A FINITE-SAMPLE ANALYSIS OF QUANTILE TEMPORAL-DIFFERENCE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantile temporal-difference learning (QTD) is an effective method for learning return distributions through quantile approximation, yet its finite-time behavior remains poorly understood. Its update is nonlinear and nonsmooth, and the stability needed for a sharp convergence rate holds only near the target. We establish a global high-probability last-iterate guarantee for synchronous tabular QTD under general positive, nonincreasing step-size sequences and arbitrary initialization in the natural parameter range. For polynomially decaying step sizes with exponent $a \in (0, 1)$, the last iterate converges to the target at rate $T^{-a/2}$ in the infinity norm, up to logarithmic and lower-order terms. A suitably tuned harmonic schedule recovers the $T^{-1/2}$ statistical rate up to logarithmic factors. For the m-quantile representation, its ∞-Wasserstein error scales as $\sqrt{m/T}$ up to logarithmic factors, matching the leading polynomial dependence on the quantile resolution and sample size of the corresponding model-based estimator. The proof uses a two-stage global-to-local argument. From arbitrary initialization, Bellman contraction and CDF monotonicity first bring the iterate close to the target, after which, a novel variance–drift matching argument sharpens the control of accumulated noise and local contraction reduces the remaining errors, yielding the sharp rate. Simulations verify the predicted polynomial decay and assess the finite-time entrance bound.

1 INTRODUCTION

Distributional reinforcement learning (DRL) learns the distribution of future outcomes rather than only their expectation. By retaining the full distribution of possible returns, it provides a richer picture of the uncertainty induced by the environment and the policy. Methods built on this idea have achieved strong performance in challenging games and control tasks, and have also been applied to real-world problems such as robotic manipulation and autonomous navigation (Bellemare et al., 2017; Dabney et al., 2018b;a; Barth-Maron et al., 2018; Bodnar et al., 2020; Bellemare et al., 2020).

Many distributional methods build on temporal-difference (TD) learning, which uses short transitions to learn long-term returns without waiting for an episode to end (Sutton & Barto, 2018). Classical TD learning predicts the expected return, whereas distributional TD learning uses the same recursive principle to estimate its full distribution (Bellemare et al., 2017). One prominent approach is quantile TD learning (QTD), which represents the return distribution through a fixed collection of quantiles, with each quantile corresponding to a value below which a specified proportion of returns falls, and learns their locations through quantile regression (Dabney et al., 2018b; Rowland et al., 2024). QTD thus reduces distributional learning to a collection of simple but interacting quantile updates.

Despite its empirical success, the finite-time behavior of QTD remains poorly understood. For nonparametric and categorical distributional TD learning, finite-sample analyses have established a minimax-optimal rate in the 1-Wasserstein distance, showing that estimating a return distribution need not sacrifice statistical efficiency (Peng et al., 2024). For QTD, almost-sure asymptotic convergence establishes that the quantile iterates eventually approach their target (Rowland et al., 2024). These results, however, do not describe how QTD behaves over a finite run. This leaves a basic question: *how quickly does QTD converge to its target?*

In this work, we develop global, high-probability finite-time convergence guarantees for the last-iterate estimator of synchronous tabular QTD under a generative model. Our results cover a general

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

class of positive, nonincreasing step-size sequences under explicit conditions, and the resulting bounds quantify both the transient effect of initialization and the stochastic error induced by the step-size schedule. In particular, they yield guarantees for three common schedules: (i) a constant step size $\alpha_t = \eta$, (ii) a polynomially decaying step size $\alpha_t = c(t+1)^{-a}$ with $0 < a < 1$, and (iii) a harmonic step size $\alpha_t = c/(t+t_0)$. We show that, up to logarithmic factors, a constant step size forgets the initialization geometrically until reaching a noise floor of order $\sqrt{\eta}$; polynomial decay yields an error of order $T^{-a/2}$; and a suitably tuned harmonic schedule recovers the canonical $T^{-1/2}$ statistical rate. Moreover, for an m-quantile representation, the harmonic schedule yields a W_∞ error scaling as $\sqrt{m/T}$ up to logarithmic factors, matching the model-based estimator of Cheng et al. (2026b) in its leading polynomial dependence on quantile resolution m and sample size.

The main technical difficulty is that sharp control of the QTD recursion is available only near the target, while the iterate may start far away. A purely local analysis therefore cannot explain why the iterate reaches this region and remains there. We address this through a two-stage argument: a global analysis first drives the iterate into a suitable local neighborhood, after which a refined concentration argument controls its stochastic fluctuations and prevents exit with high probability. Within the local regime, a key variance–drift matching property leads to a sharper rate than worst-case bounded-noise arguments. Intuitively, an extreme quantile receives a weaker pull toward its target, but its quantile score also fluctuates less. Our analysis keeps these two effects together.

Our contributions are as follows.

(i) We establish a global finite-sample, high-probability last-iterate guarantee for synchronous tabular QTD. A unified result for positive, nonincreasing step sizes yields explicit guarantees for constant, polynomial, and harmonic schedules.

(ii) We develop a two-stage analysis that connects global convergence with sharp local rate analysis: the first stage guarantees entrance into a local region, while the second controls stochastic fluctuations and possible exit from that region.

(iii) We uncover a variance–drift matching property of the local QTD dynamics. At an extreme quantile, a weak restoring drift is accompanied by smaller quantile noise, yielding sharper stochastic fluctuation control than worst-case bounded-noise analysis.

Related work. Foundational work on DRL developed distributional Bellman operators and categorical and quantile approximations (Bellemare et al., 2017; