

Appendices

A STRUCTURE OF THE APPENDICES

This appendix is organized as follows:

- Section B: Mathematical notation used throughout the paper
- Section C: Supporting lemmas for the theoretical analysis
- Section D: Convergence proofs for PM-DA-MCTS in non-stationary bandits
- Section E: Experimental details and hyperparameter selection
- Section F: Exponential Dependence on d and Guarantees Under a Capped Implementation

B NOTATION

Table 6: Principal symbols for the non-stationary multi-armed bandit model used throughout the paper.

Symbol	Set / Type	Meaning
K	\mathbb{N}	Number of arms
$T_a(t)$	\mathbb{N}	Number of times arm a has been pulled up to (and including) round t
μ_a	\mathbb{R}	True (possibly time-varying) mean reward of arm a
a_\star	\mathcal{A}	Index of an optimal arm, i.e. $a_\star \in \arg \max_a \mu_a$
μ_\star	\mathbb{R}	Optimal mean reward, $\mu_\star = \max_a \mu_a$
$\hat{\mu}_{a,n}$	\mathbb{R}	Empirical mean of arm a after n pulls, i.e. $\frac{1}{n} \sum_{s=1}^n Y_{a,s}$
$\hat{\mu}_n(p)$	\mathbb{R}	Power-mean estimator at time n , $(\sum_{a=1}^K \frac{T_a(n)}{n} \hat{\mu}_{a,T_a(n)}^p)^{1/p}$, with $p \in [1, \infty)$
Δ_a	$\mathbb{R}_{\geq 0}$	Gap of arm a : $\Delta_a = \mu_\star - \mu_a$
$A_a(n)$	\mathbb{N}	Safety index for arm a (Eq. equation 13), threshold beyond which its bonus term is $\leq \Delta_a/2$
$U_{a,n,s}$	\mathbb{R}	Upper-confidence value used by the algorithm (see Lemma 5)
p	parameter	Power-mean exponent (backup operator)

C SUPPORTING LEMMAS

In this section, we will present all necessary supporting Lemmas for the main theoretical analysis.

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

Lemma 1 (Lemma 1 Dam et al. (2024)). *For $m \in [M]$, let $(\hat{V}_{m,n})_{n \geq 1}$ be a sequence of estimator satisfying $\hat{V}_{m,n} \xrightarrow{\alpha, \zeta} V_m$, and there exists a constant L such that $\hat{V}_{m,n} \leq L, \forall n \geq 1$. Let X_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*

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

Table 7: Key symbols used in the Monte-Carlo Tree-Search (MCTS) analysis.

Symbol	Set / Type	Meaning
\mathcal{T}	tree	Search tree maintained by the algorithm
s_h	node / state	State encountered at depth h ($h = 0$ at the root)
h	\mathbb{N}	Depth index (planning step)
H	\mathbb{N}	Planning horizon / maximum depth
\mathcal{A}_s	set	Feasible actions at state s
a	action	A concrete action $a \in \mathcal{A}_s$
$N_s(n)$	\mathbb{N}	Visit count of node s after n simulations
$N_{s,a}(n)$	\mathbb{N}	Visit count of edge (s, a) after n simulations
$\hat{Q}_n(s, a)$	\mathbb{R}	Empirical Q -value at (s, a) after n simulations
$\hat{V}_n(s)$	\mathbb{R}	Backup value of node s after n simulations (power mean)
$U_n(s, a)$	\mathbb{R}	Upper confidence bound used for selection (algorithm-specific)
γ	$(0, 1]$	Discount factor for future rewards
$\text{Rollout}(s)$	random variable	Return of one simulation rollout starting from state s

Then with $2\alpha \leq \zeta, \zeta > 1$,

$$\hat{Q}_n \xrightarrow{\alpha, \zeta} \mu + \sum_{m=1}^M p_m V_m.$$

Lemma 2 (Lemma 2 Dam et al. (2024)). *Let consider non-negative variables $x, y \in \mathbb{R}^+$, and a constant m that $0 \leq m \leq 1$. Then*

$$(x + y)^m \leq x^m + y^m. \quad (10)$$

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*

$$\left(\sum_i (|x_i + y_i|)^p \right)^{\frac{1}{p}} \leq \left(\sum_i (|x_i|)^p \right)^{\frac{1}{p}} + \left(\sum_i (|y_i|)^p \right)^{\frac{1}{p}} \quad (11)$$

Proof. This is a basic result. \square

D CONVERGENCE OF PM-DA-MCTS IN NON-STATIONARY MULTI-ARMED BANDITS

Every node of an MCTS tree forms its own non-stationary multi-armed bandit: the empirical reward of an action drifts as the parent’s selection strategy evolves over time. To analyse such drift formally we prove that the *power-mean* backup estimator at the root converges and concentrates.

Road-map of the proof. The main concentration result for the power-mean estimator is stated in Theorem 1. Its proof reduces to a series of increasingly specialised lemmas:

- Lemma 4 bounds the probability that the empirical mean of the *optimal* arm deviates from μ_* after $T_{a_*}(n)$ pulls.

- Lemma 5 provides single-step tail bounds; combined with it, Lemma 6 yields a high-probability bound on how often each sub-optimal arm is selected.
- Lemma 7 decomposes the power-mean error into an *optimal-arm* term and an aggregate *sub-optimal-arm* term. The two parts are then treated separately by Lemma 8 and Lemma 9.
- These auxiliary results are assembled in Lemma 10, which in turn implies Theorem 1.

Together, the lemmas ensure that the power-mean backup operator enjoys polynomial tail bounds even under the non-stationarity induced by tree search, laying the foundation for the global convergence guarantees that follow.

Lemma 4 (Optimal-arm deviation). *In the bandit model of Section 5.1 let*

$$A(n) := \left(\frac{2Cn^{\frac{b}{\zeta}}}{\Delta} \right)^{\frac{\zeta}{\alpha}}, \quad \Delta := \min_{a \neq a_*} (\mu_* - \mu_a),$$

and fix parameters with $R \geq \epsilon \geq n^{-\alpha/\zeta}$. Then

$$\Pr\left(|\hat{\mu}_{a_*, T_{a_*}(n)} - \mu_*| > \epsilon\right) \leq \sum_{a \neq a_*} \Pr(T_a(n) > A(n) + 1) + \frac{c}{\alpha - 1} \epsilon^{-\zeta} [n - (K - 1)A(n) + 1]^{-\alpha + 1}.$$

Proof. Define the event

$$\mathcal{E} := \left\{ \sum_{a \neq a_*} T_a(n) > (K - 1)[A(n) + 1] \right\}.$$

Split the probability as

$$\Pr(|\hat{\mu}_{a_*, T_{a_*}(n)} - \mu_*| > \epsilon) \leq \Pr(\mathcal{E}) + D_1, \quad (12)$$

where

$$D_1 := \Pr(\mathcal{E}^c; |\hat{\mu}_{a_*, T_{a_*}(n)} - \mu_*| \geq \epsilon).$$

Bounding D_1 . On \mathcal{E}^c we have $T_{a_*}(n) \geq n - (K - 1)[A(n) + 1]$. Hence

$$\begin{aligned} D_1 &\leq \sum_{t=n-(K-1)(A(n)+1)}^n \Pr(|\hat{\mu}_{a_*, t} - \mu_*| \geq \epsilon) \\ &\leq c \epsilon^{-\zeta} \sum_{t=n-(K-1)(A(n)+1)}^n t^{-\alpha} \\ &\leq c \epsilon^{-\zeta} \int_{n-(K-1)(A(n)+1)-1}^{\infty} t^{-\alpha} dt \\ &= \frac{c}{\alpha - 1} \epsilon^{-\zeta} [n - (K - 1)(A(n) + 1) - 1]^{-\alpha + 1}, \quad (\alpha > 2). \end{aligned}$$

Final bound. Substituting this estimate into equation 12 gives the claim. \square

We introduce the notation $U_{a,t,s} = \hat{\mu}_{a,s} + L\varepsilon_k^\beta + C \frac{t^{\frac{b}{\zeta}}}{s^{\frac{\alpha}{\zeta}}}$ and introduce that for all a the quantity

$$A_a(t) := \inf \left\{ s \leq t : C \frac{t^{\frac{b}{\zeta}}}{s^{\frac{\alpha}{\zeta}}} \leq \frac{\Delta_a}{2} \right\} = \left(\frac{2C}{\Delta_a} \right)^{\frac{\zeta}{\alpha}} t^{\frac{b}{\alpha}}, \quad (13)$$

where $\Delta_a = \mu_* - \mu_a$, the concentration properties permits to prove the following Lemma.

Lemma 5 (Basic concentration). *Consider the non-stationary K -armed bandit of Section 5.1. Assume each empirical mean obeys the tail condition $\hat{\mu}_{a,s} \xrightarrow{\alpha, \zeta} \mu_a$. Fix*

$$U_{a,n,s} := \hat{\mu}_{a,s} + L\varepsilon_k^\beta + C \frac{n^{\frac{b}{\zeta}}}{s^{\frac{\alpha}{\zeta}}}, \quad 1 \leq s \leq n,$$

where $L\varepsilon_k^\beta$ is defined at Section 4. Let $\Delta_a := \mu_* - \mu_a > 0$.

(i) **Lower tail.** For every $s \in \{1, \dots, n\}$,

$$\Pr(U_{a,n,s} < \mu_a) \leq c C^{-\zeta} n^{-b}.$$

(ii) **Upper tail beyond the safety index.** Define the safety index

$$A_a(n) := \left(\frac{2C}{\Delta_a}\right)^{\zeta/\alpha} n^{b/\alpha}.$$

If $s \geq A_a(n)$ and $L\varepsilon_k^\beta \leq \frac{\Delta_a}{2}$ (which always holds once k is large enough), then

$$\Pr(U_{a,n,s} > \mu_*) \leq c C^{-\zeta} n^{-b}.$$

Proof. (i) **Lower tail.** Because $L\varepsilon_k^\beta > 0$, the event $U_{a,n,s} < \mu_a$ implies $\hat{\mu}_{a,s} - \mu_a < -C n^{b/\zeta} s^{-\alpha/\zeta}$. Applying the concentration assumption directly gives

$$\Pr(U_{a,n,s} < \mu_a) \leq c C^{-\zeta} n^{-b},$$

as required.

(ii) **Upper tail.** For every $s \geq A_a(n)$ we have $C n^{b/\zeta} s^{-\alpha/\zeta} \leq \Delta_a/2$. Under the additional requirement $L\varepsilon_k^\beta \leq \Delta_a/2$ (the algorithm's tuning ensures this once the tree is sufficiently refined), we obtain

$$U_{a,n,s} > \mu_* \implies \hat{\mu}_{a,s} - \mu_a > \Delta_a - (L\varepsilon_k^\beta + C n^{b/\zeta} s^{-\alpha/\zeta}) \geq \frac{\Delta_a}{2}.$$

Hence

$$\Pr(U_{a,n,s} > \mu_*) \leq \Pr(\hat{\mu}_{a,s} - \mu_a > \frac{\Delta_a}{2}) \leq c C^{-\zeta} n^{-b},$$

again by the tail assumption with $\epsilon = \Delta_a/2$. \square

In turn, Lemma 6 permits us to prove the following crucial high-probability bound on the number of selection of each sub-optimal arm.

Lemma 6 (HP-bound on visits). *Fix the bandit model of Section 5.1 with some constant $b > 1$. For every sub-optimal arm a , set*

$$A_a(n) := \left(\frac{2C}{\Delta_a}\right)^{\frac{\zeta}{\alpha}} n^{\frac{b}{\alpha}}.$$

Then, for any threshold $u \geq A_a(n)$,

$$\Pr(T_a(n) \geq u) \leq \frac{2c C^{-\zeta}}{b-1} (u-1)^{-(b-1)}.$$

Proof. First, for a real number τ define

$$\mathcal{E}_1 := \{ U_{a,t,u} \leq \tau \text{ for all } t \in [u, n] \cap \mathbb{Z} \},$$

$$\mathcal{E}_2 := \{ U_{a^*, u+t_0, t_0} > \tau \text{ for all } t_0 \in [1, n-u] \cap \mathbb{Z} \}.$$

If both events hold, arm a cannot be chosen once it has accrued u plays; hence

$$\mathcal{E}_1 \cap \mathcal{E}_2 \implies T_a(n) \leq u.$$

Thus

$$\Pr(T_a(n) \geq u) \leq \Pr(\mathcal{E}_1^c) + \Pr(\mathcal{E}_2^c).$$

Second, Using the union bound and rewriting:

$$\Pr(T_a(n) \geq u) \leq \sum_{t=u}^n \Pr(U_{a,t,u} > \tau) + \sum_{t_0=1}^{n-u} \Pr(U_{a^*, u+t_0, t_0} \leq \tau).$$

Because $u \geq A_a(n)$, Lemma 5 yields

$$\Pr(U_{a,t,u} > \mu_*) \leq c C^{-\zeta} t^{-b}, \quad \Pr(U_{a^*, u+t_0, t_0} \leq \mu_*) \leq c C^{-\zeta} (u+t_0)^{-b}.$$

Hence

$$\sum_{t=u}^n cC^{-\zeta} t^{-b} \leq cC^{-\zeta} \int_{u-1}^{\infty} t^{-b} dt = \frac{cC^{-\zeta}}{b-1} (u-1)^{-(b-1)},$$

$$\sum_{t_0=1}^{n-u} cC^{-\zeta} (u+t_0)^{-b} \leq cC^{-\zeta} \int_{u-1}^{\infty} t^{-b} dt = \frac{cC^{-\zeta}}{b-1} (u-1)^{-(b-1)}.$$

Adding the two identical bounds gives the claimed result:

$$\Pr(T_a(n) \geq u) \leq \frac{2cC^{-\zeta}}{b-1} (u-1)^{-(b-1)}.$$

□

Lemma 7 (Lemma 7 Dam et al. (2024)). *Let us define the power mean estimator $\hat{\mu}_n(p)$ as $\hat{\mu}_n(p) = \left(\sum_{a=1}^K \frac{T_a(n)}{n} \hat{\mu}_{a,T_a(n)}^p \right)^{\frac{1}{p}}$. For any $p \geq 1$, we have*

$$|\hat{\mu}_n(p) - \mu_*| \leq R \sum_{a=1, a \neq a_*}^K \frac{T_a(n)}{n} + \left(\sum_{a=1}^K \frac{T_a(n)}{n} (|\hat{\mu}_{a,T_a(n)} - \mu_a|^p) \right)^{\frac{1}{p}} \quad (14)$$

Lemma 8 (Optimal-arm term). *Let the bandit model of Section 5.1 satisfy the tail assumption $\hat{\mu}_{a,n} \xrightarrow{\alpha, \zeta} \mu_a$ with $\alpha > 2$ and $b > 1$. Recall*

$$A(n) := \left(\frac{2cC^{b/\zeta}}{\Delta} \right)^{\zeta/\alpha}, \quad \Delta := \min_{a \neq a_*} (\mu_* - \mu_a),$$

and fix a scale $R \geq \epsilon \geq n^{-\alpha/\zeta}$. Then, for every $p \geq 1$,

$$\Pr\left(\frac{T_{a_*(n)}}{n} |\hat{\mu}_{a_*,T_{a_*(n)}} - \mu_*|^p > \epsilon^p\right) \leq \frac{2cC^{-\zeta}(K-1)}{b-1} A(n)^{-(b-1)} + \frac{c}{\alpha-1} \epsilon^{-\zeta} [n - (K-1)(A(n)+1) - 1]^{-\alpha+1}.$$

Proof. Reduce to a simpler deviation. Because $\frac{T_{a_*(n)}}{n} \leq 1$,

$$\Pr\left(\frac{T_{a_*(n)}}{n} |\hat{\mu}_{a_*,T_{a_*(n)}} - \mu_*|^p > \epsilon^p\right) \leq \Pr(|\hat{\mu}_{a_*,T_{a_*(n)}} - \mu_*| > \epsilon).$$

Apply Lemma 4. Lemma 4 gives, for the right-hand probability,

$$\Pr(|\hat{\mu}_{a_*,T_{a_*(n)}} - \mu_*| > \epsilon) \leq \underbrace{\sum_{a \neq a_*} \Pr(T_a(n) > A(n) + 1)}_{F_{11}} + \underbrace{\frac{c}{\alpha-1} \epsilon^{-\zeta} [n - (K-1)(A(n) + 1) - 1]^{-\alpha+1}}_{F_{12}}.$$

Bound the visit term F_{11} . Lemma 6 yields, for every sub-optimal arm $a \neq a_*$,

$$\Pr(T_a(n) > A(n) + 1) \leq \frac{2cC^{-\zeta}}{b-1} A(n)^{-(b-1)}.$$

Summing over the $K-1$ sub-optimal arms gives

$$F_{11} \leq \frac{2cC^{-\zeta}(K-1)}{b-1} A(n)^{-(b-1)}.$$

Combine the pieces. Adding F_{11} and F_{12} completes the claimed inequality. □

Lemma 9 (Upper bound for a sub-optimal arm). *Consider the K -armed bandit introduced in Section 5.1 and fix any sub-optimal arm a . Let $R \geq \epsilon \geq n^{-\frac{\alpha}{\zeta}}$ and denote*

$$A(n) := \left(\frac{2cC^{b/\zeta}}{\Delta_a} \right)^{\frac{\zeta}{\alpha}}.$$

There exists $N_0 \in \mathbb{N}$ such that, for every $n \geq N_0$,

$$\Pr\left(\frac{T_a(n)}{n} |\hat{\mu}_{a,T_a(n)} - \mu_a|^p > \frac{\epsilon^p}{K-1}\right) \leq G_1 + G_2,$$

where the two terms admit the following explicit bounds:

(i) (**Small** p) If $1 \leq p \leq 2$ and $\alpha \leq \frac{\zeta}{p}$,

$$G_1 + G_2 \leq \frac{2cC^{-\zeta}}{b-1} A(n)^{-(b-1)} + \frac{2c(K-1)^{\frac{\zeta}{p}}}{1 - (\alpha - \frac{\zeta}{p})} \epsilon^{-\zeta} [A(n) + 1]^{-(\alpha-1)}.$$

(ii) (**Moderate** p) If $p > 2$ and $0 < \alpha - \frac{\zeta}{p} < 1$,

$$G_1 + G_2 \leq \frac{2cC^{-\zeta}}{b-1} A(n)^{-(b-1)} + \frac{c(K-1)^{\frac{\zeta}{p}}}{1 - (\alpha - \frac{\zeta}{p})} \epsilon^{-\zeta} [A(n) + 1]^{-(\alpha-1)}.$$

(iii) (**Large** p) If $p > 2$ and $\alpha - \frac{\zeta}{p} > 1$,

$$G_1 + G_2 \leq \frac{2cC^{-\zeta}}{b-1} A(n)^{-(b-1)} + \frac{c(K-1)^{\frac{\zeta}{p}} (\alpha - \frac{\zeta}{p})}{(\alpha - \frac{\zeta}{p}) - 1} \epsilon^{-\zeta} [A(n) + 1]^{-\frac{\zeta}{p}}.$$

Proof. From Lemma 6, for every $u > A_a(n) = A(n)$ we have

$$\Pr(T_a(n) > u) \leq \frac{2cC^{-\zeta}}{b-1} (u-1)^{-(b-1)}.$$

Define the events

$$\mathcal{E}_1 := \{T_a(n) > A(n) + 1\}, \quad \mathcal{E}_1^c := \{T_a(n) \leq A(n) + 1\}.$$

Using the law of total probability,

$$\Pr\left(\frac{T_a(n)}{n} \mid \hat{\mu}_{a,T_a(n)} - \mu_a \mid^p > \frac{\epsilon^p}{K-1}\right) \leq \underbrace{\Pr(\mathcal{E}_1)}_{=:G_1} + \underbrace{\sum_{t=1}^{A(n)+1} \Pr\left(\frac{t}{n} \mid \hat{\mu}_{a,t} - \mu_a \mid^p > \frac{\epsilon^p}{K-1}\right)}_{=:G_2}.$$

Bounding G_1 . Directly from the visit-probability bound,

$$G_1 = \Pr(T_a(n) > A(n) + 1) \leq \frac{2cC^{-\zeta}}{b-1} A(n)^{-(b-1)}.$$

Bounding G_2 . For all $t \leq A(n) + 1$ and $n \geq N_0$, $(\frac{n}{t(K-1)})^{1/p} \epsilon \geq \epsilon \geq n^{-\frac{\alpha}{p}}$. Applying the tail assumption on the empirical-mean error (with constant c):

$$\begin{aligned} G_2 &\leq \sum_{t=1}^{A(n)+1} c t^{-\alpha} \left[\left(\frac{n}{t(K-1)} \right)^{1/p} \epsilon \right]^{-\zeta} \\ &= c(K-1)^{\frac{\zeta}{p}} \epsilon^{-\zeta} n^{-\frac{\zeta}{p}} \sum_{t=1}^{A(n)+1} t^{-(\alpha - \frac{\zeta}{p})}. \end{aligned}$$

The remaining sum is handled in three regimes, depending on the sign of $\alpha - \frac{\zeta}{p}$:

Case (i) $\alpha - \frac{\zeta}{p} \leq 0$. Using an integral comparison,

$$\sum_{t=1}^{A(n)+1} t^{-(\alpha - \frac{\zeta}{p})} \leq \frac{2[A(n) + 1]^{1 - (\alpha - \frac{\zeta}{p})}}{1 - (\alpha - \frac{\zeta}{p})},$$

which yields the bound claimed in part (i).

Case (ii) $0 < \alpha - \frac{\zeta}{p} < 1$. Evaluating the geometric sum,

$$\sum_{t=1}^{A(n)+1} t^{-(\alpha - \frac{\zeta}{p})} = \frac{\alpha - \frac{\zeta}{p}}{(\alpha - \frac{\zeta}{p}) - 1} - \frac{[A(n) + 1]^{1 - (\alpha - \frac{\zeta}{p})}}{(\alpha - \frac{\zeta}{p}) - 1},$$

which leads to the expression in part (ii).

Case (iii) $\alpha - \frac{\zeta}{p} > 1$. The same calculation gives

$$\sum_{t=1}^{A(n)+1} t^{-(\alpha - \frac{\zeta}{p})} \leq \frac{\alpha - \frac{\zeta}{p}}{(\alpha - \frac{\zeta}{p}) - 1},$$

hence the bound in part (iii).

Combining the estimates for G_1 and G_2 completes the proof. \square

Lemma 10 (Deviation of the p -power mean). *For the K -armed bandit of Section 5.1 let*

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

be the empirical power-mean estimator. Write $\Delta_a := \mu_ - \mu_a$ and $\Delta := \min_{a \in [K]} \Delta_a$. Define*

$$A(n) := \left(\frac{2C n^{b/\zeta}}{\Delta} \right)^{\zeta/\alpha}, \quad \epsilon_0 := \frac{2^{1/p} n \epsilon}{x} + 2^{1/p} R(K-1)[3 + A(n)],$$

where the parameters satisfy

$$R \geq \epsilon \geq n^{-\alpha/\zeta}, \quad x \geq 1.$$

There exists $N_p \in \mathbb{N}$ such that for all $n \geq N_p$

$$\Pr\left(|\hat{\mu}_n(p) - \mu_*| \geq \frac{\epsilon_0 x}{n}\right) \leq \frac{8c C^{-\zeta} K R^\zeta \epsilon^{-\zeta} A(n)^{-(b-1)}}{b-1} + \frac{2c C^{-\zeta} (K-1)}{b-1} [2^{1/p} (3 + A(n)) x - 1]^{-(b-1)}.$$

Proof. **A deterministic decomposition.** Lemma 7 gives

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

Because $\frac{\epsilon_0 x}{n} = 2^{1/p} \epsilon + 2^{1/p} R(K-1)[3 + A(n)] \frac{x}{n}$, the probability event splits as

$$\begin{aligned} \Pr\left(|\hat{\mu}_n(p) - \mu_*| > \frac{\epsilon_0 x}{n}\right) &\leq \underbrace{\Pr\left(R \sum_{a \neq a_*} \frac{T_a(n)}{n} > R(K-1) \frac{2^{1/p} [3 + A(n)] x}{n}\right)}_{=: H_1} \\ &\quad + \underbrace{\Pr\left(\left(\sum_{a=1}^K \frac{T_a(n)}{n} |\hat{\mu}_{a, T_a(n)} - \mu_a|^p\right)^{1/p} \geq 2^{1/p} \epsilon\right)}_{=: H_2}. \end{aligned}$$

Bounding H_1 (large visit counts). For $x \geq 1$,

$$H_1 \leq \sum_{a \neq a_*} \Pr(T_a(n) > 2^{1/p} [3 + A(n)] x) \stackrel{\text{Lemma 6}}{\leq} \frac{2c C^{-\zeta} (K-1)}{b-1} [2^{1/p} (3 + A(n)) x - 1]^{-(b-1)}.$$

Bounding H_2 (estimation errors). Using Markov's inequality,

$$H_2 \leq F_1 + F_2, \quad \text{where} \quad \begin{cases} F_1 := \Pr\left(\frac{T_{a_*}(n)}{n} |\hat{\mu}_{a_*, T_{a_*}(n)} - \mu_{a_*}|^p > \epsilon^p\right), \\ F_2 := \sum_{a \neq a_*} \Pr\left(\frac{T_a(n)}{n} |\hat{\mu}_{a, T_a(n)} - \mu_a|^p > \frac{\epsilon^p}{K-1}\right). \end{cases}$$

(i) *Optimal arm.* Lemma 8 yields

$$F_1 \leq \frac{2c C^{-\zeta} (K-1)}{b-1} A(n)^{-(b-1)} + \frac{c}{\alpha-1} \epsilon^{-\zeta} [n - (K-1)(A(n)+1) - 1]^{-\alpha+1}.$$

(ii) *Sub-optimal arms.* Applying Lemma 9 in the two admissible parameter regimes (small p or moderate p),

$$F_2 \leq \frac{2c C^{-\zeta} (K-1)}{b-1} A(n)^{-(b-1)} + \frac{c(K-1)^{\frac{\zeta}{p}+1}}{1 - (\alpha - \frac{\zeta}{p})} \epsilon^{-\zeta} (A(n)+1)^{-(\alpha-1)}.$$

Combining F_1 and F_2 , and using $b-1 < \alpha-1$ together with $\epsilon \geq n^{-\alpha/\zeta}$, we can fix N_p large enough so that

$$H_2 \leq \frac{8c C^{-\zeta} K R^\zeta \epsilon^{-\zeta} A(n)^{-(b-1)}}{b-1}.$$

Adding the bounds for H_1 and H_2 completes the proof:

$$\Pr\left(|\hat{\mu}_n(p) - \mu_*| \geq \frac{\epsilon_0 x}{n}\right) \leq H_1 + H_2 \leq \frac{8c C^{-\zeta} K R^\zeta \epsilon^{-\zeta} A(n)^{-(b-1)}}{b-1} + \frac{2c C^{-\zeta} (K-1)}{b-1} [2^{1/p} (3+A(n))x - 1]^{-(b-1)}.$$

□

Theorem 1 (Power-mean concentration in non-stationary bandits). *Let $\{\mu_a\}_{a=1}^K$ be the (possibly drifting) arm means of the non-stationary bandit in Section 5.1. Assume each arm-estimator $\hat{\mu}_{a,n}$ satisfies the tail condition $\hat{\mu}_{a,n} \xrightarrow{\alpha, \zeta} \mu_a$ and that the algorithm uses the polynomial exploration bonus with parameters from Table 1.*

Define the empirical p -power mean

$$\hat{\mu}_n(p) := \left(\sum_{a=1}^K \frac{T_a(n)}{n} \hat{\mu}_{a, T_a(n)}^p \right)^{1/p}, \quad \mu_* = \max_a \mu_a.$$

Then $\hat{\mu}_n(p)$ concentrates around μ_* with exponents

$$\alpha' = (b-1) \left(1 - \frac{b}{\alpha}\right), \quad \zeta' = b-1,$$

i.e. there exist $c' > 0$ and N_0 such that for all $n \geq N_0$ and all $\epsilon \geq n^{-\alpha'/\zeta'}$,

$$\Pr\left(|\hat{\mu}_n(p) - \mu_*| \geq \epsilon\right) \leq c' n^{-\alpha'} \epsilon^{-\zeta'}.$$

Proof. Set $\Delta_a = \mu_* - \mu_a$ and $\Delta = \min_a \Delta_a$. For brevity let $A(n) := \left(\frac{2C n^{b/\zeta}}{\Delta}\right)^{\zeta/\alpha}$.

Parameterisation of the deviation scale. Fix $x \geq 1$, $R \geq \epsilon \geq n^{-\alpha/\zeta}$, and put

$$\epsilon_0 := \frac{2^{1/p} n \epsilon'}{x} + \frac{n R (K-1)}{x} 2^{1/p} [3 + A(n)], \quad \epsilon' := n^{-\alpha/\zeta} x.$$

With these choices

$$\frac{\epsilon_0 x}{n} = 2^{1/p} \epsilon' + 2^{1/p} R (K-1) \frac{[3+A(n)]x}{n}.$$

Decomposition. Lemma 10 implies, for $n \geq N_p$,

$$\begin{aligned} D &:= \Pr(|\hat{\mu}_n(p) - \mu_\star| \geq \epsilon) \leq \Pr(|\hat{\mu}_n(p) - \mu_\star| \geq \frac{\epsilon_0 x}{n}) \\ &\leq \frac{8cC^{-\zeta} K R^\zeta \epsilon^{-\zeta} A(n)^{-(b-1)}}{b-1} + \frac{2cC^{-\zeta} (K-1)}{b-1} [2^{1/p}(3 + A(n))x - 1]^{-(b-1)}. \end{aligned} \quad (15)$$

Replace the x -term. Because $x \geq 1$ and $A(n) \geq 3$ for sufficiently large n , $2^{1/p}(3 + A(n))x - 1 \geq A(n)x$. Using $A(n)x = \frac{n\epsilon}{2^{1/p}R(2K-1)}$ (from the definition of ϵ below) we substitute $A(n)$ and rearrange each term of equation 15; tedious but direct algebra yields

$$D \leq c_0 n^{-\alpha'} \epsilon^{-\zeta'},$$

with the constants

$$\alpha' = (b-1)\left(1 - \frac{b}{\alpha}\right), \quad \zeta' = b-1, \quad c_0 = \frac{2^{b+\zeta/p} c C^{-\zeta} K (2K-1)^\zeta R^{2\zeta}}{(b-1)} \left(\frac{2C}{\Delta}\right)^{-\frac{\zeta}{\alpha}(b-1-\zeta)}.$$

Extension to all n . For $n < N_0$ the bound holds trivially because $|\hat{\mu}_n(p) - \mu_\star| \leq R$ while $\epsilon \leq R$. Choosing the same c_0 for all n completes the probability statement.

Convergence in expectation. Since $\zeta' = b-1 > 1$ (recall $b > 2$), we have

$$|\mathbb{E} \hat{\mu}_n(p) - \mu_\star| \leq \int_0^\infty \Pr(|\hat{\mu}_n(p) - \mu_\star| \geq s) ds \xrightarrow{n \rightarrow \infty} 0,$$

establishing L^1 -convergence and finishing the proof. \square

Theorem 2 (Concentration along the search tree). *Run Stochastic-Power-UCT with depth-dependent constants $\{b_h\}_{h=0}^H$, $\{\alpha_h\}_{h=0}^H$, $\{\zeta_h\}_{h=0}^H$ satisfying Table 1. Then:*

(i) **Value nodes.** For every state s_h at depth $h \in \{0, \dots, H\}$,

$$\hat{V}_n(s_h) \xrightarrow{\alpha_h, \zeta_h} \tilde{V}(s_h).$$

(ii) **Action nodes.** For every state s_h with $h \in \{0, \dots, H-1\}$ and each $a \in \mathcal{A}_{s_h}$,

$$\hat{Q}_n(s_h, a) \xrightarrow{\alpha_{h+1}, \zeta_{h+1}} \tilde{Q}(s_h, a).$$

Proof. We proceed by induction on the tree depth H .

Base case $H = 1$. The root is s_0 and the only children are the leaf states $s_1 \sim \mathcal{P}(\cdot | s_0, a)$. Because every rollout value $\hat{V}_n(s_1) = \frac{1}{n} \sum_{t=1}^n \text{Rollout}_t(s_1)$ is an average of i.i.d. returns, Lemma 1 gives

$$\hat{V}_n(s_1) \xrightarrow{\alpha_1, \zeta_1} \tilde{V}(s_1).$$

Using the empirical Q -definition

$$\hat{Q}_n(s_0, a) = \frac{1}{n} \sum_{t=1}^n \left[r^{(t)}(s_0, a) + \gamma \hat{V}_{T_{s_0, a}^{s_1}(t)}(s_1) \right],$$

Lemma 1 yields

$$\hat{Q}_n(s_0, a) \xrightarrow{\alpha_1, \zeta_1} \tilde{Q}(s_0, a).$$

Finally, apply Theorem 1 to the value backup $\hat{V}_n(s_0) = \left(\sum_a \frac{T_{s_0, a}(n)}{n} \hat{Q}_{T_{s_0, a}(n)}^p(s_0, a) \right)^{1/p}$ to obtain $\hat{V}_n(s_0) \xrightarrow{\alpha_0, \zeta_0} \tilde{V}(s_0)$. Thus both claims hold when $H = 1$.

Induction step. Assume the theorem is valid for all trees of depth $H - 1$. Consider a tree of depth H .

Inside the depth- $(H - 1)$ subtrees. For any child state s_1 of the root s_0 , apply the induction hypothesis to the subtree rooted at s_1 :

$$\widehat{V}_n(s_h) \xrightarrow{\alpha_h, \zeta_h} \widetilde{V}(s_h), \quad h = 1, \dots, H, \quad (16)$$

$$\widehat{Q}_n(s_h, a) \xrightarrow{\alpha_{h+1}, \zeta_{h+1}} \widetilde{Q}(s_h, a), \quad h = 1, \dots, H - 1. \quad (17)$$

Back at the root. Using equation 16 in the same way as for the base case, Lemma 1 gives

$$\widehat{Q}_n(s_0, a) \xrightarrow{\alpha_1, \zeta_1} \widetilde{Q}(s_0, a), \quad a \in \mathcal{A}_{s_0}.$$

With these Q -concentration results and the power-mean backup, Theorem 1 again yields

$$\widehat{V}_n(s_0) \xrightarrow{\alpha_0, \zeta_0} \widetilde{V}(s_0).$$

Conclusion. Equations equation 16 plus the new bound on $\widehat{V}_n(s_0)$ establish part (i); eqs. equation 17 plus the new bound on $\widehat{Q}_n(s_0, a)$ establish part (ii). By induction, the theorem holds for every depth H . \square

Theorem 3 (Convergence of Expected Payoff). *We have at the root node s_0 , with the best possible parameter tuning that*

$$|\mathbb{E}[\widehat{V}_n(s_0)] - \widetilde{V}(s_0)| \leq \mathcal{O}(n^{-1/2}).$$

Proof. Using the convexity of $f(x) = |x|$ and applying Jensen's inequality we have

$$\begin{aligned} |\mathbb{E}[\widehat{V}_n(s_0)] - \widetilde{V}(s_0)| &\leq \mathbb{E}[|\widehat{V}_n(s_0) - \widetilde{V}(s_0)|] \\ &= \int_0^{+\infty} \mathbb{P}\left(|\widehat{V}_n(s_0) - \widetilde{V}(s_0)| \geq s\right) ds \\ &\leq \int_0^{n^{-\frac{\alpha_0}{\zeta_0}}} 1 ds + \int_{n^{-\frac{\alpha_0}{\zeta_0}}}^{+\infty} c_0 n^{-\alpha_0} s^{-\zeta_0} ds \\ &\leq n^{-\frac{\alpha_0}{\zeta_0}} + c_0 n^{-\alpha_0} \left(\frac{s^{-\zeta_0+1}}{-\zeta_0+1} \right) \Big|_{n^{-\frac{\alpha_0}{\zeta_0}}}^{+\infty} \\ &= \left(\frac{c_0}{\zeta_0 - 1} + 1 \right) n^{-\frac{\alpha_0}{\zeta_0}}. \end{aligned}$$

Because $\frac{\alpha_0}{\zeta_0} \leq \frac{1}{2}$ (Theorem 1), then the best possible rate we can estimate is

$$|\mathbb{E}[\widehat{V}_n(s_0)] - \widetilde{V}(s_0)| \leq \mathcal{O}(n^{-1/2}).$$

That concludes the proof. \square

Theorem 4 (Sample Complexity with Polynomial-tail Estimates). *PM-DA-MCTS returns an ε -optimal action at s_0 with probability at least $1 - \delta$ after at most*

$$N(\varepsilon, \delta) \leq C_1 \sigma^2 L^{\frac{d}{\beta}} \varepsilon^{-(\frac{d}{\beta}+2)} \log \left(\frac{C_2 L^{\frac{d}{\beta}}}{\varepsilon^{\frac{d}{\beta}} \delta} \right),$$

where $C_1, C_2 > 0$ depend only on d, β and the refinement/selection constants (but not on $\varepsilon, \delta, L, \sigma$).

Proof. Nets and discretization schedule. For each level $k = 0, 1, 2, \dots$, let $\mathcal{N}_k \subset \mathcal{A}$ be an ε_k -net of \mathcal{A} with respect to $\|\cdot\|_2$, with a geometric schedule $\varepsilon_{k+1} = \varepsilon_k/2$. There exists a constant $C_{\mathcal{A}} = C_{\mathcal{A}}(d)$ such that the net size satisfies

$$M_k \triangleq |\mathcal{N}_k| \leq C_{\mathcal{A}} \varepsilon_k^{-d} \quad \text{for all } k. \quad (18)$$

Let $a^* \in \arg \max_{a \in \mathcal{A}} Q^*(s_0, a)$ denote an optimal action at the root. By the definition of ε_k -nets and Hölder continuity,

$$\max_{a \in \mathcal{N}_k} Q^*(s_0, a) \geq Q^*(s_0, a^*) - L \varepsilon_k^\beta. \quad (19)$$

Choose k^* as the smallest level such that

$$L \varepsilon_{k^*}^\beta \leq \varepsilon/2 < L \varepsilon_{k^*-1}^\beta. \quad (20)$$

Then $\varepsilon_{k^*} \asymp (\varepsilon/L)^{1/\beta}$ and hence, by equation 18, $M_{k^*} \lesssim (L/\varepsilon)^{d/\beta}$.

Median-of-Means (MOM) estimator and deviation bound. Fix any action $a \in \mathcal{A}$. Let $X_i(a)$ be i.i.d. single-simulation returns with $\mathbb{E}[X_i(a)] = Q^*(s_0, a)$ and $\text{Var}(X_i(a)) \leq \sigma^2$. Given a total of n samples at (s_0, a) , partition them into B blocks of equal size $m = \lfloor n/B \rfloor$, form block means $\bar{X}_1, \dots, \bar{X}_B$, and define the MOM estimator

$$\hat{Q}_n^{\text{MOM}}(s_0, a) \triangleq \text{median}\{\bar{X}_1, \dots, \bar{X}_B\}.$$

We will use the following standard claim.

Claim (MOM deviation). There exist absolute constants $c_0, c_1, c_2 > 0$ such that for any confidence $\delta \in (0, 1)$, if $B \geq c_0 \log(2/\delta)$ and $m \geq c_1 \sigma^2/t^2$, then

$$\Pr \left\{ |\hat{Q}_n^{\text{MOM}}(s_0, a) - Q^*(s_0, a)| > t \right\} \leq \delta \quad \text{for } n = Bm. \quad (21)$$

Proof of the claim. For each block mean \bar{X}_j , by Chebyshev's inequality, $\Pr\{|\bar{X}_j - Q^*(s_0, a)| > t\} \leq \sigma^2/(mt^2)$. If $m \geq c_1 \sigma^2/t^2$ with $c_1 \geq 4$, then the failure probability per block is at most $1/4$, hence each block is “good” with probability at least $3/4$ independently. By a Chernoff bound, the probability that fewer than $B/2$ blocks are good is at most $\exp(-c_2 B)$ for a universal $c_2 > 0$. If $B \geq c_0 \log(2/\delta)$ with $c_0 \geq 1/c_2$, then $\exp(-c_2 B) \leq \delta/2$. Conditioned on the event that at least $B/2$ blocks are good, the median of the B block means must lie within t of $Q^*(s_0, a)$. Combining the two events gives equation 21. \square

