t,t}\exp(\frac{\eta r_{I_t,t}+3\eta\beta}{p_{I_t,t}})]}{1-\mathbb{E}[(1-c_{I_t,t})\exp(\frac{\eta r_{I_t,t}}{p_{I_t,t}})]} - \frac{\eta\mathbb{E}[r_{I_t,t}]}{p_{I_t,t}\mathbb{E}[c_{It,t}]} \\
&\quad - \frac{3\eta\beta}{p_{I_t,t}} - 1\Big) + \sum_{i\neq I_t} w_{i,t}\Big(\exp\Big(\frac{3\eta\beta}{p_{i,t}}\Big) - \frac{3\eta\beta}{p_{i,t}} - 1\Big) \\
&\triangleq w_{I_t,t}A_1 + \sum_{i\neq I_t} w_{i,t}A_2.
\end{aligned}
\tag{6}
$$

For convenience, we omit the index I_t and t without ambiguity, and denote $x = \frac{\eta r_{I_t,t}}{p_{I_t,t}}, b = \frac{3\eta\beta}{p_{I_t,t}}, c = c_{I_t,t}$. We rewrite the term A_1 to:

$$
\begin{aligned}
A_1 &= \frac{\mathbb{E}[ce^{x+b}]}{1-\mathbb{E}[(1-c)e^x]} - \frac{\mathbb{E}[x]}{\mathbb{E}[c]} - b - 1 \\
&= \frac{\mathbb{E}[c]\mathbb{E}[ce^{x+b}] + \mathbb{E}[(1-c)e^x](\mathbb{E}[x] + b\mathbb{E}[c] + \mathbb{E}[c])}{\mathbb{E}[c](1-\mathbb{E}[(1-c)e^x])} - \frac{(\mathbb{E}[x] + b\mathbb{E}[c] + \mathbb{E}[c])}{\mathbb{E}[c](1-\mathbb{E}[(1-c)e^x])} \\
&\leq \frac{\mathbb{E}[c]\mathbb{E}[c((x+b)^2 + (x+b) + 1)] - (\mathbb{E}[x] + b\mathbb{E}[c] + \mathbb{E}[c])}{\mathbb{E}[c](1-\mathbb{E}[(1-c)(x^2+x+1)])} + \frac{\mathbb{E}[(1-c)(x^2+x+1)](\mathbb{E}[x] + b\mathbb{E}[c] + \mathbb{E}[c])}{\mathbb{E}[c](1-\mathbb{E}[(1-c)(x^2+x+1)])} \\
&= \frac{(\mathbb{E}[x] + b\mathbb{E}[c])\mathbb{E}[x^2 - x^2c + x + bc] + b\mathbb{E}[xc]\mathbb{E}[c]}{\mathbb{E}[c](1-\mathbb{E}[(1-c)(x^2+x+1)])} - \frac{b\mathbb{E}[x]\mathbb{E}[c] + \mathbb{E}[c]\mathbb{E}[x^2] - \mathbb{E}[x]\mathbb{E}[xc]}{\mathbb{E}[c](1-\mathbb{E}[(1-c)(x^2+x+1)])} \\
&\leq \frac{(\mathbb{E}[x] + b\mathbb{E}[c])\mathbb{E}[x^2 - x^2c + x + bc] + \mathbb{E}[c]\mathbb{E}[x^2]}{\mathbb{E}[c](1-\mathbb{E}[(1-c)(x^2+x+1)])} \\
&\leq \frac{(\mathbb{E}[x] + b\mathbb{E}[c])\mathbb{E}[x^2 - x^2c + x + bc + \mathbb{E}[c]\frac{\eta}{p_{I_t,t}}]}{\mathbb{E}[c](1-\mathbb{E}[(1-c)(x^2+x+1)])} \\
&\leq \frac{(\mathbb{E}[x] + b\mathbb{E}[c])\mathbb{E}[(1-c)cx\frac{\eta}{p_{I_t,t}} + x + bc + \mathbb{E}[c]\frac{\eta}{p_{I_t,t}}]}{\mathbb{E}[c](1-\mathbb{E}[(1-c)(x^2+x+1)])}.
\end{aligned}
$$

