

Appendix Contents

A	Notations	14
B	Supporting Lemmas	14
C	Convergence of CATSO and PATSO in Non-stationary multi-armed bandits	18
D	Convergence of CATSO and PATSO in Monte-Carlo Tree Search	28
E	Distributional MCTS and Wasserstein Robust Optimization	32
F	Experimental Setup	??
G	Limitations	38

A NOTATIONS

Table 2: List of all notations for Non-stationary Multi-armed Bandits.

Notation	Type	Description
K	\mathbb{N}	Number of arms
$T_a(t)$	\mathbb{N}	Number of visitations at arm a after t timesteps
μ_a	\mathbb{R}	Mean value of arm a
a_\star	\mathcal{A}	Optimal action
μ_\star	\mathbb{R}	Mean value of the optimal arm (assumed unique)
$\hat{\mu}_n(p)$	\mathbb{R}	Power mean estimator with parameter $p \in [1, +\infty)$
$\hat{\mu}_{a,n}$	\mathbb{R}	Mean estimator of arm a after n visitations
F_a	Distribution	Cumulative distribution function of arm a
F_a^n	Distribution	Empirical CDF of arm a after n visitations
$\mathcal{K}_{\text{inf}}(F_a, \mu_\star)$	\mathbb{R}_+	KL divergence between F_a and optimal arm
R	\mathbb{R}_+	Maximum reward value

B SUPPORTING LEMMAS

We start with a result of the following lemma which plays an important role in the analysis of our MCTS algorithm.

Lemma 1. For $m \in [M]$, let $(\hat{V}_{m,n})_{n \geq 1}$ be a sequence of estimator satisfying $\text{plim}_{n \rightarrow \infty} \hat{V}_{m,n} = V_m$.

Assume that there exists a constant $L > 0$ such that $L = \sup_{n \geq 1} \{\hat{V}_{m,n}\}$. Let R_i be an iid sequence with mean μ and S_i be an iid sequence from a distribution $p = (p_1, \dots, p_M)$ supported on $\{1, \dots, M\}$. Introducing the random variables $N_m^n = \#\{i \leq n : S_i = s_m\}$, we define the sequence of estimator

$$\hat{Q}_n = \frac{1}{n} \sum_{i=1}^n R_i + \gamma \sum_{m=1}^M \frac{N_m^n}{n} \hat{V}_{m, N_m^n}.$$

Then there exists some constant c' (which depends on p_i ($i=1,2,\dots,M$), γ , μ) such that

$$\text{plim}_{n \rightarrow \infty} \hat{Q}_n = \mu + \sum_{m=1}^M p_m V_m.$$

Proof. Let $p = (p_1, p_2, \dots, p_M)$, $p \in \Delta^M$ where $\Delta^M = \{x \in \mathbb{R}^M : \sum_{i=1}^M x_i = 1, x_i \geq 0\}$ is the $(M-1)$ -dimensional simplex. Let us study a random vector $\hat{p}_n = (\frac{N_1^n}{n}, \frac{N_2^n}{n}, \dots, \frac{N_M^n}{n})$. Let us define $V = (V_1, V_2, \dots, V_M)$. Let $\hat{R}_n = \frac{1}{n} \sum_{i=1}^n R_i$, $\hat{V}_n = (\hat{V}_{1, N_1^n}, \hat{V}_{2, N_2^n}, \dots, \hat{V}_{M, N_M^n})$, $\sum_{i=1}^M N_i^n = n$, N_i^n is the number of times that population i was observed. We have $\hat{Q}_n = \hat{R}_n + \gamma \langle \hat{p}_n, \hat{V}_n \rangle$. Therefore,

$$\begin{aligned} \mathbb{P}\left(\hat{Q}_n - (\mu + \gamma \langle p, V \rangle) \geq \varepsilon\right) &\leq \mathbb{P}\left(\hat{R}_n - \mu \geq \frac{1}{2}\varepsilon\right) + \mathbb{P}\left(\gamma \langle \hat{p}_n, \hat{V}_n \rangle - \gamma \langle p, V \rangle \geq \frac{1}{2}\varepsilon\right) \\ &\leq \exp\{-2n \frac{\varepsilon^2}{4}\} + \underbrace{\mathbb{P}\left(\langle \hat{p}_n, \hat{V}_n \rangle - \langle p, V \rangle \geq \frac{1}{2\gamma}\varepsilon\right)}_A. \end{aligned}$$

Table 3: List of all notations for Monte-Carlo Tree Search.

Notation	Type	Description
γ	$\mathbb{R} \in [0, 1)$	Discount factor
N	\mathbb{N}	Number of atoms (for categorical distribution)
s_h	\mathcal{S}	State at depth h
$\widehat{V}_t(s)$	\mathbb{R}	Estimated value function at state s after t visitations
$\widetilde{V}(s)$	\mathbb{R}	True value function at state s
$T_s(t)$	\mathbb{N}	Number of visitations at state s after t timesteps
$T_{s,a}(t)$	\mathbb{N}	Number of visitations at (s, a) after t timesteps
$T_{s,a}^{s'}(t)$	\mathbb{N}	Number of transitions from (s, a) to s' after t timesteps
$\widehat{Q}_t(s, a)$	\mathbb{R}	Estimated Q-value at (s, a) after t visitations
$\widetilde{Q}(s, a)$	\mathbb{R}	True Q-value at (s, a)
$Q_{\min}(s, a)$	\mathbb{R}	Minimum support of Q-value distribution at (s, a)
$Q_{\max}(s, a)$	\mathbb{R}	Maximum support of Q-value distribution at (s, a)
$\mathcal{R}(s, a)$	Distribution	Reward distribution at (s, a)
$\mathcal{V}(s)$	Distribution	Value distribution at state s
$\mathcal{Q}(s, a)$	Distribution	Q-value distribution at (s, a)
$p_i(s, a)$	$[0, 1]$	Probability of the i -th atom in Q-distribution at (s, a)
Δz	\mathbb{R}_+	Distance between consecutive atoms
$z_i(s, a)$	\mathbb{R}	Value of the i -th atom at (s, a)
$\overline{Q}_t(s, a)$	\mathbb{R}	Intermediate Q-value at time t at (s, a)
H	\mathbb{N}	Maximum depth of the search tree

To upper bound A, let us consider $\langle \widehat{p}_n, \widehat{V} \rangle - \langle p, V \rangle = \langle (\widehat{p}_n - p), \widehat{V}_n \rangle + \langle p, (\widehat{V} - V) \rangle$. Then,

$$A \leq \underbrace{\mathbb{P}\left(\langle (\widehat{p}_n - p), \widehat{V}_n \rangle \geq \frac{1}{4\gamma}\varepsilon\right)}_{A_1} + \underbrace{\mathbb{P}\left(\langle p, (\widehat{V}_n - V) \rangle \geq \frac{1}{4\gamma}\varepsilon\right)}_{A_2}.$$

By applying a Hölder inequality to $\widehat{p}_n - p$ and \widehat{V} , we obtain

$$\langle (\widehat{p}_n - p), \widehat{V}_n \rangle \leq \|\widehat{p}_n - p\|_1 \|\widehat{V}_n\|_\infty = \|\widehat{p}_n - p\|_1 L,$$

with L is the supremum of \widehat{V} . Then we can derive

$$\begin{aligned} A_1 &= \mathbb{P}\left(\langle (\widehat{p}_n - p), \widehat{V}_n \rangle \geq \frac{1}{4\gamma}\varepsilon\right) \leq \mathbb{P}\left(\|\widehat{p}_n - p\|_1 L \geq \frac{1}{4\gamma}\varepsilon\right) \\ &= \mathbb{P}\left(\|\widehat{p}_n - p\|_1 \geq \frac{1}{4\gamma L}\varepsilon\right). \end{aligned}$$

According to Weissman et al. (2003), we have for any $M \geq 2$ and $\delta \in [0, 1]$

$$\mathbb{P}\left(\|\widehat{p}_n - p\|_1 \geq \sqrt{\frac{2M \ln(2/\delta)}{n}}\right) \leq \delta.$$

Define $\varepsilon = \sqrt{\frac{2M \ln(2/\delta)}{n}}$, therefore $\delta = 2 \exp\{-\frac{n\varepsilon^2}{2M}\}$, we have

$$\mathbb{P}\left(\|\hat{p}_n - p\|_1 \geq \varepsilon\right) \leq 2 \exp\left\{-\frac{n\varepsilon^2}{2M}\right\}.$$

Therefore,

$$A_1 \leq \mathbb{P}\left(\|\hat{p}_n - p\|_1 \geq \varepsilon\right) \leq 2 \exp\left\{-\frac{n\varepsilon^2}{32M\gamma^2 L^2}\right\}.$$

We also have

$$\begin{aligned} A_2 &= \mathbb{P}\left(\sum_{m=1}^M p_m (\hat{V}_{m, N_m^n} - V_m) \geq \frac{1}{4\gamma} \varepsilon\right) \\ &\leq \sum_{m=1}^M \mathbb{E}\left[\mathbb{P}\left(\frac{1}{N_m^n} \sum_{t=1}^{N_m^n} V_{m,t} - V_m \geq \frac{1}{4\gamma p_m} \varepsilon \mid N_m^n\right)\right] \\ &\leq \sum_{m=1}^M \mathbb{E}\left[c(N_m^n)^{-1} \left(\frac{\varepsilon}{4\gamma p_m}\right)^{-1}\right]. \end{aligned}$$

Let us define an event $\mathcal{E} = \left\{N_m^n \geq \frac{np_m}{2}\right\}$. Therefore,

$$\begin{aligned} A_2 &\leq \sum_{m=1}^M \mathbb{E}\left[c\left(\frac{np_m}{2}\right)^{-1} \left(\frac{\varepsilon}{4\gamma p_m}\right)^{-1}\right] \\ &+ \sum_{m=1}^M \mathbb{E}\left[\mathbb{P}\left(N_m^n < \frac{np_m}{2}\right)\right] = \sum_{m=1}^M (c2^{1+2}\gamma^1 p_m^{-1+1}) n^{-1} \varepsilon^{-1} \\ &+ \sum_{m=1}^M \mathbb{E}\left[\mathbb{P}\left(N_m^n - p_m n \leq -\frac{p_m n}{2}\right)\right] \\ &\leq \sum_{m=1}^M (c2^3\gamma) n^{-1} \varepsilon^{-1} + \sum_{m=1}^M \exp\left\{-2n\left(\frac{p_m n}{2}\right)^2\right\} \end{aligned}$$

We consider $p_m > 0$ only since if $p_m = 0$, $p_m(\hat{V}_{m, N_m^n} - V_m) = 0$, and has been eliminated. Therefore,

$$A \leq A_1 + A_2 \leq 2 \exp\left\{-\frac{n\varepsilon^2}{32M\gamma^2 L^2}\right\} + \sum_{m=1}^M (c2^3\gamma) n^{-1} \varepsilon^{-1} + \sum_{m=1}^M \exp\left\{-2n\left(\frac{p_m n}{2}\right)^2\right\}.$$

That leads to

$$\begin{aligned} &\mathbb{P}\left(\hat{Q}_n - (\mu + \gamma \langle p, V \rangle) \geq \varepsilon\right) \leq \exp\left\{-2n\frac{\varepsilon^2}{4}\right\} \\ &+ 2 \exp\left\{-\frac{n\varepsilon^2}{32M\gamma^2 L^2}\right\} + \sum_{m=1}^M (c2^3\gamma) n^{-1} \varepsilon^{-1} + \sum_{m=1}^M \exp\left\{-2n\left(\frac{p_m n}{2}\right)^2\right\} \leq c' n^{-1} \varepsilon^{-2}, \end{aligned}$$

with $c' > 0$ depends on c, M, p_i . We have the last inequality because to argue that $\exp(-cn\varepsilon^2) = \mathcal{O}(n^{-1}\varepsilon^{-2})$. Furthermore, with $0 < \varepsilon < 1$, $n^{-1}\varepsilon^{-1} \leq n^{-1}\varepsilon^{-2}$, and $\exp(-cn^3) \leq \mathcal{O}(n^{-1}\varepsilon^{-2})$. So that

$$\mathbb{P}\left(\hat{Q}_n - (\mu + \gamma \langle p, V \rangle) \geq \varepsilon\right) \leq c' n^{-1} \varepsilon^{-2},$$

By following the same steps, we can derive

$$\mathbb{P}\left(\hat{Q}_n - (\mu + \gamma \langle p, V \rangle) \leq -\varepsilon\right) \leq c' n^{-1} \varepsilon^{-2}.$$

Therefore, with $n \geq 1, \varepsilon > 0$,

$$\mathbb{P}\left(\left|\hat{Q}_n - (\mu + \gamma \langle p, V \rangle)\right| \geq \varepsilon\right) \leq c' n^{-1} \varepsilon^{-2}.$$

Furthermore,

$$\begin{aligned} \hat{Q}_n - (\mu + \gamma \langle p, V \rangle) &= (\hat{R}_n - \mu) + \left(\gamma \langle \hat{p}_n, \hat{V}_n \rangle - \gamma \langle p, V \rangle\right) \\ &= (\hat{R}_n - \mu) + \gamma \left(\langle \hat{p}_n - p, \hat{V}_n \rangle + \langle p, \hat{V} - V \rangle\right) \end{aligned}$$

Therefore,

$$\begin{aligned} \Rightarrow \left|\mathbb{E}[\hat{Q}_n] - (\mu + \gamma \langle p, V \rangle)\right| &\leq \left|\mathbb{E}[\hat{R}_n - \mu]\right| + \gamma \left(\left|\mathbb{E}[\hat{p}_n - p]\right| \left|\mathbb{E}[\hat{V}_n]\right| + p \left|\mathbb{E}[\hat{V} - V]\right|\right) \\ \Rightarrow \left|\mathbb{E}[\hat{Q}_n] - (\mu + \gamma \langle p, V \rangle)\right| &\leq \left|\mathbb{E}[\hat{R}_n - \mu]\right| + \gamma \left(L \left|\mathbb{E}[\hat{p}_n - p]\right| + p \left|\mathbb{E}[\hat{V} - V]\right|\right) \end{aligned}$$

Also because $\lim_{n \rightarrow \infty} \mathbb{E}[\hat{V}_{m,n}] = V_m$, $\lim_{n \rightarrow \infty} \frac{\hat{N}_m^n}{n} = p_m$, and $\mathbb{E}[(\hat{R}_n - \mu)] = 0$ so that,

$$\lim_{n \rightarrow \infty} \mathbb{E}[\hat{Q}_n] = \mu + \gamma \sum_{m=1}^M p_m V_m.$$

That mean

$$\text{plim}_{n \rightarrow \infty} \hat{Q}_n = \mu + \gamma \sum_{m=1}^M p_m V_m,$$

which concludes the proof. \square

Results from Lemma 1 is important as it shows the concentration for the Q value estimation given the concentration of V value of the children nodes.

Lemma 2. Let consider non-negative variables $x, y \in \mathbb{R}^+$, and a constant m that $0 \leq m \leq 1$. Then

$$(x + y)^m \leq x^m + y^m.$$

We use Minkowski's inequality as shown below

Lemma 3. (Minkowski's inequality) Given $p \geq 1, \{x_i, y_i\} \in \mathbb{R}, i = 1, 2, \dots, n$, then we have the following inequality

$$\left(\sum_i (|x_i + y_i|)^p\right)^{\frac{1}{p}} \leq \left(\sum_i (|x_i|)^p\right)^{\frac{1}{p}} + \left(\sum_i (|y_i|)^p\right)^{\frac{1}{p}}.$$

Proof. This is a basic result. \square

Lemma 4. (Markov's inequality) If X is a nonnegative random variable and $a > 0$, then the probability that X is at least a is at most the expectation of X divided by a :

$$\mathbb{P}(X > a) \leq \frac{\mathbb{E}[X]}{a}.$$

Proof. This is a well-known result. \square

C CONVERGENCE OF CATSO AND PATSO IN NON-STATIONARY MULTI-ARMED BANDITS

We note that in an MCTS tree, each node is considered a non-stationary multi-armed bandit where the average mean drifts due to the given action selection strategy. Therefore, we first study the convergence of CATSO and PATSO in non-stationary multi-armed bandits where the action selection is Thompson sampling, with the power mean backup operator at the root node. Detailed descriptions of the CATSO and PATSO in Non-stationary multi-armed bandits settings can be found in the main article in the Theoretical Analysis section.

We first establish the convergence and concentration properties for the power mean backup operator in non-stationary bandits, detailed in Theorem 1 for CATSO and Theorem 2 for PATSO.

To achieve these results, we demonstrate that the expected payoff of the power mean backup operator decays polynomially at a rate of $O(\frac{\log n}{n})$. This is supported by Lemma 7 for CATSO and Lemma 8 for PATSO. Critical to this analysis are Lemma 5 and Lemma 6, which establish an upper bound of $\log(n)$ for the expected number of suboptimal arm pulls.

We introduce some important definitions. F_a^n represents the empirical cumulative distribution function of arm a after n visitations, and F_a represents the cumulative distribution function of arm a . We employ the following distance measure: If P and Q are two distributions characterized by parameters $p = (p_0, p_1, \dots, p_N)$ and $q = (q_0, q_1, \dots, q_N)$ respectively, then the distance is defined as

$$d(P, Q) := \|p - q\|_\infty = \sup_{i \in [0, N]} |p_i - q_i|$$

This represents the L^∞ distance between p and q in \mathbb{R}^{N+1} . We also denote $\text{KL}(P \parallel Q)$ as the Kullback–Leibler divergence between P and Q , and denote $\mathcal{K}_{\text{inf}}(F_a, \mu_\star) = \inf_{G: \mathbb{E}[G] > \mu_\star} \text{KL}(F_a \parallel G)$. In addition, we denote $\mathcal{K}_{\text{inf}}^{(N)}(F_a, \mu_\star) = \inf \left\{ \text{KL}(F_a \parallel G) \mid \text{the support of } G \in \left\{ 0, \frac{R(t)}{N}, \frac{2R(t)}{N}, \dots, R(t) \right\}, \mathbb{E}[G] > \mu_\star \right\}$.

We see that the definition of $\mathcal{K}_{\text{inf}}(F_a, \mu_\star)$ and $\mathcal{K}_{\text{inf}}^{(N)}(F_a, \mu_\star)$ is only difference in the support set.

We denote the true parameter of arm a by $p_a = (p_a^0, p_a^1, \dots, p_a^N)$ with $p_a^i = \mathbb{P}_{X \sim F_a}[X = \frac{i}{N}]$. We denote the parameter of the posterior distribution of arm a as $\alpha_a = (\alpha_a^0, \alpha_a^1, \dots, \alpha_a^N)$. Since each arm a is non-stationary, we also denote the parameter of arm a after n visitations by $p_a(n) = (p_a^0(n), p_a^1(n), \dots, p_a^N(n))$ with $p_a^i(n) = \mathbb{P}_{X \sim F_a^n}[X = \frac{i}{N}]$. The parameter of the posterior distribution of arm a denoted as $\alpha_a(n) = (\alpha_a^0(n), \alpha_a^1(n), \dots, \alpha_a^N(n))$. We first show the results of an important Lemma 5. The proof follows closely to the Proof of Proposition 7 (Riou and Honda, 2020). The only difference is that in our settings, we study non-stationary bandits.

Lemma 5. Assume a non-stationary bandit setting satisfies Assumption 1, and that the CATSO (Categorical Thompson Sampling with Optimistic Bonus) algorithm is applied. Let $T_a(n)$ denote the number of times arm a is pulled by time step n . If a indexes a suboptimal arm, then for any $\varepsilon_0 \geq 0$, the expected count of arm a satisfies

$$\mathbb{E}[T_a(n)] \leq \frac{(1 + \varepsilon_0) \log n}{\mathcal{K}_{\text{inf}}^{(N)}(F_a, \mu_\star)} + \mathcal{O}(\sqrt{n}).$$

Proof. We have $\bar{\theta}_a(t) = [0, \frac{R(t)}{N}, \frac{2R(t)}{N}, \dots, R(t)]^\top L_{a,t}$, with $L_{a,t} \sim \text{Dir}(\alpha_a^0(t), \dots, \alpha_a^N(t))$.

To analyze the expectation associated with selecting a suboptimal arm a , we follow the approach of (Riou and Honda, 2020) and decompose the term into two components:

$$\begin{aligned} \mathbb{E} \left[\sum_{t=1}^n \mathbb{1}(I(t) = a) \right] &= \underbrace{\mathbb{E} \left[\sum_{t=1}^n \mathbb{1}(I(t) = a, \bar{\theta}_a(t) \geq \mu_* - C \frac{t^{\frac{1}{4}}}{T_a(t)^{\frac{1}{2}}}, d(\hat{F}_{I(t)}, F_{I(t)}) \leq \varepsilon_2) \right]}_{A_1} \\ &\quad + \underbrace{\mathbb{E} \left[\sum_{t=1}^n \mathbb{1} \left(I(t) = a, \left\{ \bar{\theta}_a(t) < \mu_* - C \frac{t^{\frac{1}{4}}}{T_a(t)^{\frac{1}{2}}} \right\} \cup \left\{ d(\hat{F}_{I(t)}, F_{I(t)}) > \varepsilon_2 \right\} \right) \right]}_{A_2} \end{aligned}$$

We first find an upper bound for A_1 :

$$A_1 = \sum_{t=1}^n \sum_{m=1}^n \mathbb{E} \left[\mathbb{1} \left(I(t) = a, \bar{\theta}_a(t) \geq \mu_* - C \frac{t^{\frac{1}{4}}}{T_a(t)^{\frac{1}{2}}}; \left\| \frac{\alpha_a(t)}{T_a(t) + N + 1} - p_a(t) \right\|_{\infty} \leq \varepsilon_2, T_a(t) = m \right) \right]$$

We first note that if the event

$$\left\{ I(t) = a, \bar{\theta}_a(t) \geq \mu_* - C \frac{t^{\frac{1}{4}}}{T_a(t)^{\frac{1}{2}}}; \left\| \frac{\alpha_a(t)}{T_a(t) + N + 1} - p_a(t) \right\|_{\infty} \leq \varepsilon_2, T_a(t) = m \right\}$$

occurs at step t for a certain $m \in [1, n]$, then $T_a(t') > T_a(t) = m$ for any $t' > t$.

Thus, for any $m \in [n]$

$$\sum_{t=1}^n \mathbb{1} \left(I(t) = a, \bar{\theta}_a(t) \geq \mu_* - \varepsilon_1; \left\| \frac{\alpha_a(t)}{T_a(t) + N + 1} - p_a(t) \right\|_{\infty} \leq \varepsilon_2, T_a(t) = m \right) \leq 1.$$

We deduce that for any $m_0 \in [n]$,

$$\begin{aligned} A_1 &\leq m_0 + \sum_{t=1}^n \sum_{m=m_0}^n \mathbb{E} \left[\mathbb{1} \left(I(t) = a, \bar{\theta}_a(t) \geq \mu_* - C \frac{t^{\frac{1}{4}}}{T_a(t)^{\frac{1}{2}}}; \left\| \frac{\alpha_a(t)}{T_a(t) + N + 1} - p_a(t) \right\|_{\infty} \leq \varepsilon_2, T_a(t) = m \right) \right] \\ &\leq m_0 + \sum_{t=1}^n \sum_{m=m_0}^n \mathbb{P} \left(\bar{\theta}_a(t) \geq \mu_* - C \frac{t^{\frac{1}{4}}}{T_a(t)^{\frac{1}{2}}}; \left\| \frac{\alpha_a(t)}{T_a(t) + N + 1} - p_a(t) \right\|_{\infty} \leq \varepsilon_2, T_a(t) = m \right) \\ &\leq m_0 + \sum_{t=1}^n \sum_{m=m_0}^n \mathbb{P} \left(\bar{\theta}_a(t) \geq \mu_* - C \frac{t^{\frac{1}{4}}}{T_a(t)^{\frac{1}{2}}} \mid \left\| \frac{\alpha_a(t)}{T_a(t) + N + 1} - p_a(t) \right\|_{\infty} \leq \varepsilon_2, T_a(t) = m \right) \\ &\quad \times \mathbb{P} \left(\left\| \frac{\alpha_a(t)}{T_a(t) + N + 1} - p_a(t) \right\|_{\infty} \leq \varepsilon_2, T_a(t) = m \right). \end{aligned}$$

For any fixed $\varepsilon_1 > 0$, if we choose m_0 such that

$$m_0 \geq \frac{C}{\varepsilon_1} \sqrt{n},$$

then

$$\begin{aligned} A_1 &\leq m_0 + \sum_{t=1}^n \sum_{m=m_0}^n \mathbb{P} \left(\bar{\theta}_a(t) \geq \mu_* - \varepsilon_1 \mid \left\| \frac{\alpha_a(t)}{T_a(t) + N + 1} - p_a(t) \right\|_{\infty} \leq \varepsilon_2, T_a(t) = m \right) \\ &\quad \times \mathbb{P} \left(\left\| \frac{\alpha_a(t)}{T_a(t) + N + 1} - p_a(t) \right\|_{\infty} \leq \varepsilon_2, T_a(t) = m \right). \end{aligned}$$

By applying the results of Lemma 13, Appendix F (Riou and Honda, 2020), we have

$$\mathbb{P}\left(\bar{\theta}_a(t) \geq \mu_\star - \varepsilon_1 \mid \alpha_a(t), T_a(t) = m\right) \leq C_0(m+N+1)^{N/2} \exp\{-(m+N+1)\text{KL}(P_{\alpha_a(t)} \parallel P_{\mu_\star - \varepsilon_1}^*)\}$$

where

$$P_{\mu_\star - \varepsilon_1}^* = \arg \min_{x: u^\top x \geq \mu_\star - \varepsilon_1} \text{KL}(P_{\alpha_a} \parallel x) \quad \text{and} \quad P_{\alpha_a(t)} = \frac{1}{n+N+1} \alpha_a(t).$$

By definition, we have $\text{KL}(P_{\alpha_a(t)} \parallel P_{\mu_\star - \varepsilon_1}^*) = \mathcal{K}_{\text{inf}}(P_{\alpha_a(t)}, \mu_\star - \varepsilon_1)$, and

$$\mathbb{P}\left(\bar{\theta}_a(t) \geq \mu_\star - \varepsilon_1 \mid \alpha_a(t), T_a(t) = m\right) \leq C_0(m+N+1)^{N/2} \exp\{-(m+N+1)\mathcal{K}_{\text{inf}}(P_{\alpha_a(t)}, \mu_\star - \varepsilon_1)\},$$

$$\text{where } C_0 = \frac{\exp\{1/12\}}{\Gamma(N+1)} \left(\frac{1}{\sqrt{2\pi}}\right)^N.$$

On the other hand, $\mathcal{K}_{\text{inf}}(x, \mu_\star - \varepsilon_1)$ is continuous in $x \in [0, 1]^{N+1}$ on the probability simplex with respect to the L^∞ distance from ((Honda and Takemura, 2010), Theorem 7) and Lemma 18 in Appendix H (Riou and Honda, 2020). Therefore, for any $\varepsilon_3 > 0$, there exists $\varepsilon_2 > 0$ and constant $C' > 0$ such that

$$\begin{aligned} \mathbb{P}\left(\bar{\theta}_a(t) \geq \mu_\star - \varepsilon_1 \mid \left\| \frac{\alpha_a(t)}{T_a(t) + N + 1} - p_a(t) \right\|_\infty \leq \varepsilon_2, T_a(t) = m\right) \\ \leq C' \exp\{-(m+N+1)(\mathcal{K}_{\text{inf}}(p_a, \mu_\star - \varepsilon_1) - \varepsilon_3)\} \end{aligned}$$

Note that $\mathbb{P}\left(\left\| \frac{\alpha_a(t)}{T_a(t) + N + 1} - p_a(t) \right\|_\infty \leq \varepsilon_2, T_a(t) = m\right) \leq 1$. We have

$$\begin{aligned} A1 &\leq m_0 + C'_1 \sum_{t=1}^n \exp\{-(m_0 + N + 1)(\mathcal{K}_{\text{inf}}(p_a, \mu_\star - \varepsilon_1) - \varepsilon_3)\} \\ &\leq m_0 + C'_1 n \exp\{-(m_0 + N + 1)(\mathcal{K}_{\text{inf}}(p_a, \mu_\star - \varepsilon_1) - \varepsilon_3)\} \end{aligned}$$

Furthermore, as from ((Honda and Takemura, 2010), Theorem 7), it is proven that $\mu \rightarrow \mathcal{K}_{\text{inf}}(F, \mu)$ is continuous for $\mu < 1$, when we scale reward from $[0, 1]$ to $[0, R]$ therefore μ from $[0, 1]$ to $[0, R]$. We have $\mu \rightarrow \mathcal{K}_{\text{inf}}(F, \mu)$ is continuous for $\mu < R$. Therefore, $\forall \varepsilon_4 > 0, \exists \varepsilon_1 > 0$, such that

$$|\mathcal{K}_{\text{inf}}(p_a, \mu^* - \varepsilon_1) - \mathcal{K}_{\text{inf}}(p_a, \mu^*)| \leq \varepsilon_4$$

which implies that for any ε_0 , there exists $\varepsilon_1, \varepsilon_2$ such that

$$\mathcal{K}_{\text{inf}}(p_a, \mu^* - \varepsilon_1) - \varepsilon_3 \geq \frac{1}{1 + \varepsilon_0} \mathcal{K}_{\text{inf}}(p_a, \mu^*).$$

Thus, if we choose

$$\begin{aligned} m_0 &= \max \left\{ \frac{C}{\varepsilon_1} \sqrt{n}, \frac{\log n}{\mathcal{K}_{\text{inf}}(p_a, \mu_\star - \varepsilon_1) - \varepsilon_3} - N - 1 \right\} \\ &\leq \mathcal{O}(\sqrt{n}) + \frac{(1 + \varepsilon_0) \log n}{\mathcal{K}_{\text{inf}}^{(N)}(F_a, \mu_\star)}. \end{aligned}$$

then $A1 \leq m_0 + C'_1$.

Upper bound for A_2 :

To provide an upper bound for A_2 , we decompose the term into two parts:

$$\begin{aligned} A_2 &= \mathbb{E} \left[\underbrace{\sum_{t=1}^n \mathbb{1} \left(I(t) = a, \left\{ \bar{\theta}_a(t) < \mu_\star - C \frac{t^{\frac{1}{4}}}{T_a(t)^{\frac{1}{2}}} \right\} \right)}_{R1} \right] \\ &\quad + \mathbb{E} \left[\underbrace{\sum_{t=1}^n \mathbb{1} \left(I(t) = a, \left\{ d(\hat{F}_{I(t)}, F_{I(t)}) > \varepsilon_2 \right\} \right)}_{R2} \right] \end{aligned}$$

According to Proposition 8 (Riou and Honda, 2020), for any $\varepsilon_2 > 0$, we have

$$R2 \leq O(1).$$

To bound B_1 , we note that

$$R1 \leq \mathbb{E} \left[\sum_{t=1}^n \mathbb{1} \left(\bar{\theta}_a(t) < \mu_* - C \frac{t^{\frac{1}{4}}}{T_a(t)^{\frac{1}{2}}} \right) \right]$$

We decompose the set

$$\{1, 2, \dots, n\} = \bigcup V_x, \quad \text{where } V_x = \left\{ t : x \leq \frac{T_a(t)}{\sqrt{t}} < x+1 \right\}, \quad x \in \{0, 1, \dots, \sqrt{n}\}$$

We have

$$\begin{aligned} \mathbb{E} \left[\sum_{t \in V_x} \mathbb{1} \left(\bar{\theta}_a(t) < \mu_* - C \frac{t^{\frac{1}{4}}}{T_a(t)^{\frac{1}{2}}} \right) \right] &\leq \mathbb{E} \left[\sum_{t \in V_x} \mathbb{1} \left(\bar{\theta}_a(t) < \mu_* - \frac{C}{\sqrt{x+1}} \right) \right] \\ &\leq \mathbb{E} \left[\sum_{t=1}^n \mathbb{1} \left(\bar{\theta}_a(t) < \mu_* - \frac{C}{\sqrt{x+1}} \right) \right] = \mathcal{O}(1) \end{aligned}$$

by Proposition 8 (Riou and Honda, 2020). Since there is at most $\sqrt{n} + 1$ such non-empty V_x , we deduce that

$$R1 = \mathcal{O}(\sqrt{n}).$$

By combining the upper bounds on A_1 , R_1 and R_2 , we have

$$\mathbb{E}[T_a(n)] \leq \frac{(1 + \varepsilon_0) \log n}{\mathcal{K}_{\inf}^{(N)}(F_a, \mu_*)} + \mathcal{O}(\sqrt{n}).$$

This completes the proof. \square

Lemma 6. Suppose the payoff sequence satisfies Assumption 1, and consider a non-stationary bandit setting in which Particle Thompson Sampling with Optimistic Bonus (PATSO) is employed. Let $T_a(n)$ denote the number of times a suboptimal arm a is played up to time n . For any $\varepsilon_0 \geq 0$, the expected number of pulls of each suboptimal arm a satisfies

$$\mathbb{E}[T_a(n)] \leq \frac{\log n}{\mathcal{K}_{\inf}(F_a, \mu_*) - \varepsilon_0} + \mathcal{O}(\sqrt{n}).$$

where μ_* is the (unique) optimal mean reward, F_a is the distribution associated with arm a .

Proof. In this Theorem, we use the Levy distance. Recall that the Levy distance between two cumulative distribution functions F and G on $[0, 1]$ is defined as

$$D_L(F, G) = \inf\{\varepsilon > 0 : \forall x \in [0, 1], F(x - \varepsilon) - \varepsilon \leq G(x) \leq F(x + \varepsilon) + \varepsilon\}.$$

The proof follows the same steps as in Lemma 5. We derive

$$\begin{aligned} \mathbb{E} \left[\sum_{t=1}^n \mathbb{1}(I(t) = a) \right] &= \underbrace{\mathbb{E} \left[\sum_{t=1}^n \mathbb{1}(I(t) = a), \bar{\varphi}_{a,t} \geq \mu_* - C \frac{t^{\frac{1}{4}}}{T_a(t)^{\frac{1}{2}}}, D_L(\hat{F}_{I(t)}, F_{I(t)}) \leq \varepsilon_2 \right]}_{B1} \\ &\quad + \underbrace{\mathbb{E} \left[\sum_{t=1}^n \mathbb{1}(I(t) = a), \left\{ \bar{\varphi}_{a,t} < \mu_* - C \frac{t^{\frac{1}{4}}}{T_a(t)^{\frac{1}{2}}} \right\} \cup \left\{ D_L(\hat{F}_{I(t)}, F_{I(t)}) > \varepsilon_2 \right\} \right]}_{B2} \end{aligned}$$

As in Lemma 5, when for any fixed $\varepsilon_1 > 0$, if we choose m_0 such that

$$m_0 \geq \frac{C}{\varepsilon_1} \sqrt{n},$$

then

$$B1 \leq m_0 + \sum_{t=1}^n \sum_{m=m_0}^n \mathbb{P} \left(\bar{\theta}_a(t) \geq \mu_* - \varepsilon_1 \middle| D_L(\hat{F}_a(t), F_a(t)) \leq \varepsilon_2, T_a(t) = m \right) \\ \times \mathbb{P} \left(D_L(\hat{F}_a(t), F_a(t)) \leq \varepsilon_2, T_a(t) = m \right)$$

According to Lemma 15 in Appendix G.1 (Riou and Honda, 2020) on conditional probabilities, for any $\nu \in (0, 1)$ we have

$$\mathbb{P} \left(\bar{\theta}_a(t) \geq \mu_* - \varepsilon_1 \middle| D_L(\hat{F}_a(t), F_a(t)) \leq \varepsilon_2, T_a(t) = m \right) \\ \leq \frac{1}{\nu} \exp \left\{ -m \left(\mathcal{K}_{\inf}(\hat{F}_a(t), \mu_* - \varepsilon_1) - \nu \frac{\mu_* - \varepsilon_1}{1 - (\mu_* - \varepsilon_1)} \right) \right\}$$

Because $\mathcal{K}_{\inf}(F, \mu)$ is continuous in F with respect to the Levy distance from (Honda and Takemura, 2010), Theorem 7, for any $\varepsilon_3 > 0$ there exists $\varepsilon_2 > 0$ such that

$$D_L(\hat{F}_a(t), F_a) \leq \varepsilon_2 \Rightarrow \left| \mathcal{K}_{\inf}(\hat{F}_a(t), \mu_* - \varepsilon_1) - \mathcal{K}_{\inf}(F_a, \mu_* - \varepsilon_1) \right| \leq \varepsilon_3$$

Therefore, $\forall \nu \in (0, 1)$ and for any $\varepsilon_5 > 0$, there exists $\varepsilon_1, \varepsilon_2 > 0$ such that

$$\mathbb{P} \left(\bar{\theta}_a(t) \geq \mu_* - \varepsilon_1 \middle| D_L(\hat{F}_a(t), F_a(t)) \leq \varepsilon_2, T_a(t) = m \right) \\ \leq \frac{1}{\nu} \left(-m \left(\mathcal{K}_{\inf}(F_a, \mu_* - \varepsilon_1) - \varepsilon_3 - \nu \frac{\mu_* - \varepsilon_1}{1 - (\mu_* - \varepsilon_1)} \right) \right) \\ \stackrel{(\text{Theorem 6 (Honda and Takemura, 2010)})}{\leq} \frac{1}{\nu} \left(-m \left(\mathcal{K}_{\inf}(F_a, \mu_*) - \frac{\varepsilon_1}{1 - \mu_*} - \varepsilon_3 - \nu \frac{\mu_* - \varepsilon_1}{1 - (\mu_* - \varepsilon_1)} \right) \right)$$

This implies that $\forall \varepsilon_0 > 0$, there exists $\nu \in (0, 1)$, $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ such that

$$\mathbb{P} \left(\bar{\theta}_a(t) \geq \mu_* - \varepsilon_1 \middle| D_L(\hat{F}_a(t), F_a(t)) \leq \varepsilon_2, T_a(t) = m \right) \leq \frac{1}{\nu} \exp \{ -m(\mathcal{K}_{\inf}(F_a, \mu_*) - \varepsilon_0) \}$$

Therefore, according to inequality equation C and the fact that

$$\mathbb{P} \left(D_L(\hat{F}_a(t), F_a(t)) \leq \varepsilon_2, T_a(t) = m \right) \leq 1,$$

we have

$$B1 \leq m_0 + \sum_{t=1}^n \frac{1}{\nu} \exp \{ -m_0(\mathcal{K}_{\inf}(F_a, \mu_*) - \varepsilon_0) \} \leq m_0 + \frac{1}{\nu} T \exp \{ -m_0(\mathcal{K}_{\inf}(F_a, \mu_*) - \varepsilon_0) \}.$$

Choose

$$m_0 = \max \left\{ \frac{C}{\varepsilon_1} \sqrt{n}, \frac{\log n}{\mathcal{K}_{\inf}(F_a, \mu_*) - \varepsilon_0} \right\} \leq \mathcal{O}(\sqrt{n}) + \frac{\log n}{\mathcal{K}_{\inf}(F_a, \mu_*) - \varepsilon_0}.$$

we have

$$B1 \leq \frac{\log n}{\mathcal{K}_{\inf}(F_a, \mu_*) - \varepsilon_0} + \mathcal{O}(\sqrt{n}).$$

Using the same argument as in the proof of Lemma 5, we have $B2 \leq \mathcal{O}(\sqrt{n})$.

That leads us to

$$\mathbb{E}[T_a(n)] \leq \frac{\log n}{\mathcal{K}_{\inf}(F_a, \mu_*) - \varepsilon_0} + \mathcal{O}(\sqrt{n}),$$

which completes the proof. \square

Lemma 7. Let **CATSO** (Categorical Thompson Sampling with Optimistic Bonus) be applied to a non-stationary bandit problem whose payoff sequence satisfies Assumption 1. Define the power-mean estimator

$$\hat{\mu}_n(p) = \left(\sum_{a=1}^K \frac{T_a(n)}{n} [\hat{\mu}_{a,T_a(n)}]^p \right)^{\frac{1}{p}},$$

and let $\delta_{*,n} = \mu_* - \mu_{*,n}$. For any $p \geq 1$ and $\varepsilon_0 > 0$, the following upper bound on $|\mathbb{E}[\hat{\mu}_n(p)] - \mu_*|$ holds:

$$|\mathbb{E}[\hat{\mu}_n(p)] - \mu_*| \leq |\delta_{*,n}| + \frac{R}{n} \sum_{\substack{a=1 \\ a \neq a_*}}^K \left\{ \frac{(1 + \varepsilon_0) \log n}{\mathcal{K}^{(N)}(F_a, \mu_*)} + \mathcal{O}(\sqrt{n}) \right\}.$$

Proof. We observe that

$$|\hat{\mu}_n(p) - \mu_*| \leq |\hat{\mu}_n(p) - \mu_{*,n}| + |\mu_* - \mu_{*,n}| = |\hat{\mu}_n(p) - \mu_{*,n}| + |\delta_{*,n}|$$

Furthermore,

$$\hat{\mu}_{a,T_a(n)} \leq \mu_{a,n} + |\hat{\mu}_{a,T_a(n)} - \mu_{a,n}|.$$

Since $\mu_{*,n} = \max_{a \in [K]} \{\mu_{a,n}\}$, we have

$$\begin{aligned} \hat{\mu}_n(p) - \mu_{*,n} &= \hat{\mu}_n(p) - \sum_{a=1}^K \frac{T_a(n)}{n} \mu_{*,n} \leq \left(\sum_{a=1}^K \frac{T_a(n)}{n} (\hat{\mu}_{a,T_a(n)})^p \right)^{\frac{1}{p}} - \left(\sum_{a=1}^K \frac{T_a(n)}{n} (\mu_{a,n})^p \right)^{\frac{1}{p}} \\ &= \frac{\left(\sum_{a=1}^K T_a(n) (\hat{\mu}_{a,T_a(n)})^p \right)^{\frac{1}{p}} - \left(\sum_{a=1}^K T_a(n) (\mu_{a,n})^p \right)^{\frac{1}{p}}}{n^{\frac{1}{p}}} \end{aligned}$$

Applying Minkowski's inequality from Lemma 3, and the result of equation C, we have

$$\begin{aligned} \hat{\mu}_n(p) - \mu_{*,n} &\leq \frac{\left(\sum_{a=1}^K T_a(n) (\mu_a + |\hat{\mu}_{a,T_a(n)} - \mu_{a,n}|)^p \right)^{\frac{1}{p}} - \left(\sum_{a=1}^K T_a(n) (\mu_{a,n})^p \right)^{\frac{1}{p}}}{n^{\frac{1}{p}}} \\ &\leq \frac{\left(\sum_{a=1}^K T_a(n) (|\hat{\mu}_{a,T_a(n)} - \mu_{a,n}|)^p \right)^{\frac{1}{p}}}{n^{\frac{1}{p}}} \end{aligned}$$

On the other hand,

$$\begin{aligned} \mu_{*,n} - \hat{\mu}_n(p) &= \frac{n\mu_{*,n} - n\hat{\mu}_n(p)}{n} = \frac{n\mu_{*,n} - (\sum_{a=1}^K T_a(n)\mu_{a,n}) + \sum_{a=1}^K T_a(n)\mu_{a,n} - n\hat{\mu}_n(p)}{n} \\ &= \frac{\sum_{a=1, a \neq a_*}^K T_a(n) |\mu_{*,n} - \mu_{a,n}| + \sum_{a=1}^K T_a(n)\mu_{a,n} - n\hat{\mu}_n(p)}{n} \\ &\leq R \sum_{a=1, a \neq a_*}^K \frac{T_a(n)}{n} + \sum_{a=1}^K \frac{T_a(n)}{n} \mu_{a,n} - \hat{\mu}_n(p) \end{aligned}$$

Because power mean is an increasing function of p , so that

$$\sum_{a=1}^K \frac{T_a(n)}{n} \mu_{a,n} \leq \left(\sum_{a=1}^K \frac{T_a(n)}{n} (\mu_{a,n})^p \right)^{1/p}.$$

Furthermore, we observe that

$$\mu_{a,n} \leq \hat{\mu}_{a,T_a(n)} + |\hat{\mu}_{a,T_a(n)} - \mu_{a,n}|.$$

So that, from equation equation C we have

$$\begin{aligned}
\mu_{\star,n} - \hat{\mu}_n(p) &\leq R \sum_{a=1, a \neq a_*}^K \frac{T_a(n)}{n} + \left(\sum_{a=1}^K \frac{T_a(n)}{n} (\mu_{a,n})^p \right)^{1/p} - \hat{\mu}_n(p) \\
&\leq R \sum_{a=1, a \neq a_*}^K \frac{T_a(n)}{n} \\
&\quad + \frac{\left(\sum_{a=1}^K T_a(n) (\hat{\mu}_{a,T_a(n)} + |\hat{\mu}_{a,T_a(n)} - \mu_{a,n}|)^p \right)^{\frac{1}{p}} - \left(\sum_{a=1}^K T_a(n) (\hat{\mu}_{a,T_a(n)})^p \right)^{\frac{1}{p}}}{n^{\frac{1}{p}}} \\
&\stackrel{\text{(Minkovski's inequality)}}{\leq} R \sum_{a=1, a \neq a_*}^K \frac{T_a(n)}{n} + \frac{\left(\sum_{a=1}^K T_a(n) (|\hat{\mu}_{a,T_a(n)} - \mu_{a,n}|)^p \right)^{\frac{1}{p}}}{n^{\frac{1}{p}}} \\
&\stackrel{\text{(Properties of } L^p \text{ norm)}}{\leq} R \sum_{a=1, a \neq a_*}^K \frac{T_a(n)}{n} + \frac{\left(\sum_{a=1}^K T_a(n) (|\hat{\mu}_{a,T_a(n)} - \mu_{a,n}|) \right)}{n^{\frac{1}{p}}} \\
&= R \sum_{a=1, a \neq a_*}^K \frac{T_a(n)}{n} + \frac{\sum_{a=1}^K \left(\left| \sum_t^{T_a(n)} R_{a,t} - T_a(n) \mu_{a,n} \right| \right)}{n^{\frac{1}{p}}}
\end{aligned}$$

Therefore

$$\begin{aligned}
|\mathbb{E}[\hat{\mu}_n(p) - \mu_{\star,n}]| &\leq R \sum_{a=1, a \neq a_*}^K \frac{\mathbb{E}[T_a(n)]}{n} + \frac{\mathbb{E} \left[\left| \sum_{a=1}^K \sum_t^{T_a(n)} R_{a,t} - T_a(n) \mu_{a,n} \right| \right]}{n^{\frac{1}{p}}} \\
&= R \sum_{a=1, a \neq a_*}^K \frac{\mathbb{E}[T_a(n)]}{n}
\end{aligned}$$

Please note that because we study non-stationary bandits, $\mathbb{E}[\sum_t^n R_{a,t}] = n\mu_{a,n}$, therefore,

$$\frac{\mathbb{E} \left[\left| \sum_{a=1}^K \sum_t^{T_a(n)} R_{a,t} - T_a(n) \mu_{a,n} \right| \right]}{n^{\frac{1}{p}}} = 0$$

According to Lemma 5, we have

$$|\mathbb{E}[\hat{\mu}_n(p) - \mu_{\star,n}]| \leq R \sum_{a=1, a \neq a_*}^K \frac{\mathbb{E}[T_a(n)]}{n} \leq \frac{R}{n} \sum_{a=1, a \neq a_*}^K \left\{ \frac{(1 + \varepsilon_0) \log n}{\mathcal{K}^{(N)}(F_a, \mu_{\star})} + \mathcal{O}(\sqrt{n}) \right\},$$

which concludes the proof. \square

Lemma 8. Let PATSO (Particle Thompson Sampling with Optimistic Bonus) be applied to a non-stationary bandit problem whose payoff process satisfies Assumption 1. Define the power-mean estimator

$$\hat{\mu}_n(p) = \left(\sum_{a=1}^K \frac{T_a(n)}{n} [\hat{\mu}_{a,T_a(n)}]^p \right)^{1/p},$$

and let $\delta_{\star,n} = \mu_{\star} - \mu_{\star,n}$. For any $p \geq 1$ and $\varepsilon_0 > 0$, the following holds:

$$\left| \mathbb{E}[\hat{\mu}_n(p)] - \mu_{\star} \right| \leq |\delta_{\star,n}| + \frac{R}{n} \sum_{\substack{a=1 \\ a \neq a_*}}^K \left\{ \frac{\log n}{\mathcal{K}_{\inf}(F_a, \mu_{\star}) - \varepsilon_0} + \mathcal{O}(\sqrt{n}) \right\}.$$

Proof. Similar to Lemma 7, we can derive

$$|\mathbb{E}[\hat{\mu}_n(p)] - \mu_{*,n}| \leq |\delta_{*,n}| + R \sum_{a=1, a \neq a_*}^K \frac{\mathbb{E}[T_a(n)]}{n}.$$

and the bound is obtained as a direct corollary of Lemma 6. \square

Theorem 1. (CATSO in Non-Stationary Bandits) Let $(\hat{\mu}_{a,n})_{n \geq 1}$ be a bounded sequence in $[0, R]$ satisfying Assumption 1, and let $\mu_* = \max_{a \in [K]} \mu_a$. Assume CATSO chooses each arm once initially, then follows the exploration strategy in Fig. 2. For any $p \geq 1$, the power-mean estimator $\hat{\mu}_n(p)$ converges to μ_* in the sense of Definition 1, i.e.

$$\text{plim}_{n \rightarrow \infty} \hat{\mu}_n(p) = \mu_*.$$

Proof. We first prove that $\lim_{n \rightarrow \infty} \mathbb{E}[\hat{\mu}_n(p)] = \mu_*$. According to the result of Lemma 7, we have

$$\begin{aligned} |\mathbb{E}[\hat{\mu}_n(p)] - \mu_*| &\leq |\delta_{*,n}| + R \sum_{a=1, a \neq a_*}^K \frac{\mathbb{E}[T_a(n)]}{n} \\ &\leq |\delta_{*,n}| + \frac{R}{n} \sum_{a=1, a \neq a_*}^K \left\{ \frac{(1 + \varepsilon_0) \log n}{\mathcal{K}^{(N)}(F_a, \mu_*)} + o(\sqrt{n}) + O(1) \right\} \end{aligned}$$

with $\delta_{*,n} = \mu_* - \mu_{*,n}$, and because $\lim_{n \rightarrow \infty} \mu_{*,n} = \mu_*$, we can conclude that

$$\lim_{n \rightarrow \infty} \mathbb{E}[\hat{\mu}_n(p)] = \mu_*.$$

Second, we prove that

$$\forall n \geq 1, \forall 0 < \varepsilon \leq n^{-1/2} (\log n)^{-1/2}, \exists c > 0 \text{ that } \mathbb{P}(|\hat{\mu}_n(p) - \mu_*| > \varepsilon) \leq cn^{-1} \varepsilon^{-2}.$$