Setting $t = \alpha L \varepsilon_k^\beta$ with a constant $\alpha \in (0, 1)$ to be fixed, the claim shows that if

$$B_k \geq c_0 \log \left(\frac{2}{\delta_{k,a}} \right), \quad m_k \geq \frac{c_1 \sigma^2}{\alpha^2 L^2 \varepsilon_k^{2\beta}}, \quad n_k = B_k m_k, \quad (22)$$

then

$$\Pr \left\{ |\hat{Q}_{n_k}^{\text{MOM}}(s_0, a) - Q^*(s_0, a)| \leq \alpha L \varepsilon_k^\beta \right\} \geq 1 - \delta_{k,a}. \quad (23)$$

Uniform confidence over all actions and levels up to k^* . We will visit only actions in $\bigcup_{k \leq k^*} \mathcal{N}_k$ before termination (since levels are refined monotonically down to ε_{k^*} by construction). Set confidence budgets

$$\delta_{k,a} = \frac{\delta}{4(k^*+1)M_k} \quad \text{for each } a \in \mathcal{N}_k, \quad k = 0, 1, \dots, k^*. \quad (24)$$

By a union bound across all $k \leq k^*$ and $a \in \mathcal{N}_k$, the event

$$\mathcal{E} \triangleq \bigcap_{k=0}^{k^*} \bigcap_{a \in \mathcal{N}_k} \left\{ |\hat{Q}_{n_k}^{\text{MOM}}(s_0, a) - Q^*(s_0, a)| \leq \alpha L \varepsilon_k^\beta \right\}$$

satisfies $\Pr(\mathcal{E}) \geq 1 - \delta/4$ provided n_k obey equation 22 with $\delta_{k,a}$ as in equation 24.

Using equation 22 and equation 24, there exist constants $C_b, C_m > 0$ such that

$$B_k \leq C_b \log \left(\frac{(k^*+1)M_k}{\delta} \right), \quad m_k \leq C_m \frac{\sigma^2}{L^2 \varepsilon_k^{2\beta}}. \quad (25)$$

Hence

$$n_k = B_k m_k \leq C \frac{\sigma^2}{L^2 \varepsilon_k^{2\beta}} \log \left(\frac{(k^*+1)M_k}{\delta} \right), \quad (26)$$

for a constant $C = C_b C_m$ depending only on the universal constants from Step 1.

Alignment with the selection/refinement rule. The selection rule (Eq. equation 5) uses a polynomially decaying exploration bonus $b_h N_{s,a}(t)^{-\zeta_h}$ together with a fixed discretization bias term $L \varepsilon_k^\beta$. Thus the algorithm *continues sampling* an action $a \in \mathcal{N}_k(s)$ at level k while

$$b_h N_{s,a}(t)^{-\zeta_h} > \alpha L \varepsilon_k^\beta,$$

and it *refines* to level $k+1$ only once, for all currently active actions in $\mathcal{N}_k(s)$,

$$N_{s,a}(t) \geq n_k^{(\text{poly})} \triangleq \left(\frac{b_h}{\alpha L \varepsilon_k^\beta} \right)^{\frac{1}{\zeta_h}}. \quad (27)$$

In addition, to guarantee high-probability estimation accuracy for the value estimator we require a per-action lower bound

$$N_{s,a}(t) \geq n_k^{(\text{est})} \triangleq C \frac{\sigma^2}{L^2 \varepsilon_k^{2\beta}} \log\left(\frac{(k^*+1) M_k}{\delta}\right), \quad (28)$$

where $M_k = |\mathcal{N}_k(s)|$ and $C > 0$ is the constant from the estimator's deviation inequality. We therefore set

$$n_k \triangleq \max\{n_k^{(\text{poly})}, n_k^{(\text{est})}\}. \quad (29)$$

Under the uniform event \mathcal{E} (all per-action concentration events hold), the bonus is monotone in $N_{s,a}$ and any under-sampled action remains optimistic; hence before leaving level k each active action has been sampled at most n_k times (up to a constant factor absorbed into C).

Correctness at the stopping level k^* . On the event \mathcal{E} , when the algorithm reaches level k^* and each active action has at least n_{k^*} samples, for every $a \in \mathcal{N}_{k^*}$ we have

$$|\hat{Q}_{n_{k^*}}^{\text{MOM}}(s_0, a) - Q^*(s_0, a)| \leq \alpha L \varepsilon_{k^*}^\beta.$$

Let $\hat{a} \in \arg \max_{a \in \mathcal{N}_{k^*}} \hat{Q}_{n_{k^*}}^{\text{MOM}}(s_0, a)$ be the selected action. Let $a_{k^*}^* \in \arg \max_{a \in \mathcal{N}_{k^*}} Q^*(s_0, a)$ denote a net-optimal action at level k^* . Then

$$\begin{aligned} Q^*(s_0, a^*) - Q^*(s_0, \hat{a}) &\leq \underbrace{[Q^*(s_0, a^*) - Q^*(s_0, a_{k^*}^*)]}_{\leq L \varepsilon_{k^*}^\beta \text{ by equation 19}} + \underbrace{[Q^*(s_0, a_{k^*}^*) - \hat{Q}_{n_{k^*}}^{\text{MOM}}(s_0, a_{k^*}^*)]}_{\leq \alpha L \varepsilon_{k^*}^\beta} \\ &\quad + \underbrace{[\hat{Q}_{n_{k^*}}^{\text{MOM}}(s_0, a_{k^*}^*) - \hat{Q}_{n_{k^*}}^{\text{MOM}}(s_0, \hat{a})]}_{\leq 0 \text{ by definition of } \hat{a}} \\ &\quad + \underbrace{[\hat{Q}_{n_{k^*}}^{\text{MOM}}(s_0, \hat{a}) - Q^*(s_0, \hat{a})]}_{\leq \alpha L \varepsilon_{k^*}^\beta}. \end{aligned}$$

Choosing $\alpha = 1/4$ and recalling equation 20 gives

$$Q^*(s_0, a^*) - Q^*(s_0, \hat{a}) \leq \left(1 + \frac{1}{4} + \frac{1}{4}\right) L \varepsilon_{k^*}^\beta \leq \frac{3}{2} \cdot \frac{\varepsilon}{2} < \varepsilon.$$

Thus \hat{a} is ε -optimal at the root on the event \mathcal{E} .

Sample complexity bound. Let N_k denote the total number of samples taken at level k across all actions. By Step 3 and equation 26,

$$N_k \leq M_k n_k \leq C M_k \frac{\sigma^2}{L^2 \varepsilon_k^{2\beta}} \log\left(\frac{(k^*+1) M_k}{\delta}\right).$$

Using equation 18 and the geometric schedule $\varepsilon_k = \varepsilon_0 2^{-k}$, the sum over $k = 0$ to k^* is dominated by the finest level:

$$\sum_{k=0}^{k^*} N_k \leq C' \sigma^2 \sum_{k=0}^{k^*} \varepsilon_k^{-(d+2\beta)} \log\left(\frac{(k^*+1) C_A \varepsilon_k^{-d}}{\delta}\right) \leq C'' \sigma^2 \varepsilon_{k^*}^{-(d+2\beta)} \log\left(\frac{C''' \varepsilon_{k^*}^{-d}}{\delta}\right).$$

Finally, by equation 20, $\varepsilon_{k^*} \asymp (\varepsilon/L)^{1/\beta}$, hence

$$\varepsilon_{k^*}^{-(d+2\beta)} = \left((\varepsilon/L)^{1/\beta}\right)^{-(d+2\beta)} = L^{\frac{d}{\beta}} \varepsilon^{-(\frac{d}{\beta}+2)}, \quad \log\left(\frac{C''' \varepsilon_{k^*}^{-d}}{\delta}\right) = \log\left(\frac{C_2 L^{\frac{d}{\beta}}}{\varepsilon^{\frac{d}{\beta}} \delta}\right).$$

This yields

$$N(\varepsilon, \delta) \leq C_1 \sigma^2 L^{\frac{d}{\beta}} \varepsilon^{-(\frac{d}{\beta}+2)} \log\left(\frac{C_2 L^{\frac{d}{\beta}}}{\varepsilon^{\frac{d}{\beta}} \delta}\right),$$

as claimed.

Total failure probability. We allocated $\delta/4$ to the uniform estimation event \mathcal{E} . The remaining probability budget covers (i) the high-probability coupling used to argue Step 3 (choice of exploration constants to prevent premature refinement) and (ii) standard stopping events. Choosing constants so that each of these has probability at least $1 - \delta/4$ yields an overall success probability at least $1 - \delta$ by a union bound. \square

E EXPERIMENTAL SETUP AND HYPERPARAMETER SELECTION

E.1 ENVIRONMENT CONFIGURATION

Our experimental evaluation encompasses diverse continuous control environments with configurable stochasticity to assess PM-DA-MCTS performance across varying action dimensionalities and noise levels. The test suite includes: Continuous CartPole, Stochastic Pendulum, Stochastic Mountain Car Continuous, and Stochastic Continuous Acrobot and MuJoCo-style Locomotion Improved Hopper.

All environments incorporate configurable stochasticity through three noise components:

$$\text{action_noise_scale} \in [0.02, 0.05] \tag{30}$$

$$\text{dynamics_noise_scale} \in [0.01, 0.7] \tag{31}$$

$$\text{obs_noise_scale} \in [0.0, 0.01] \tag{32}$$

E.2 ALGORITHM CONFIGURATION

PM-DA-MCTS Parameters. We systematically evaluate power mean exponents $p \in \{1.5, 2.0, 3.0, 4.0, 5.0\}$ and find optimal performance with $p = 2.0$ across most environments, providing the best balance between exploration and exploitation. The discretization parameters are set as $\varepsilon_1 = 0.5$ and $\beta = 1.0$ (assuming unit Hölder continuity). We set the exploration constant $C = 30$.

For the polynomial exploration bonus, we use algorithmic constants satisfying the conditions in Table 1 with the optimal configuration:

$$\frac{\alpha_h}{\zeta_h} = \frac{1}{2} \tag{33}$$

$$\frac{b_h}{\zeta_h} = \frac{1}{4} \tag{34}$$

yielding the exploration bonus $C \frac{N_s(n)^{1/4}}{N_{s,\alpha}(n)^{1/2}}$.

Baseline Method Configuration. We compare against several state-of-the-art continuous action planning methods:

- **UCT with Discretization:** Fixed uniform grid discretization with $C = 2.0$.
- **Progressive Widening (PW):** Dimension-adaptive expansion using $K(n) = \lfloor \alpha n^\beta \rfloor$ where $\alpha = 0.5$ for $d \leq 3$, $\alpha = 0.3$ for $d > 3$, and $\beta = 0.5$

- **HOOT (HOO over Trees):** Dimension-adaptive HOO parameters:

$$\rho = \begin{cases} 2^{-2/d} & \text{if } d = 1 \\ 2^{-1.5/d} & \text{if } d \leq 3 \\ 2^{-1/d} & \text{if } d \leq 8 \\ 2^{-0.7/d} & \text{if } d > 8 \end{cases} \quad (35)$$

$$\nu = \begin{cases} 4d & \text{if } d = 1 \\ 3.5d & \text{if } d \leq 3 \\ 2.5d & \text{if } d \leq 8 \\ 1.8d & \text{if } d > 8 \end{cases} \quad (36)$$

- **Polynomial HOOT (P-HOOT):** Power mean backup with polynomial exploration bonuses, using dimension-adaptive parameters:

$$\alpha = \begin{cases} 5 & \text{if } d = 1 \\ 3 & \text{if } d \leq 3 \\ 2 & \text{if } d \leq 8 \\ 1.5 & \text{if } d > 8 \end{cases} \quad (37)$$

$$\xi = \begin{cases} 20 & \text{if } d = 1 \\ 15 & \text{if } d \leq 3 \\ 10 & \text{if } d \leq 8 \\ 8 & \text{if } d > 8 \end{cases} \quad (38)$$

$$\eta = \begin{cases} 0.5 & \text{if } d = 1 \\ 0.4 & \text{if } d \leq 3 \\ 0.3 & \text{if } d \leq 8 \\ 0.25 & \text{if } d > 8 \end{cases} \quad (39)$$

E.3 EXPERIMENTAL PROTOCOL

Planning and Evaluation. Each experiment uses the following protocol:

- **Planning budget:** 1000 MCTS simulations per re-planning step
- **Planning horizon:** $H = 150$ steps with discount factor $\gamma = 0.99$
- **Evaluation:** 20 independent runs with different random seeds
- **Re-planning frequency:** After each environment step during evaluation
- **Performance metric:** Cumulative discounted reward over 150 evaluation steps

Iteration Budget Analysis. We evaluate performance across multiple computational budgets using a geometric progression:

$$\text{base} = 1000^{1/15} \quad (40)$$

$$\text{samples}_i = \lfloor 3 \times \text{base}^i \rfloor, \quad i \in \{0, 1, 2, 3, 4, 5\} \quad (41)$$

This yields iteration budgets approximately in $\{3, 4, 6, 8, 11, 16\} \times 100$ simulations.

Dimension-Adaptive Rollout Strategy. For computational efficiency in high-dimensional environments, we implement adaptive rollout strategies:

- **Low-dimensional** ($d \leq 6$): Full tree search to maximum depth
- **High-dimensional** ($d > 6$): Hybrid approach with rollout depth $\min(20, H/2)$
- **Re-planning budget:** Full budget for $d \leq 6$, reduced to $\max(\text{budget}/2, 100)$ for $d > 6$

Computational Resources. All experiments were conducted on Intel(R) Core(TM) i9-14900K 3.20 GHz with 32 cores per CPU.

F EXPONENTIAL DEPENDENCE ON d AND GUARANTEES UNDER A CAPPED IMPLEMENTATION

Unavoidable exponential term in d . Under β -Hölder regularity of $Q^*(s_0, \cdot)$ with constant L on a d -dimensional action set, any *planning* procedure that returns an ε -optimal action with high probability must localize the maximizer within a ball of radius $\Theta((\varepsilon/L)^{1/\beta})$. The metric entropy of such balls scales as $\Theta((L/\varepsilon)^{d/\beta})$; with finite-variance rollouts, one additionally needs the usual ε^{-2} factor for mean estimation. Our analysis makes this precise at the root:

$$N(\varepsilon, \delta) \leq C_1 \sigma^2 L^{\frac{d}{\beta}} \varepsilon^{-(\frac{d}{\beta}+2)} \log \left(\frac{C_2 L^{\frac{d}{\beta}}}{\varepsilon^{\frac{d}{\beta}} \delta} \right),$$

i.e., an **exponential dependence in d** through the covering factor $L^{d/\beta} \varepsilon^{-d/\beta}$ is *intrinsic* to continuous-action planning under β -Hölder assumptions; the extra ε^{-2} is the unavoidable Monte Carlo cost.

Why fixed grids are insufficient; why *progressive refinement* is necessary. A static ε -grid incurs an approximation floor of order $L\varepsilon^\beta$ that does not vanish with more samples; hence it learns (at best) the grid optimum. In contrast, our schedule refines ε_k geometrically and triggers refinement only when the geometric term $L\varepsilon_k^\beta$ and the statistical term balance. This is what yields the end-to-end bound above.

Theorem 4 (Sample Complexity with Polynomial-tail Estimates). *PM-DA-MCTS returns an ε -optimal action at s_0 with probability at least $1 - \delta$ after at most*

$$N(\varepsilon, \delta) \leq C_1 \sigma^2 L^{\frac{d}{\beta}} \varepsilon^{-(\frac{d}{\beta}+2)} \log \left(\frac{C_2 L^{\frac{d}{\beta}}}{\varepsilon^{\frac{d}{\beta}} \delta} \right),$$

where $C_1, C_2 > 0$ depend only on d, β and the refinement/selection constants (but not on $\varepsilon, \delta, L, \sigma$).

Lower bound. The minimax rate for continuum-armed bandits with β -Hölder smoothness already appears in the *X-armed bandits* analysis of Bubeck et al. (2011). Let (\mathcal{X}, ℓ) be a metric space with packing numbers $N(\mathcal{X}, \ell, \varepsilon) \gtrsim \varepsilon^{-D}$. Theorem 13 in Bubeck et al. (2011) gives a cumulative-regret lower bound of order $n^{(D+1)/(D+2)}$; for $\ell(x, y) \asymp \|x - y\|^\beta$ on $[0, 1]^d$ this yields $D = d/\beta$. Standard conversion to fixed-confidence simple regret (or a direct Fano/packing argument with “bump” functions of radius $\Theta((\varepsilon/L)^{1/\beta})$ and height $\Theta(\varepsilon)$) then implies that any planner returning an ε -optimal action with probability at least $1 - \delta$ must use at least

$$N(\varepsilon, \delta) \geq c \sigma^2 L^{\frac{d}{\beta}} \varepsilon^{-(\frac{d}{\beta}+2)}$$

samples, up to logarithmic factors. Hence the $L^{d/\beta} \varepsilon^{-d/\beta}$ dependence is information-theoretically necessary. Our upper bound (Theorem 4) matches this rate up to logarithms.

Capped implementation and coverage along the realized search. The analysis indexes children at level k via an abstract ε_k -net to guarantee the *covering property* in each *visited* cell v :

$$\forall x \in \text{cell}(v) \exists a \in N_k : \|x - a\| \leq \varepsilon_k. \quad (\star)$$

To enforce (\star) *without materializing a full net*, we instantiate *capped random* anchors: draw M_k i.i.d. points uniformly in the current cell and keep at most a cap M_k^{\max} . The following standard fact controls coverage.

Lemma 11 (High-probability coverage of a random capped net). *Let $\mathcal{R}_k \subset \mathbb{R}^d$ be a measurable region of diameter $D_k \in (0, \infty)$ and let $\varepsilon_k \in (0, D_k]$. Draw M_k i.i.d. anchors A_1, \dots, A_{M_k} uniformly from \mathcal{R}_k , and let*

$$\tilde{\varepsilon}(M_k) := \sup_{x \in \mathcal{R}_k} \min_{1 \leq j \leq M_k} \|x - A_j\|_2$$

be the random covering radius of the anchor set. There exist constants $c_d, C_d \in (0, \infty)$, depending only on d and on a shape-regularity constant of \mathcal{R}_k , such that:

(i) For any confidence parameters $\delta \in (0, 1)$ and $T \geq 1$, if

$$M_k \geq c_d \left(\frac{D_k}{\varepsilon_k} \right)^d \left(\log \left(\frac{D_k}{\varepsilon_k} \right) + \log \left(\frac{T}{\delta} \right) \right),$$

then $\tilde{\varepsilon}(M_k) \leq \varepsilon_k$ with probability at least $1 - \delta/T$; equivalently, the anchors form an ε_k -net of \mathcal{R}_k w.p. $\geq 1 - \delta/T$.

(ii) (Rate) There exist $c'_d, C'_d > 0$ such that, for all large M_k ,

$$\Pr \left\{ \tilde{\varepsilon}(M_k) \leq C'_d D_k \left(\frac{\log M_k}{M_k} \right)^{1/d} \right\} \geq 1 - M_k^{-2}.$$

In particular, $\tilde{\varepsilon}(M_k) = \Theta(D_k (\log M_k / M_k)^{1/d})$ w.h.p.

Proof. 1: A volumetric discretization. Partition \mathcal{R}_k into m axis-aligned cubes of side $s := \varepsilon_k / (2\sqrt{d})$ (discarding partial cubes outside \mathcal{R}_k); by shape regularity there exist constants $0 < c_1 \leq c_2 < \infty$ (depending on d) such that

$$c_1 \left(\frac{D_k}{\varepsilon_k} \right)^d \leq m \leq c_2 \left(\frac{D_k}{\varepsilon_k} \right)^d.$$

Let C_1, \dots, C_m denote these microcells and write $p_j := \text{vol}(C_j) / \text{vol}(\mathcal{R}_k)$. Then $p_j \geq c_3 (\varepsilon_k / D_k)^d$ for some $c_3 = c_3(d) > 0$.

2: Empty-cell probability and union bound. For any fixed j , the probability C_j is empty is $\Pr\{C_j \text{ empty}\} = (1 - p_j)^{M_k} \leq \exp(-p_j M_k) \leq \exp(-c_3 M_k (\varepsilon_k / D_k)^d)$. By a union bound over the m cells,

$$\Pr\{\exists j : C_j \text{ empty}\} \leq m \exp(-c_3 M_k (\varepsilon_k / D_k)^d).$$

Therefore, if $M_k \geq (1/c_3) (D_k / \varepsilon_k)^d (\log m + \log(T/\delta))$, then $\Pr\{\text{all cells hit}\} \geq 1 - \delta/T$. Using $\log m \leq C_4 \log(D_k / \varepsilon_k)$ and $m \leq c_2 (D_k / \varepsilon_k)^d$ yields the stated condition with c_d large enough.

3: From nonempty cells to an ε_k -net. If every C_j is hit by at least one anchor, then for any $x \in \mathcal{R}_k$ the unique cell C_j containing x also contains some anchor A_j ; since the cell diameter in ℓ_2 is $\sqrt{d} s = \varepsilon_k / 2$, we have $\|x - A_j\|_2 \leq \varepsilon_k / 2$, hence $\tilde{\varepsilon}(M_k) \leq \varepsilon_k$.

4: The rate in (ii). Let $\varepsilon > 0$ be chosen so that $m \asymp (D_k / \varepsilon)^d$ and $p_j \gtrsim (\varepsilon / D_k)^d$. Repeating Step 2 with failure target M_k^{-2} shows

$$\Pr\{\tilde{\varepsilon}(M_k) > \varepsilon\} \lesssim (D_k / \varepsilon)^d \exp(-c M_k (\varepsilon / D_k)^d).$$

Solving $(D_k / \varepsilon)^d \exp(-c M_k (\varepsilon / D_k)^d) \leq M_k^{-2}$ gives $\varepsilon \lesssim D_k (\log M_k / M_k)^{1/d}$ and yields the claim. \square

Corollary 5 (Finite-variance guarantee under a capped implementation). *Run PM-DA-MCTS with the usual selection/refinement rule. At each visited node whose current level is k , instantiate at most M_k^{\max} random anchors uniformly in the node's region \mathcal{R}_k , where*

$$M_k^{\max} \geq c_d \left(\frac{D_k}{\varepsilon_k} \right)^d \left(\log \left(\frac{T}{\delta} \right) + \log \left(\frac{D_k}{\varepsilon_k} \right) \right)$$

for all levels $k \leq k_*$ actually reached by the search (constants as in Lemma 11). Then, with probability at least $1 - \delta$, the covering property

$$(\star) \quad \forall x \in \mathcal{R}_k \exists a \text{ (instantiated at level } k) : \|x - a\|_2 \leq \varepsilon_k$$

holds simultaneously for every visited node/level up to k_* ; consequently the geometric optimism term at that node remains $L\varepsilon_k^\beta$, and the sample-complexity bound of Theorem 4 continues to hold up to absolute constants and polylogarithmic factors.

Proof. Apply Lemma 11(i) independently at every *visited* node/level, with per-event failure budget $\delta/(2T)$, where T upper-bounds the total number of such (node,level) pairs reachable under the rollout budget (a standard polynomial envelope in the main proof). A union bound yields that (\star) holds at all visited nodes with probability at least $1 - \delta/2$.

On this high-probability event, for each visited node the discretization (geometric) error is uniformly bounded by $L\varepsilon_k^\beta$ exactly as in the idealized analysis that assumes a global ε_k -net. The selection rule is unchanged; only actions that are instantiated can be chosen, but (\star) guarantees that an ε_k -near-optimal anchor exists among them. The concentration of the power-mean backups with polynomial bonuses (node-wise polynomial tails and their depth-wise propagation) depends on visit counts and tail exponents, not on how the anchors were *generated*. Therefore, the coupling used in the proof of Theorem 4 goes through verbatim: the algorithm refines only when the statistical and geometric terms balance, and when it stops at level k_\star the selected action is ε -optimal at the root with probability at least $1 - \delta$. The only change is an additional (benign) logarithmic factor $\log(D_k/\varepsilon_k)$ from Lemma 11, which is absorbed by the global $\log(C_2 L^{d/\beta}/(\varepsilon^{d/\beta}\delta))$ term already present in Theorem 4. \square

Takeaways for complexity and practice. (i) The *exponential* $L^{d/\beta}\varepsilon^{-d/\beta}$ factor is the geometric price of localizing the maximizer in d dimensions under β -Hölder smoothness; the ε^{-2} factor is the statistical price of finite-variance estimation. (ii) The capped, on-demand anchor generation avoids materializing a full net while *preserving* the covering guarantees needed by the proof along the actually visited cells; in code we use small caps (e.g., 1.5–2K per level for $d \leq 8/d > 8$) because k_\star is limited by the budget, and the optimistic path touches few cells.