Since $(1-c)c \leq 0.25$, we have,

$$
\begin{aligned}
A_1 &\leq \frac{\eta^2}{p_{I_t,t}}\frac{(2.25+3\beta)\mathbb{E}[c_{I_t,t}]}{1-\mathbb{E}[(1-c_{I_t,t})(1+\frac{\eta}{p_{I_t,t}}r + (\frac{\eta}{p_{I_t,t}}r)^2)]} \times \Big(\frac{\mathbb{E}[r_{I_t,t}]}{\mathbb{E}[c_{I_t,t}]p_{I_t,t}} + \frac{3\beta}{p_{I_t,t}}\Big) \\
&\leq \frac{\eta^2}{p_{I_t,t}}\frac{(2.25+3\beta)\mathbb{E}[c_{I_t,t}]}{(1-\frac{\eta}{p_{I_t,t}})\mathbb{E}[c_{I_t,t}]}\Big(\frac{\mathbb{E}[r_{I_t,t}]}{\mathbb{E}[c_{I_t,t}]p_{I_t,t}} + \frac{3\beta}{p_{I_t,t}}\Big) \\
&= \frac{\eta^2}{p_{I_t,t}}\frac{(2.25+3\beta)p_{I_t,t}}{p_{I_t,t}-\eta}\Big(\frac{\mathbb{E}[r_{I_t,t}]}{\mathbb{E}[c_{I_t,t}]p_{I_t,t}} + \frac{3\beta}{p_{I_t,t}}\Big) \\
&= \frac{\eta^2}{p_{I_t,t}}\frac{(2.25+3\beta)\gamma/K}{\gamma/K-\eta}\Big(\frac{\mathbb{E}[r_{I_t,t}]}{\mathbb{E}[c_{I_t,t}]p_{I_t,t}} + \frac{3\beta}{p_{I_t,t}}\Big).
\end{aligned}
\tag{7}
$$

At the same time,

$$
\begin{aligned}
A_2 &= \exp\Big(\frac{3\eta\beta}{p_{i,t}}\Big) - \frac{3\eta\beta}{p_{i,t}} - 1 \\
&\le \Big(\frac{3\eta\beta}{p_{i,t}}\Big)^2 + \frac{3\eta\beta}{p_{i,t}} + 1 - \frac{3\eta\beta}{p_{i,t}} - 1 \\
&\le \frac{\eta^2}{p_{i,t}} \frac{(2.25+\beta)\gamma/K}{\gamma/K-\eta}\Big(\frac{3\beta}{p_{i,t}}\Big).
\end{aligned} \tag{8}
$$

Combine (7) and (8), we have

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\lambda}\Big[\ln \mathop{\mathbb{E}}_{i\sim w_t} \exp(\eta\tilde{R}_{i,t}) - \mathop{\mathbb{E}}_{i\sim w_t} \eta\tilde{R}_{i,t}\Big] &\le \sum_{i=1}^{K} \frac{w_i}{p_i}\eta^2 \frac{(2.25+3\beta)\gamma/K}{\gamma/K-\eta}\Big(\frac{\mathbb{1}_{i=I_t}+3\beta}{p_{i,t}}\Big) \\
&\le \frac{\eta^2}{1-\gamma}\frac{(2.25+3\beta)\gamma/K}{\gamma/K-\eta}\sum_{i=1}^{K}\frac{\mathbb{1}_{i=I_t}+3\beta}{p_{i,t}} \\
&= \frac{\eta^2}{1-\gamma}\frac{(2.25+3\beta)\gamma/K}{\gamma/K-\eta}\sum_{I_t=1}^{K} p_{I_t}\sum_{i=1}^{K}\frac{\mathbb{1}_{i=I_t}+3\beta}{p_{i,t}} \\
&= \frac{\eta^2}{1-\gamma}\frac{(2.25+3\beta)\gamma/K}{\gamma/K-\eta}\Big(K + \sum_{I_t=1}^{K}\sum_{i=1}^{K}\frac{3p_{I_t}\beta}{p_{i,t}}\Big) \\
&= \frac{\eta^2}{1-\gamma}\frac{(2.25+3\beta)\gamma/K}{\gamma/K-\eta}\Big(K + \sum_{i=1}^{K}\frac{\beta}{p_{i,t}}\Big) \\
&\le \frac{\eta^2}{1-\gamma}\frac{(2.25+3\beta)\gamma/K}{\gamma/K-\eta}\Big(K + \sum_{i=1}^{K}\frac{K\beta}{\gamma}\Big) \\
&\le \frac{\eta^2}{1-\gamma}\frac{(2.25+3\beta)\gamma/K}{\gamma/K-\eta}\Big(K + K\frac{K\beta}{\gamma}\Big).
\end{aligned}
$$

To sum up,

$$
\begin{aligned}
\mathbb{E}[\mathrm{Reg}(n)] &\le \gamma n + \frac{(1-\gamma)(3+\ln K)}{\beta} + \frac{(1-\gamma)\ln K}{\eta} + 3\beta nK + \sum_{t=1}^{n}\mathop{\mathbb{E}}_{\lambda}\Big[\frac{1-\gamma}{\eta}\Big(\ln \mathop{\mathbb{E}}_{i\sim w_t}\exp(\eta\tilde{R}_{i,t}) - \mathop{\mathbb{E}}_{i\sim w_t}\eta\tilde{R}_{i,t}\Big)\Big] \\
&\le \gamma n + \frac{(1-\gamma)(3+\ln K)}{\beta} + \frac{(1-\gamma)\ln K}{\eta} + 3\beta nK + n\frac{1-\gamma}{\eta}\frac{\eta^2}{1-\gamma}\frac{(2.25+3\beta)\gamma/K}{\gamma/K-\eta}\Big(K + K\frac{3K\beta}{\gamma}\Big) \\
&\le \gamma n + \frac{(1-\gamma)(3+\ln K)}{\beta} + \frac{(1-\gamma)\ln K}{\eta} + 3\beta nK + \eta nK\frac{(2.25+3\beta)\gamma/K}{\gamma/K-\eta}\Big(1 + \frac{3K\beta}{\gamma}\Big) \\
&\le 17.33\sqrt{nK\ln K}.
\end{aligned}
$$

A.3. The details of the modification of the Exp3.P.MS algorithm

Algorithm 1 Modification of Exp3.P.MS

Parameters: $\eta \in \mathbb{R}^+$ and $\gamma, \beta \in [0, 1]$.
Let p_1 be the uniform distribution over $1, \dots, K$.
For each round $t = 1, \dots, n$
(1) Draw an arm I_t from the probability distribution p_t, and let $C_{I_t,t} = 0, R_{i,t} = 0, a_{i,t} = 0$.
while $C_{I_t,t} < 2\ln(n+1)$ **do**
Pull arm I_t and record the reward $r_{I_t,t}$ and the cost $c_{I_t,t}$ ($c_{I_t,t} \le \tau_{I_t}$),

$$R_{I_t,t} \leftarrow R_{I_t,t} + r_{I_t,t}, \quad C_{I_t,t} \leftarrow C_{I_t,t} + c_{I_t,t}.$$

For each arm $i = 1, \dots, (I_t - 1)$,

$$R_{i,t} \leftarrow R_{i,t} + a_{i,t} \cdot r_{I_t,t} \mathbb{1}_{c_{I_t,t} \le \tau_i},$$
$$a_{i,t} \leftarrow a_{i,t} \max\left(0, \frac{(\tau_i - c_{I_t,t})\tau_{I_t}}{(\tau_{I_t} - c_{I_t,t})\tau_i}\right).$$

For each arm $i = (I_t + 1), \dots, K$,

$$R_{i,t} = 0.$$

Break the loop with probability $\frac{c_{I_t,t}}{\tau_{I_t}}$.
end while
(2) Compute the estimated gain for each arm:

$$\tilde{R}_{i,t} = \frac{1}{\tau_i}\Big(R_{1,t} + \sum_{j=2}^{\min(I_t,i)} \frac{(R_{j,t} - R_{j-1,t})}{\sum_{k=i}^K p_{k,t}}\Big) + \frac{3\beta}{\sum_{k=i}^K p_{k,t}},$$

and update the estimated cumulative gain for each arm:

$$q_{i,t} = \sum_{s=1}^{t} \tilde{R}_{i,s}. \tag{9}$$

(3) Compute the new probability distribution over the arms $p_{t+1} = (p_{1,t+1}, \dots, p_{K,t+1})$ where:

$$p_{i,t+1} = (1-\gamma)\frac{\exp(\eta q_{i,t})}{\sum_{i=1}^K \exp(\eta q_{i,t})} + \gamma \mathbb{1}_{i=K}$$

As at each round, some information of arms $1, \dots, I_t - 1$ can also be monitored, we assign a value instead of 0 for these arms. By this way, the variance of the estimate $\tilde{R}_{i,t}$ can be reduced. For the multi-sampling process at round t, after pulling an arm with a cost of $c_{I_t,t}$, the bandit agent will continue to play the current arm with a probability of $1 - \frac{c_{I_t,t}}{\tau_{I_t}}$. This means that if the last game level is not completed but interrupted, then the current arm will be abandoned, and a new round will start. We use $\lambda_j = (r_j, c_j, \epsilon_j)$ denoting a variable vector, where the index j is used to count the number of sampling, and we omit the notations of I_t and t without ambiguity. ϵ_j is a random variable depending on c_j, satisfying that $\mathbb{P}(\epsilon_j = 1) = 1 - \frac{c_j}{\tau_{I_t}}$, and $\mathbb{P}(\epsilon_j = 0) = \frac{c_j}{\tau_{I_t}}$. If we don't take the constraint $C_{I_t,t} < 2\ln(n+1)$ into account and let the sampling process naturally end, then we can obtain that,

$$\mathop{\mathbb{E}}_{\lambda_{[1:\infty]}}[C_{I_t,t}] = \tau_{I_t}. \tag{10}$$

Let

$$\hat{R}_{i,t} = \frac{1}{\tau_i}\left(R_{1,t} + \sum_{j=2}^{\min(I_t,i)} \frac{(R_{j,t} - R_{j-1,t})}{\sum_{k=i}^K p_{k,t}}\right).$$

Then we can verify that $\hat{R}_{i,t}$ is the unbiased estimator of $\frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]}$.

$$\mathop{\mathbb{E}}_{\lambda,I_t}[\hat{R}_{i,t}] = \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]}. \tag{11}$$

Proof: At round t, we pull arm I_t and receive $c_{I_t,t}$ and $r_{I_t,t}$. If instead, we pull arm i ($i < I_t$) and receive $c_{i,t}$ and $r_{i,t}$, then 1) if $c_{I_t,t} > \tau_i$, we have $c_{i,t} = \tau_i$, $r_{i,t} = 0$; 2) if $c_{I_t,t} \leq \tau_i$, we have $c_{i,t} = c_{I_t,t}$, $r_{i,t} = r_{I_t,t}$.

Thus,

$$\begin{aligned}
&\mathbb{E}[r_{I_t,t}|c_{I_t,t} \leq \tau_i] \cdot \mathbb{P}(c_{I_t,t} \leq \tau_i) \\
&\quad = \mathbb{E}[r_{i,t}|c_{I_t,t} \leq \tau_i] \cdot \mathbb{P}(c_{I_t,t} > \tau_i) + \mathbb{E}[r_{I_t,t}|c_{I_t,t} \leq \tau_i] \cdot \mathbb{P}(c_{I_t,t} > \tau_i) \\
&\quad = \mathbb{E}[r_i],
\end{aligned} \tag{12}$$