We observe that

$$\begin{aligned} |\hat{\mu}_n(p) - \mu_*| &\leq |\hat{\mu}_n(p) - \mu_{*,n}| + |\mu_* - \mu_{*,n}| = |\hat{\mu}_n(p) - \mu_{*,n}| + |\delta_{*,n}| \\ \implies \mathbb{P}(|\hat{\mu}_n(p) - \mu_*| \geq \varepsilon) &\leq \mathbb{P}(|\hat{\mu}_n(p) - \mu_{*,n}| \geq \varepsilon/2) + \mathbb{P}(|\delta_{*,n}| \geq \varepsilon/2). \end{aligned}$$

Because $\lim_{n \rightarrow \infty} |\delta_{*,n}| = 0$, therefore, $\exists N_0 > 0$ such that $\forall n \geq N_0$, we have $|\delta_{*,n}| < \varepsilon/2$ that means

$$\forall n > N_0, \mathbb{P}(|\delta_{*,n}| \geq \varepsilon/2) = 0.$$

Next, according to Lemma 7,

$$|\mathbb{E}[\hat{\mu}_n(p)] - \mu_{*,n}| \leq \frac{R}{n} \sum_{a=1, a \neq a_*}^K \left\{ \frac{(1 + \varepsilon_0) \log n}{\mathcal{K}^{(N)}(F_a, \mu_*)} + o(\sqrt{n}) + O(1) \right\} = O(n^{-1/2}),$$

that leads to

$$\mathbb{P}(|\hat{\mu}_n(p) - \mu_{*,n}| \geq \varepsilon/2) \leq \frac{|\mathbb{E}[\hat{\mu}_n(p)] - \mu_{*,n}|}{\varepsilon/2} = \frac{O(n^{-1/2})}{\varepsilon/2}.$$

Therefore, $\exists c > 0, \forall 0 < \varepsilon \leq n^{-1/2}$ such that

$$\mathbb{P}(|\hat{\mu}_n(p) - \mu_{*,n}| \geq \varepsilon/2) \leq cn^{-1/2}(\varepsilon)(\varepsilon^{-2}) \leq cn^{-1/2} \left(n^{-1/2} \right) \varepsilon^{-2} = cn^{-1} \varepsilon^{-2},$$

which means

$$\forall n \geq N_0, \forall 0 < \varepsilon < n^{-1/2}, \exists c > 0 \text{ that } \mathbb{P}(|\hat{\mu}_n(p) - \mu_*| > \varepsilon) \leq cn^{-1} \varepsilon^{-2}.$$

With $0 < \varepsilon \leq n^{-1/2} \Rightarrow \varepsilon^{-2} \leq n^{-1}$, so that $n^{-1} \varepsilon^{-2} \leq n^{-2}$.

With $1 \leq n < N_0 \Rightarrow n^{-1} \varepsilon^{-2} \leq N_0^{-2}$. Therefore

$$\forall C > N_0^{-2} \Rightarrow \mathbb{P}(|\hat{\mu}_n(p) - \mu_*| > \varepsilon) \leq 1 < Cn^{-1} \varepsilon^{-2},$$

which means

$$\forall n \geq 1, \forall 0 < \varepsilon \leq n^{-1/2}, \exists C > 0 \text{ that } \mathbb{P}(|\hat{\mu}_n(p) - \mu_\star| > \varepsilon) \leq C n^{-1} \varepsilon^{-2}.$$

Next, we prove the following claim

Claim. For any power-mean order $p \geq 1$,

$$\mathbb{P}(|\hat{\mu}_n(p) - \mu_\star| > \varepsilon) \leq C n^{-1} \varepsilon^{-2}, \quad \forall \varepsilon > n^{-1/2},$$

for a universal constant $C > 0$ depending only on problem parameters (K, R , etc.).

Setup & decomposition. By Lemma 7 we have

$$|\hat{\mu}_n(p) - \mu_\star| \leq R \sum_{a \neq a_\star} \frac{T_a(n)}{n} + \left(\sum_{a=1}^K \frac{T_a(n)}{n} |\hat{\mu}_{a, T_a(n)} - \mu_a|^p \right)^{1/p}.$$

Define

$$A_n := R \sum_{a \neq a_\star} \frac{T_a(n)}{n}, \quad B_n := \left(\sum_{a=1}^K \frac{T_a(n)}{n} |\hat{\mu}_{a, T_a(n)} - \mu_a|^p \right)^{1/p}.$$

1. Bound on A_n . Let $S_n = \sum_{a \neq a_\star} T_a(n)$. Lemma 5 implies $\mathbb{E}[S_n] \leq c_1(K-1)\sqrt{n}$. By Azuma–Hoeffding,

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq \exp\left(-\frac{t^2}{2n}\right).$$

Choose $t = \frac{\varepsilon n}{2R} - c_1(K-1)\sqrt{n}$ (valid since $\varepsilon > n^{-1/2}$), then

$$= \mathbb{P}\left(S_n > \frac{\varepsilon n}{2R}\right) \leq \exp(-c_3 n \varepsilon^2) \leq \frac{1}{2} n^{-1} \varepsilon^{-2}.$$

2. Bound on B_n . Rewards in $[0, R]$ imply $\text{Var}(\hat{\mu}_{a, m}) \leq R^2/(4m)$. By total expectation,

$$\mathbb{E}[|\hat{\mu}_{a, T_a(n)} - \mu_a|^2] \leq \mathbb{E}\left[\frac{R^2}{4T_a(n)}\right].$$

For any $p \geq 1$ we can bound the power mean by the ℓ_2 -mean:

$$B_n^2 \leq \sum_{a=1}^K \frac{T_a(n)}{n} |\hat{\mu}_{a, T_a(n)} - \mu_a|^2 \Rightarrow \mathbb{E}[B_n^2] \leq \frac{KR^2}{4n}.$$

Chebyshev then gives

$$\mathbb{P}(B_n > \varepsilon/2) \leq \frac{4\mathbb{E}[B_n^2]}{\varepsilon^2} \leq \frac{KR^2}{n\varepsilon^2}.$$

3. Union bound. Combining (1)–(2),

$$\mathbb{P}(|\hat{\mu}_n(p) - \mu_\star| > \varepsilon) \leq \mathbb{P}(A_n > \varepsilon/2) + \mathbb{P}(B_n > \varepsilon/2) \leq \frac{\frac{1}{2} + KR^2}{n\varepsilon^2} = C n^{-1} \varepsilon^{-2},$$

which completes the proof. \square

Theorem 2. (*PATSO* in Non-Stationary Bandits) Under the same assumptions as Theorem 1, if instead *PATSO* is employed, the identical convergence result holds:

$$\text{plim}_{n \rightarrow \infty} \hat{\mu}_n(p) = \mu_\star.$$

Proof. The proof follows the same steps as Theorem 1. We first prove that $\lim_{n \rightarrow \infty} \mathbb{E}[\hat{\mu}_n(p)] = \mu_*$. According to the result of Lemma 8, we have

$$\begin{aligned} |\mathbb{E}[\hat{\mu}_n(p)] - \mu_*| &\leq |\delta_{*,n}| + R \sum_{a=1, a \neq a_*}^K \frac{\mathbb{E}[T_a(n)]}{n} \\ &\leq |\delta_{*,n}| + \frac{R}{n} \sum_{a=1, a \neq a_*}^K \left\{ \frac{\log n}{\mathcal{K}_{\inf}(F_a, \mu_*) - \varepsilon_0} + o(\sqrt{n}) + O(1) \right\} \end{aligned}$$

with $\delta_{*,n} = \mu_* - \mu_{*,n}$, and because $\lim_{n \rightarrow \infty} \mu_{*,n} = \mu_*$, we can conclude that

$$\lim_{n \rightarrow \infty} \mathbb{E}[\hat{\mu}_n(p)] = \mu_*.$$

Second, we prove that

$$\forall n \geq 1, \forall 0 < \varepsilon < n^{-1/2}, \exists c > 0 \text{ that } \mathbb{P}(|\hat{\mu}_n(p) - \mu_*| > \varepsilon) \leq cn^{-1}\varepsilon^{-2}.$$

We observe that

$$\begin{aligned} |\hat{\mu}_n(p) - \mu_*| &\leq |\hat{\mu}_n(p) - \mu_{*,n}| + |\mu_* - \mu_{*,n}| = |\hat{\mu}_n(p) - \mu_{*,n}| + |\delta_{*,n}| \\ \implies \mathbb{P}(|\hat{\mu}_n(p) - \mu_*| \geq \varepsilon) &\leq \mathbb{P}(|\hat{\mu}_n(p) - \mu_{*,n}| \geq \varepsilon/2) + \mathbb{P}(|\delta_{*,n}| \geq \varepsilon/2). \end{aligned}$$

Because $\lim_{n \rightarrow \infty} |\delta_{*,n}| = 0$, therefore, $\exists N_0 > 0$ such that $\forall n \geq N_0$, we have $|\delta_{*,n}| < \varepsilon/2$ that means

$$\forall n > N_0, \mathbb{P}(|\delta_{*,n}| \geq \varepsilon/2) = 0.$$

Next, according to Lemma 8,

$$|\mathbb{E}[\hat{\mu}_n(p)] - \mu_{*,n}| \leq \frac{R}{n} \sum_{a=1, a \neq a_*}^K \left\{ \frac{\log n}{\mathcal{K}_{\inf}(F_a, \mu_*) - \varepsilon_0} + o(\sqrt{n}) + O(1) \right\} = O(n^{-1/2}),$$

that leads to

$$\mathbb{P}(|\hat{\mu}_n(p) - \mu_{*,n}| \geq \varepsilon/2) \leq \frac{|\mathbb{E}[\hat{\mu}_n(p)] - \mu_{*,n}|}{\varepsilon/2} = \frac{O(n^{-1/2})}{\varepsilon/2}.$$

Therefore, $\exists c > 0, \forall 0 < \varepsilon < n^{-1/2}$ such that

$$\mathbb{P}(|\hat{\mu}_n(p) - \mu_{*,n}| \geq \varepsilon/2) \leq cn^{-1/2}(\varepsilon)(\varepsilon^{-2}) \leq cn^{-1/2} \left(n^{-1/2} \right) \varepsilon^{-2} = cn^{-1}\varepsilon^{-2},$$

which means

$$\forall n \geq N_0, \forall 0 < \varepsilon < n^{-1/2}, \exists c > 0 \text{ that } \mathbb{P}(|\hat{\mu}_n(p) - \mu_*| > \varepsilon) \leq cn^{-1}\varepsilon^{-2}.$$

With $0 < \varepsilon < n^{-1/2} \Rightarrow \varepsilon^{-2} > n^{-1}$, so that $n^{-1}\varepsilon^{-2} > n^{-2}$.

With $1 \leq n < N_0 \Rightarrow n^{-1}\varepsilon^{-2} > N_0^{-2}$. Therefore

$$\forall C > N_0^{-2} \Rightarrow \mathbb{P}(|\hat{\mu}_n(p) - \mu_*| > \varepsilon) \leq 1 < Cn^{-1}\varepsilon^{-2},$$

which means

$$\forall n \geq 1, \forall 0 < \varepsilon < n^{-1/2}, \exists C > 0 \text{ that } \mathbb{P}(|\hat{\mu}_n(p) - \mu_*| > \varepsilon) \leq Cn^{-1}\varepsilon^{-2}.$$

Similarly to CATSO, we prove the following claim

Claim. For any power-mean order $p \geq 1$,

$$\mathbb{P}(|\hat{\mu}_n(p) - \mu_*| > \varepsilon) \leq Cn^{-1}\varepsilon^{-2}, \quad \forall \varepsilon > n^{-1/2},$$

for a universal constant $C > 0$ depending only on problem parameters (K, R , etc.).

Setup & decomposition. By Lemma 7 we have

$$|\hat{\mu}_n(p) - \mu_*| \leq R \sum_{a \neq a_*} \frac{T_a(n)}{n} + \left(\sum_{a=1}^K \frac{T_a(n)}{n} |\hat{\mu}_{a,T_a(n)} - \mu_a|^p \right)^{1/p}.$$

Define

$$A_n := R \sum_{a \neq a_*} \frac{T_a(n)}{n}, \quad B_n := \left(\sum_{a=1}^K \frac{T_a(n)}{n} |\hat{\mu}_{a,T_a(n)} - \mu_a|^p \right)^{1/p}.$$

1. Bound on A_n . Let $S_n = \sum_{a \neq a_*} T_a(n)$. Lemma 5 implies $\mathbb{E}[S_n] \leq c_1(K-1)\sqrt{n}$. By Azuma–Hoeffding,

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) \leq \exp\left(-\frac{t^2}{2n}\right).$$

Choose $t = \frac{\varepsilon n}{2R} - c_1(K-1)\sqrt{n}$ (valid since $\varepsilon > n^{-1/2}$), then

$$= \mathbb{P}\left(S_n > \frac{\varepsilon n}{2R}\right) \leq \exp(-c_3 n \varepsilon^2) \leq \frac{1}{2} n^{-1} \varepsilon^{-2}.$$

2. Bound on B_n . Rewards in $[0, R]$ imply $\text{Var}(\hat{\mu}_{a,m}) \leq R^2/(4m)$. By total expectation,

$$\mathbb{E}[|\hat{\mu}_{a,T_a(n)} - \mu_a|^2] \leq \mathbb{E}\left[\frac{R^2}{4T_a(n)}\right].$$

For any $p \geq 1$ we can bound the power mean by the ℓ_2 -mean:

$$B_n^2 \leq \sum_{a=1}^K \frac{T_a(n)}{n} |\hat{\mu}_{a,T_a(n)} - \mu_a|^2 \Rightarrow \mathbb{E}[B_n^2] \leq \frac{KR^2}{4n}.$$

Chebyshev then gives

$$\mathbb{P}(B_n > \varepsilon/2) \leq \frac{4\mathbb{E}[B_n^2]}{\varepsilon^2} \leq \frac{KR^2}{n\varepsilon^2}.$$

3. Union bound. Combining (1)–(2),

$$\mathbb{P}(|\hat{\mu}_n(p) - \mu_*| > \varepsilon) \leq \mathbb{P}(A_n > \varepsilon/2) + \mathbb{P}(B_n > \varepsilon/2) \leq \frac{\frac{1}{2} + KR^2}{n\varepsilon^2} = C n^{-1} \varepsilon^{-2},$$

which completes the proof. \square

D CONVERGENCE OF CATSO AND PATSO IN MONTE-CARLO TREE SEARCH

Based upon the results of CATSO and PATSO using power mean as the value backup operator on the described non-stationary multi-armed bandit problem, we derive theoretical results for CATSO in an MCTS tree.

We derive Theorem 3 for CATSO and Theorem 4 for PATSO, which show concentration and convergence for any internal node in the tree. These proofs utilize induction, leveraging the results of Lemma 7 for CATSO and Lemma 8 for PATSO, and Lemma 5 for CATSO and Lemma 6 for PATSO. Additionally, we use Lemma 1, which demonstrates the concentration and convergence of an estimated Q-value based on the child V-value node, applying it recursively throughout the tree.

Our main results, Theorem 5 for CATSO and Theorem 5 for PATSO, show that the simple regret converges non-asymptotically at a rate of $O(n^{-1/2})$.

Theorem 3. Let CATSO run on a tree of depth H . For each node s_h at level h , the following statements hold:

i) For any action $a_k \in \mathcal{A}_{s_h}$,

$$\lim_{n \rightarrow \infty} \hat{Q}_n(s_h, a_k) = \tilde{Q}(s_h, a_k).$$

ii)

$$\text{plim}_{n \rightarrow \infty} \hat{V}_n(s_h) = \tilde{V}(s_h).$$

Proof. We prove this by induction on the maximum depth D of the search tree.

Base Case ($D = 1$). Consider a tree of depth 1 whose root node is s_0 . When taking action a_k from s_0 , the algorithm observes an immediate reward $r_t(s_0, a_k)$ and transitions to a leaf s_1 . Let

$$R(s_0, a_k) = \mathbb{E}[r_t(s_0, a_k)]$$

denote the mean immediate reward, and recall

$$\tilde{Q}(s_0, a_k) = R(s_0, a_k) + \gamma \sum_{s_1 \in \mathcal{A}_{s_0}} \mathbb{P}(s_1 | s_0, a_k) \tilde{V}(s_1).$$

We can see CATSO estimate Q value (this is straight forward as simple derivation from pseudocode) as

$$\hat{Q}_n(s_0, a_k) = \frac{1}{n} \sum_{t=1}^n r_t(s_0, a_k) + \gamma \sum_{s_1 \sim \tau(s_0, a_k)} \frac{T_{s_0, a_k}^{s_1}(n)}{n} \hat{V}_{T_{s_0, a_k}^{s_1}(n)}(s_1),$$

where $T_{s_0, a_k}^{s_1}(n)$ is the number of visits at s_0, a_k, s_1

First, by Lemma 1 (treating $\{r_t\}$ as i.i.d. intermediate rewards with child-state probabilities p_m), one obtains

$$\text{plim}_{n \rightarrow \infty} \hat{Q}_n(s_0, a_k) = \tilde{Q}(s_0, a_k).$$

Second, from Theorem 1 (or directly from plugging in the limit above), we deduce

$$\text{plim}_{n \rightarrow \infty} \hat{V}_n(s_0) = \tilde{V}(s_0).$$

since $\tilde{V}(s_0) = \max_{a \in \mathcal{A}_{s_0}} \tilde{Q}(s_0, a)$ and

$$\hat{V}_n(s_0) = \left(\sum_{a \in \mathcal{A}_{s_0}} \frac{T_{s_0, a}(n)}{n} \left[\hat{Q}_{T_{s_0, a}(n)}(s_0, a) \right]^p \right)^{\frac{1}{p}}.$$

Hence, the result holds for $D = 1$.

Inductive Step. Suppose the statements (i) and (ii) are valid for all nodes in any tree of depth $\leq D$. We now verify that they also hold for a tree of depth $D + 1$.

At the root s_0 , after taking action a_k , the algorithm transitions to a subtree of depth D . By the inductive hypothesis, for every node in that subtree (including its root $s_1 \sim \tau(s_0, a_k)$), the limit properties of Q and V hold. Hence, each $\hat{V}_n(s_1)$ converges to its corresponding $\tilde{V}(s_1)$, ensuring that

$$\text{plim}_{n \rightarrow \infty} \hat{Q}_n(s_0, a_k) = R(s_0, a_k) + \gamma \sum_{s_1 \sim \tau(s_0, a_k)} \mathbb{P}(s_1 | s_0, a_k) \text{ and } \text{plim}_{n \rightarrow \infty} \hat{V}_n(s_1) = \tilde{Q}(s_0, a_k).$$

To conclude (ii), note that

$$\tilde{V}(s_0) = \max_{a \in \mathcal{A}_{s_0}} \tilde{Q}(s_0, a), \quad \text{while} \quad \hat{V}_n(s_0) = \left(\sum_{a \in \mathcal{A}_{s_0}} \frac{T_{s_0, a}(n)}{n} \left[\hat{Q}_{T_{s_0, a}(n)}(s_0, a) \right]^p \right)^{\frac{1}{p}}.$$

Since each $\hat{Q}_n(s_0, a)$ converges to the correct $\tilde{Q}(s_0, a)$, it follows immediately that

$$\text{plim}_{n \rightarrow \infty} \hat{V}_n(s_0) = \tilde{V}(s_0).$$

Therefore, by induction, both statements (i) and (ii) hold for every node in a tree of any finite depth H . This completes the proof. \square

Similarly we can derive the following Theorem

Theorem 4. Consider a tree of maximum depth H on which **PATSO** is run. Then, for any node s_h at depth h and any action $a_k \in \mathcal{A}_{s_h}$:

i)

$$\text{plim}_{n \rightarrow \infty} \widehat{Q}_n(s_h, a_k) = \widetilde{Q}(s_h, a_k).$$

ii)

$$\text{plim}_{n \rightarrow \infty} \widehat{V}_n(s_h) = \widetilde{V}(s_h).$$

Proof. The proof follows the same steps as Theorem 3 by applying the results of Lemma 1 and Theorem 2. \square

Notation. At a node s , $T_s(n) = \sum_a T_{s,a}(n)$ is the visit count and the p -power-mean backup is

$$\widehat{V}_n(s) = \left(\sum_a \frac{T_{s,a}(n)}{T_s(n)} (\widehat{Q}_n(s, a))^p \right)^{1/p}, \quad p \geq 1.$$

Let $\mu_{a,n}(s)$ denote the *drifting bandit mean* at (s, a) and $\mu_n^*(s) = \max_a \mu_{a,n}(s)$. Define the *drift* $\delta_n^*(s) := \widetilde{V}(s) - \mu_n^*(s) \geq 0$.

Theorem 5. (Convergence of Expected Payoff of CATSO) We have at the root node s_0 , there exists a constant $C' > 0$ such that

$$\mathbb{E}[|\widehat{V}_n(s_0) - \widetilde{V}(s_0)|] \leq C' n^{-\frac{1}{2}},$$

Proof. We argue in four steps.

1 (One-step decomposition at a node). For any node s ,

$$\mathbb{E}|\widehat{V}_n(s) - \widetilde{V}(s)| \leq \underbrace{\mathbb{E}|\widehat{V}_n(s) - \mu_n^*(s)|}_{\text{statistical term}} + \underbrace{\delta_n^*(s)}_{\text{drift term}}.$$

2 (Control of the statistical term via Lemma 7 at the node). Lemma 7 (Appendix C) gives, for CATSO and any $p \geq 1$,

$$\mathbb{E}|\widehat{V}_n(s) - \mu_n^*(s)| \leq \frac{R}{T_s(n)} \sum_{a \neq a^*} \mathbb{E}[T_{s,a}(n)] + \delta_n^*(s),$$

where $a^* \in \arg \max_a \mu_{a,n}(s)$ and rewards are in $[0, R]$. Combining equation D and equation D and bounding the (duplicated) $\delta_n^*(s)$ only once, we obtain

$$\mathbb{E}|\widehat{V}_n(s) - \widetilde{V}(s)| \leq \frac{R}{T_s(n)} \sum_{a \neq a^*} \mathbb{E}[T_{s,a}(n)] + \delta_n^*(s).$$

By Lemma 5 (Appendix C), for any suboptimal arm $a \neq a^*$,

$$\mathbb{E}[T_{s,a}(n)] \leq \frac{(1 + \varepsilon_0) \log T_s(n)}{K_{\inf}^{(N)}(F_a, \mu^*)} + O(\sqrt{T_s(n)}).$$

At the root s_0 we have $T_{s_0}(n) = n$, hence

$$\mathbb{E}|\widehat{V}_n(s_0) - \mu_n^*(s_0)| \leq C_1 \frac{\log n}{n} + C_2 n^{-1/2} = O(n^{-1/2}).$$

3 (Drift contraction down the tree via Lemma 1). Pick an optimal root action $a^* \in \arg \max_a Q^e(s_0, a)$. Since $\mu_n^*(s_0) \geq \mu_{a^*,n}(s_0)$,

$$0 \leq \delta_n^*(s_0) = \tilde{V}(s_0) - \mu_n^*(s_0) \leq \tilde{V}(s_0) - \mu_{a^*,n}(s_0).$$

By Lemma 1 (Appendix B),

$$\mu_{a^*,n}(s_0) = r(s_0, a^*) + \gamma \sum_{s_1} P(s_1|s_0, a^*) \mathbb{E}[\hat{V}_n(s_1)],$$

whence

$$\delta_n^*(s_0) \leq \gamma \sum_{s_1} P(s_1|s_0, a^*) \mathbb{E}|\hat{V}_n(s_1) - \tilde{V}(s_1)|.$$

Under Lemma 5, suboptimal root actions get only $O(\sqrt{n}) + O(\log n)$ pulls, so $T_{s_0, a^*}(n) = n - O(\sqrt{n})$ in expectation. Thus each child s_1 on a^* 's branch is visited $\Theta(n)$ times in expectation, and the nodewise bound equation D applies at s_1 as well, giving $\mathbb{E}|\hat{V}_n(s_1) - \mu_n^*(s_1)| = O(n^{-1/2})$. Iterating equation D down the tree (base case at leaves has zero drift by definition),

$$\delta_n^*(s_0) \leq \gamma C n^{-1/2} + \gamma^2 C n^{-1/2} + \dots \leq \frac{C}{1-\gamma} n^{-1/2},$$

or $CH n^{-1/2}$ for finite horizon H if $\gamma = 1$.

4 (Combine). From equation D and the drift bound,

$$\mathbb{E}|\hat{V}_n(s_0) - \tilde{V}(s_0)| \leq O(n^{-1/2}) + O(n^{-1/2}) = O(n^{-1/2}).$$

□

Alternative Proof of Theorem 5 - Direct Bias Propagation. We prove that $\mathbb{E}[\hat{V}_n(s_0) - \tilde{V}(s_0)] \leq C' n^{-1/2}$ by directly analyzing the bias at each node level.

1: Base case - Leaf nodes. At leaf nodes (depth H), we have $\hat{V}_n(s_H) = V_0(s_H)$ by definition, so there's no estimation error.