$$\begin{aligned}
&\mathbb{E}\Big[\frac{\tau_i - c_{I_t,t}}{\tau_i}\Big|c_{I_t,t} \leq \tau_i\Big]\mathbb{P}(c_{I_t,t} \leq \tau_i) \\
&\quad = \mathbb{E}\Big[\frac{\tau_i - c_{i,t}}{\tau_i}\Big|c_{I_t,t} \leq \tau_i\Big]\mathbb{P}(c_{I_t,t} \leq \tau_i) \\
&\quad = \mathbb{E}\Big[\frac{\tau_i - c_{i,t}}{\tau_i}\Big|c_{I_t,t} \leq \tau_i\Big]\mathbb{P}(c_{I_t,t} \leq \tau_i) \\
&\quad + \mathbb{E}\Big[\frac{\tau_i - c_{i,t}}{\tau_i}\Big|c_{I_t,t} > \tau_i\Big]\mathbb{P}(c_{I_t,t} > \tau_i) \\
&\quad = \mathbb{E}\Big[\frac{\tau_i - c_{i,t}}{\tau_i}\Big].
\end{aligned} \tag{13}$$

Based on (12) and (13),

$$\begin{aligned}
\mathbb{E}[R_{i,t}] &= \mathbb{E}(r_{I_t,t,1} \leq \tau_i|c_{I_t,t,1} \leq \tau_i)\mathbb{P}(c_{I_t,t,1} \leq \tau_i) + \sum_{J=2}^{\infty}\Big(\mathbb{E}[r_{I_t,t,J} \leq \tau_i|c_{I_t,t,J} \leq \tau_i]\mathbb{P}(c_{I_t,t,J} \leq \tau_i) \\
&\prod_{j=1}^{J-1}\mathbb{E}\Big[\frac{\tau_{I_t} - c_{I_t,t,j}}{\tau_{I_t}}\frac{(\tau_i - c_{I_t,t,j})\tau_{I_t}}{(\tau_{I_t} - c_{I_t,t,j})\tau_i}\Big|c_{I_t,t,j} \leq \tau_i]\mathbb{P}(c_{I_t,t,j} \leq \tau_i)\Big) \\
&= \mathbb{E}[r_{i,t,1}] + \sum_{J=2}^{\infty}\Big(\mathbb{E}[r_{i,t,J}]\prod_{j=1}^{J-1}\mathbb{E}\Big[\frac{\tau_i - c_{I_t,t,j}}{\tau_i}\Big|c_{I_t,t,j} \leq \tau_i\Big]\mathbb{P}(c_{I_t,t,j} \leq \tau_i)\Big) \\
&= \mathbb{E}[r_{i,t}] + \mathbb{E}[r_{i,t}]\sum_{J}\mathbb{E}\big[1 - \frac{c_{i,t}}{\tau_i}\big]^J \\
&= \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]}\tau_i.
\end{aligned} \tag{14}$$

And finally,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\lambda,I_t}[\hat{\mu}_{i,t}] &= \frac{1}{\tau_i}\mathop{\mathbb{E}}_{\lambda}\Big(\sum_{I_t=1}^{K}p_{I_t,t}R_{1,t} + \sum_{I_t=1}^{K}p_{I_t,t}\sum_{j=2}^{min(i,I_t)}\frac{(R_{j,t} - R_{j-1,t})}{\sum_{k=i}^{K}p_{k,t}}\Big) \\
&= \frac{1}{\tau_i}\Big(\frac{\mathbb{E}[r_{1,t}]}{\mathbb{E}[c_{1,t}]}\tau_1 + \sum_{j=2}^{i}\sum_{I_t=j}^{K}p_{I_t,t}\frac{(\frac{\mathbb{E}[r_{j,t}]}{\mathbb{E}[c_{j,t}]}\tau_j - \frac{\mathbb{E}[r_{j-1,t}]}{\mathbb{E}[c_{j-1,t}]}\tau_{j-1})}{\sum_{k=i}^{K}p_{k,t}}\Big) \\
&= \frac{\mathbb{E}[r_{i,t}]}{\mathbb{E}[c_{i,t}]}.
\end{aligned}$$