2: Inductive step - Internal nodes. Consider any internal node s_h at depth $h < H$. By Theorem 3, we know:

$$\text{plim}_{n \rightarrow \infty} \hat{Q}_n(s_h, a) = \tilde{Q}(s_h, a) \quad \forall a \in \mathcal{A}_{s_h}$$

From Lemma 7 (adapted to the MCTS setting), for each Q-node:

$$\left| \mathbb{E}[\hat{Q}_n(s_h, a)] - \tilde{Q}(s_h, a) \right| \leq \frac{C_1}{n^{1/2}}$$

3: Power-mean is 1-Lipschitz. The power-mean operator \mathcal{P}_p defined by

$$\mathcal{P}_p(q_1, \dots, q_K; w_1, \dots, w_K) = \left(\sum_{i=1}^K w_i q_i^p \right)^{1/p}$$

is 1-Lipschitz with respect to the ℓ_∞ norm on the q_i 's when the weights w_i form a probability distribution. That is:

$$|\mathcal{P}_p(q) - \mathcal{P}_p(q')| \leq \|q - q'\|_\infty$$

4: Bias propagation. At node s_h , the V-value estimate is:

$$\hat{V}_n(s_h) = \mathcal{P}_p \left(\hat{Q}_{T_{s_h, a_1}(n)}(s_h, a_1), \dots, \hat{Q}_{T_{s_h, a_K}(n)}(s_h, a_K); \frac{T_{s_h, a_1}(n)}{n}, \dots, \frac{T_{s_h, a_K}(n)}{n} \right)$$

By the 1-Lipschitz property and the bias bound on each Q-value:

$$\begin{aligned} \left| \mathbb{E}[\hat{V}_n(s_h)] - \tilde{V}(s_h) \right| &\leq \left| \mathbb{E}[\hat{V}_n(s_h)] - \mathcal{P}_p(\tilde{Q}(s_h, a_1), \dots, \tilde{Q}(s_h, a_K)) \right| \\ &\leq \max_{a \in \mathcal{A}_{s_h}} \left| \mathbb{E}[\hat{Q}_{T_{s_h, a}(n)}(s_h, a)] - \tilde{Q}(s_h, a) \right| \\ &\leq \frac{C_1}{n^{1/2}} \end{aligned}$$

5: Root node conclusion. Since the same $O(n^{-1/2})$ bias bound holds at every node level, and there are at most H levels, we get at the root:

$$\left| \mathbb{E}[\widehat{V}_n(s_0)] - \tilde{V}(s_0) \right| \leq \frac{C_1}{n^{1/2}}$$

By Jensen's inequality:

$$\mathbb{E}[|\widehat{V}_n(s_0) - \tilde{V}(s_0)|] \leq \mathbb{E}[|\widehat{V}_n(s_0) - \mathbb{E}[\widehat{V}_n(s_0)]|] + |\mathbb{E}[\widehat{V}_n(s_0)] - \tilde{V}(s_0)|$$

The first term is $O(n^{-1/2})$ by concentration (Theorem 3), and the second term is $O(n^{-1/2})$ by the above analysis. Therefore:

$$\mathbb{E}[|\widehat{V}_n(s_0) - \tilde{V}(s_0)|] \leq C'n^{-1/2}$$

□

Corollary 1 (PATSO). Replacing Lemma 7 by Lemma 8 (Appendix C) in Step 2 yields the same rate for PATSO: there exists $C' > 0$ such that $\mathbb{E}|\widehat{V}_n(s_0) - \tilde{V}(s_0)| \leq C'n^{-1/2}$.

Remark 4 (Why this avoids the tail-bound pitfall). The argument works purely in expectation at each node by combining: (i) the visit-weighted bandit bounds (Lemmas 7/8) and visit counts (Lemmas 5/6), and (ii) the one-step coupling (Lemma 1). No global extension of a polynomial tail beyond its stated range is needed.

Next, we present the results of Theorem 6. The proof follows the same steps as Theorem 5.

Theorem 6. (Convergence of Expected Payoff of PATSO) We have at the root node s_0 , there exists a constant $C' > 0$ such that

$$\mathbb{E}[|\widehat{V}_n(s_0) - \tilde{V}(s_0)|] \leq C' n^{-\frac{1}{2}},$$

Proof. The proof follows the same analysis as in Theorem 5. □

E DISTRIBUTIONAL MCTS AND WASSERSTEIN ROBUST OPTIMIZATION

The distributional MCTS approach presented in our paper extends naturally beyond sequential decision-making to the broader domain of robust planning under uncertainty. In this section, we establish formal connections between our algorithms (CATSO and PATSO) and the field of Wasserstein Distributionally Robust Optimization (WDRO), which offers a principled framework for decision-making under distributional uncertainty.

In our paper, both CATSO and PATSO maintain distributional representations of Q-values:

- **CATSO:** Uses a categorical distribution with atoms at locations $\{z_i(s, a)\}_{i=0}^{N-1}$ and probabilities $\{p_i(s, a)\}_{i=0}^{N-1}$
- **PATSO:** Uses a particle-based representation with particles $\mathcal{S}(s, a)$ and weights $\alpha(s, a)$

These representations capture empirical distributions $\widehat{P}_{s,a}$ of returns for each state-action pair (s, a) . We now show that our approach can be reinterpreted within the Wasserstein Distributionally Robust MDP (WDRMDP) framework.

The standard WDRMDP formulation is:

$$\max_{\pi} \min_{P \in \mathcal{B}_{\varepsilon}(\widehat{P})} \mathbb{E}_P[V^{\pi}]$$

where:

- π is the policy being optimized
- \widehat{P} is the empirical distribution (from samples)
- $\mathcal{B}_{\varepsilon}(\widehat{P}) = \{P : W_p(P, \widehat{P}) \leq \varepsilon\}$ is the Wasserstein ball of radius ε around \widehat{P}

- W_p is the p -Wasserstein distance

This connection provides a theoretical foundation for applying our methods to robust planning problems beyond standard MCTS settings.

A key advantage of our distributional approach is that it enables rigorous sample complexity guarantees for robust planning. We formalize this through the concept of (ε, δ) -robust policies.

Definition 2 ((ε, δ) -Robust Policy). A policy π is (ε, δ) -robust if, with probability at least $1 - \delta$, its performance is within ε of the optimal robust policy:

$$\mathbb{P} \left[\min_{P \in \mathcal{B}_\varepsilon(\hat{P})} \mathbb{E}_P[V^\pi] \geq \max_{\pi'} \min_{P \in \mathcal{B}_\varepsilon(\hat{P})} \mathbb{E}_P[V^{\pi'}] - \varepsilon \right] \geq 1 - \delta$$

Theorem 7 (Sample Complexity for Distributionally Robust Planning). *Let \mathcal{M} be an MDP with state space \mathcal{S} , action space \mathcal{A} , bounded rewards in $[0, R_{\max}]$, and discount factor γ . To learn an (ε, δ) -robust policy using the distributional representations from CATSO/PATSO combined with a concentration-based Wasserstein radius, the required number of samples is:*

$$O \left(\left[\frac{R_{\max}^3}{\varepsilon(1-\gamma)^3} \log \left(\frac{H|\mathcal{S}|^2|\mathcal{A}|}{\delta} \right) \right]^{2H} \right).$$

Proof. We provide a complete proof through several carefully developed steps. The proof applies to both CATSO and PATSO, with an additional consideration for discretization error in CATSO.

1: Wasserstein Concentration Bounds. For distributions with bounded support in $[0, R_{\max}]$, we can apply concentration inequalities for the Wasserstein distance between the empirical distribution $\hat{P}_{s,a}$ (based on $n_{s,a}$ i.i.d. samples) and the true distribution $P_{s,a}$.

For the p -Wasserstein distance ($p \geq 1$), Fournier and Guillin (2015) provide the following concentration inequality

$$\mathbb{P}[W_p(\hat{P}_{s,a}, P_{s,a}) > t] \leq c_1 \exp(-c_2 n_{s,a} t^2)$$

when $n_{s,a}$ is sufficiently large and t is sufficiently small. Here, c_1 and c_2 are constants that depend on the diameter of the support. For rewards bounded in $[0, R_{\max}]$, the diameter of the support for discounted returns is at most $\frac{R_{\max}}{1-\gamma}$. Simplifying the constants, we get

$$\mathbb{P} \left[W_p(\hat{P}_{s,a}, P_{s,a}) > \sqrt{\frac{cR_{\max}^2 \log(1/\delta)}{(1-\gamma)^2 n_{s,a}}} \right] \leq \delta$$

for an appropriate constant $c > 0$. This means that with probability at least $1 - \delta$, the Wasserstein distance between the empirical and true distributions is bounded by the right-hand side.

2: Confidence-Adjusted Wasserstein Radius. We need to ensure that for all state-action pairs (s, a) , the true distribution is contained within a Wasserstein ball centered at the empirical distribution with high probability. Using the concentration bound from Step 1 and applying a union bound over all $|\mathcal{S}||\mathcal{A}|$ state-action pairs, we set

$$\varepsilon_{s,a}(n) = \sqrt{\frac{cR_{\max}^2 \log(|\mathcal{S}||\mathcal{A}|/\delta)}{(1-\gamma)^2 n_{s,a}}}$$

With this choice, we have

$$\mathbb{P} \left[\forall (s, a) \in \mathcal{S} \times \mathcal{A} : W_p(P_{s,a}, \hat{P}_{s,a}) \leq \varepsilon_{s,a}(n) \right] \geq 1 - \delta$$

This means that with probability at least $1 - \delta$, all true distributions are contained within their respective Wasserstein balls.

3: Value Function Error Propagation. We now analyze how errors in the empirical distributions propagate to errors in the value functions. We consider two MDPs with reward distributions P and Q where $W_p(P, Q) \leq \varepsilon$.

Let V_1, V_2 be the value function for policy π under P and Q , respectively. For any state s , the Bellman equation gives

$$V_1(s) = \mathbb{E}_{(r,s') \sim P_{s,\pi(s)}} [r + \gamma V_1(s')]$$

Similarly, for the empirical MDP,

$$V_2(s) = \mathbb{E}_{(r,s') \sim Q_{s,\pi(s)}} [r + \gamma V_2(s')].$$

Let $\Delta(s) = V_1(s) - V_2(s)$. We have

$$\begin{aligned} \Delta(s) &= \mathbb{E}_{P_{s,\pi(s)}} [r + \gamma V_1(s')] - \mathbb{E}_{Q_{s,\pi(s)}} [r + \gamma V_2(s')] \\ &= \mathbb{E}_{P_{s,\pi(s)}} [r + \gamma V_1(s')] - \mathbb{E}_{Q_{s,\pi(s)}} [r + \gamma V_1(s')] \\ &\quad + \mathbb{E}_{Q_{s,\pi(s)}} [\gamma V_1(s')] - \mathbb{E}_{Q_{s,\pi(s)}} [\gamma V_2(s')] \\ &= \underbrace{\mathbb{E}_{P_{s,\pi(s)}} [r + \gamma V_1(s')] - \mathbb{E}_{Q_{s,\pi(s)}} [r + \gamma V_1(s')]}_{\text{Distribution approximation error}} + \gamma \mathbb{E}_{Q_{s,\pi(s)}} [V_1(s') - V_2(s')]. \end{aligned}$$

For the first term, we use the key property of Wasserstein distances: if two distributions P and Q satisfy $W_p(P, Q) \leq \varepsilon$, then for any L -Lipschitz function f ,

$$|\mathbb{E}_P[f] - \mathbb{E}_Q[f]| \leq L\varepsilon.$$

The function $f(r, s') = r + \gamma V_1(s')$ is Lipschitz with constant $L = 1 + \gamma \frac{R_{\max}}{1-\gamma} = \frac{1}{1-\gamma}$ (since V_1 is bounded by $\frac{R_{\max}}{1-\gamma}$ and has Lipschitz constant $\frac{1}{1-\gamma}$). Therefore

$$|\mathbb{E}_{P_{s,\pi(s)}} [r + \gamma V_1(s')] - \mathbb{E}_{Q_{s,\pi(s)}} [r + \gamma V_1(s')]| \leq \frac{\varepsilon_{s,\pi(s)}(n)}{1-\gamma}.$$

For the second term, we have

$$\gamma |\mathbb{E}_{Q_{s,\pi(s)}} [V_1(s') - V_2(s')]| \leq \gamma \max_{s'} |\Delta(s')|.$$

Combining these results, we obtain

$$|\Delta(s)| \leq \frac{\varepsilon_{s,\pi(s)}(n)}{1-\gamma} + \gamma \max_{s'} |\Delta(s')|.$$

Since this inequality holds for all states, it also holds for the state with maximum error

$$\max_s |\Delta(s)| \leq \frac{\max_{s,a} \varepsilon_{s,a}(n)}{1-\gamma} + \gamma \max_s |\Delta(s)|$$

which implies

$$\max_s |\Delta(s)| \leq \frac{1}{1-\gamma} \cdot \frac{\max_{s,a} \varepsilon_{s,a}(n)}{1-\gamma} = \frac{\max_{s,a} \varepsilon_{s,a}(n)}{(1-\gamma)^2}.$$

In other words, we have

$$\max_s |V_1(s) - V_2(s)| \leq \frac{R_{\max}}{(1-\gamma)^2} \max_{s,a} \varepsilon_{s,a}(n).$$

4: Robust Value Function Error. Now we consider the robust value functions. We define

$$V^{\pi, \text{rob}}(s) = \min_{P \in \mathcal{B}_\varepsilon(P_{\text{true}})} \mathbb{E}_P[V^\pi(s)], \quad \hat{V}^{\pi, \text{rob}}(s) = \min_{P \in \mathcal{B}_\varepsilon(\hat{P})} \mathbb{E}_P[V^\pi(s)].$$

Using the results obtained from Step 2 and Step 3, for all π , we have

$$|V^{\pi, \text{rob}}(s) - \widehat{V}^{\pi, \text{rob}}(s)| \leq \frac{3R_{\max}}{(1-\gamma)^2} \max_{s,a} \varepsilon_{s,a}(n)$$

and

$$|V^{\pi, \text{rob}}(s) - \mathbb{E}_{P_{true}}[V^{\pi}(s)]| \leq \frac{R_{\max}}{(1-\gamma)^2} \max_{s,a} \varepsilon_{s,a}(n).$$

Moreover, for **PATSO** policy (which we will denote by π_0), the results from previous sections dictate

$$\mathbb{E}_{P_{true}}[V^{\pi_0}(s)] \geq \max_{\pi'} \mathbb{E}_{P_{true}}[V^{\pi'}(s)] - \frac{C'}{\sqrt{n}}.$$

We conclude that

$$\mathbb{E}_{P_{true}}[V^{\pi_0}(s)] \geq \max_{\pi'} \min_{P \in \mathcal{B}_\varepsilon(\hat{P})} \mathbb{E}_P[V^{\pi'}] - \frac{4R_{\max}}{(1-\gamma)^2} \max_{s,a} \varepsilon_{s,a}(n) - \frac{C'}{\sqrt{n}}$$

with probability at least $1 - \delta$.

5: Sample Allocation. We note that the bounds derived from the previous steps depend on $n_{s,a} = T_{s,a}(n)$, the (random) number of times the state-action pairs (s, a) have been visited after n steps.

We first note that for some $C_0 > 0$, we have

$$\mathbb{E} \left[\sum_{a \neq a^*} T_{s,a}(n) \right] \leq C_0(|\mathcal{A}| - 1) \sqrt{T_s(n)}$$

which implies

$$\mathbb{P} \left[\sum_{a \neq a^*} T_{s,a}(n) \geq \frac{C_0(|\mathcal{A}| - 1)}{\delta} \sqrt{T_s(n)} \right] \leq \delta$$

via Markov's inequality. Thus, if we define the event

$$\mathcal{E}_n = \{T_{s,a^*}(n) \geq \frac{1}{2}T_s(n)\}$$

then $\mathbb{P}[\mathcal{E}_n] \geq 1 - \delta$ for

$$T_s(n) \geq \left(\frac{2C_0(|\mathcal{A}| - 1)}{\delta} \right)^2.$$

Under the event \mathcal{E}_n , we define K as the last time (up until step n) that a^* is the selected action. We note that

$$T_{s,a^*}(K) \geq \frac{1}{2}T_s(n),$$

and that for all a

$$\varphi_k(s, a) + C \frac{K^{1/4}}{T_{s,a}(K)^{1/2}} \leq \varphi_k(s, a^*) + C \frac{K^{1/4}}{T_{s,a^*}(K)^{1/2}} \leq R_{\max} + 2C \frac{T_s(n)^{1/4}}{T_s(n)^{1/2}} = R_{\max} + 2C.$$

We deduce that

$$T_{s,a}(n) \geq T_{s,a}(K) \geq \frac{C^2 K^{1/2}}{(R_{\max} + 2C)^2} \geq \frac{C^2}{\sqrt{2}(R_{\max} + 2C)^2} T_s(n)^{1/2}.$$

In other words, we have for all s ,

$$\mathbb{P} \left[\min_a T_{s,a}(n) \geq \frac{C^2}{\sqrt{2}(R_{\max} + 2C)^2} T_s(n)^{1/2} \right] \geq 1 - \delta.$$

Using an induction argument on the depth of the decision tree, we deduce that for any state s_h reachable from depth h , we have

$$\mathbb{P} \left[\min_a T_{s_h,a}(n) \geq \frac{C^2}{\sqrt{2}(R_{\max} + 2C)^2} n^{1/(2h)} \right] \geq (1 - \delta)^h$$

for

$$n \geq \left(\frac{2C_0(|\mathcal{A}| - 1)}{\delta} \right)^{2h}.$$

This implies

$$\mathbb{P} \left[\max_{s_h, a} \varepsilon_{s, a}(n) \leq \sqrt{\frac{cR_{\max}^2(R_{\max} + 2C)^2 \log(|\mathcal{S}||\mathcal{A}|/\delta)}{\sqrt{2}C^2(1 - \gamma)^2 n^{1/h}}} \right] \geq (1 - \delta)^h.$$

6: Total Sample Complexity.

For a policy to be $(\varepsilon_0, \delta_0)$ -robust, we need:

$$\frac{4R_{\max}}{(1 - \gamma)^2} \max_{s, a} \varepsilon_{s, a}(n) + \frac{C'}{\sqrt{n}} \leq \varepsilon_0 \quad \text{and} \quad |\mathcal{S}|(1 - (1 - \delta)^H) + \delta \leq \delta_0.$$

Substituting and solving for n while simplifying constant and dominated terms, we deduce that the sample complexity for the **PATSO** policy to be $(\varepsilon_0, \delta_0)$ -robust is of order

$$O \left(\left[\frac{R_{\max}^3}{\varepsilon_0(1 - \gamma)^3} \log \left(\frac{H|\mathcal{S}|^2|\mathcal{A}|}{\delta_0} \right) \right]^{2H} \right).$$

6: Additional Discretization Error for CATSO. For **CATSO**, there is an additional source of error due to the discretization of the return distribution into N atoms. This error arises because:

- (i) The categorical representation divides the range $[Q_{\min}, Q_{\max}]$ into N equal intervals.
- (ii) Any continuous return value must be approximated by the nearest atom.
- (iii) The maximum approximation error for any single return is half the distance between consecutive atoms.

With N atoms spanning a range of diameter $D = Q_{\max} - Q_{\min}$, the maximum discretization error is:

$$e_{\text{disc}} = \frac{D}{2(N - 1)} \approx \frac{D}{2N}$$

For discounted returns bounded in $[0, R_{\max}]$, the effective diameter is:

$$D = \frac{R_{\max}}{1 - \gamma}$$

Thus, the discretization error is:

$$e_{\text{disc}} = \frac{R_{\max}}{2N(1 - \gamma)}$$

When this error propagates through the Bellman operator, it leads to an additional value function error of:

$$e_{\text{disc}, V} = \frac{e_{\text{disc}}}{1 - \gamma} = \frac{R_{\max}}{2N(1 - \gamma)^2}$$

To ensure this discretization error is at most $\varepsilon/2$ (allowing the remaining $\varepsilon/2$ for statistical estimation error), we need:

$$\frac{R_{\max}}{2N(1 - \gamma)^2} \leq \frac{\varepsilon}{2}$$

Solving for N :

$$N \geq \frac{R_{\max}}{\varepsilon(1 - \gamma)^2}$$

Simplifying to big-O notation:

$$N = O\left(\frac{R_{\max}}{\varepsilon(1-\gamma)}\right).$$

Unlike CATSO, the PATSO algorithm does not suffer from discretization error because:

- (i) PATSO directly stores each distinct observed return as a separate particle.
- (ii) When a new return is observed, it either increments the weight of an existing identical particle or adds a new particle with that exact value.

This representation is mathematically identical to the empirical distribution:

$$W_p(Q_n^{\text{PATSO}}(s, a), Q_{\text{empirical}}(s, a)) = 0 \quad \forall n$$

Therefore, there is no approximation error in representing the empirical distribution, and PATSO only needs to account for the statistical estimation error (i.e., the difference between the empirical and true distributions). The particle-based representation provides an exact representation of the empirical distribution, making it particularly well-suited for capturing complex, multi-modal return distributions without introducing additional approximation errors. \square

F FULL EXPERIMENTAL DETAILS

All the experiments were done on 8 Intel Xeon Gold 6130 (Skylake), x86_64, 2.10GHz, 2 CPUs/n-ode, 16 cores/CPU. Whenever feasible, we opted for open-source implementations of algorithms and environments.

Parameters selection We search the number of atoms from $\{10, 20, \dots, 100\}$ and choose the results with best performances. We set the discount factor $\gamma = .99$. For UCT, we use the exploration constant $C = \sqrt{2} \times (R_{\max} - R_{\min})$.

Table 4: Experimental settings.

Aspect	Precise setting now documented
Game list	12 games: Breakout, Enduro, Frostbite, Atlantis, Seaquest, Phoenix, BankHeist, Assault, QBert, Asterix, Amidar, Freeway. Line-269 typo (17 \rightarrow 12) fixed.
Random seeds	10 seeds \times method \times game; we report mean $\pm 95\%$ CI.
Rollout budget	512 MCTS simulations per action, search depth 50.
Discount	$\gamma = 0.99$ for all planners.
Exploration constant (PUCT baseline)	$c = \sqrt{2} (R_{\max} - R_{\min})$.
Temperature grid (for ENT regularised baselines)	τ grid search in $(0, 1]$; best τ chosen per game.
Hyper-parameter grids (ours)	Atoms $N \in \{10, 20, 30, \dots, 100\}$; power- $p \in \{1, 2, 4\}$; bonus- $C \in \{4, 8, 16\}$.

Pre-trained value network (details formerly missing). We follow exactly the DQN recipe (Mnih et al., 2015):

- **Architecture:** 3 Conv (32×8 , 64×4 , 64×3) + 2 FC ($512 \rightarrow |A|$).
- **Training:** 10M ALE frames / game; ε -greedy $1 \rightarrow 0.1$ in 1M steps; replay 1M; RMSProp learning rate 2×10^{-4} .
- **Saved heads:** After training, for any state s we store $Q(s, a)$ and $V(s) = \max_a Q(s, a)$.

Node initialisation in search (as in Xiao et al., 2019).

Table 5: Initial logits used by each planner at node expansion.

Planner	$Q_{\text{init}}(s, a)$
MENTS / TENTS	$(Q(s, a) - V(s)) / \tau$
RENTS	$\log P_{\text{prior}}(a s) + (Q(s, a) - V(s)) / \tau$, with $P_{\text{prior}} = \text{softmax}(Q(s, \cdot))$
CATSO / PATSO	Same shifted logits (with τ tuned as above).

G LIMITATIONS

Computational Demands: The **CATSO** distributional Monte Carlo Tree Search (MCTS) faces challenges in managing computational demands while maintaining and updating probability distributions, leading to a slightly increased complexity.

Fixed precision: The **PATSO** set of particles can increase in size if the observed value are different. We prevent this in the implementation by fixing the float precision.

Number of atoms: Our approach’s performance is slightly influenced by hyperparameters, with the number of atoms being a critical factor. Suboptimal choices may affect performance.

H DETAILS FOR SECTION 4.4: MEMORY CAPPING AND W_1 CONTROL

Merge-on-insert (precise). We keep a sorted container of pairs $\{(Q_i, \alpha_i)\}_{i=1}^M$ with $M \leq K$. Upon a new sample z :

- (i) **Match:** If $\min_i |z - Q_i| \leq \varepsilon_{\text{tol}}$, let $i^* = \arg \min_i |z - Q_i|$ and set $\alpha_{i^*} \leftarrow \alpha_{i^*} + 1$.
- (ii) **Insert:** Else if $M < K$, insert $(z, 1)$ and keep the list sorted.
- (iii) **Merge:** Else (overflow): find j with smallest gap $|Q_{j+1} - Q_j|$, replace the pair by

$$(Q^{\text{new}}, \alpha^{\text{new}}) = \left(\frac{\alpha_j Q_j + \alpha_{j+1} Q_{j+1}}{\alpha_j + \alpha_{j+1}}, \alpha_j + \alpha_{j+1} \right),$$

then insert $(z, 1)$ in sorted order.

Bound on W_1 error under capping. Let \hat{P} be the (uncapped) empirical distribution and \tilde{P} the capped distribution after applying the procedure above. With nearest-neighbor merges,

$$W_1(\hat{P}, \tilde{P}) \leq \sum_{\text{merges}} (\alpha_j + \alpha_{j+1}) |Q_{j+1} - Q_j| = O\left(\frac{\text{range}}{K}\right).$$

Since the Bellman operator is $1/(1-\gamma)$ -Lipschitz in W_1 , the root-value error contributed by capping is $O(\text{range}/(K(1-\gamma)))$ and the simple-regret additive term is $O(\text{range}/(K(1-\gamma)^2))$, which is dominated by the proven $O(n^{-1/2})$ term for moderate K .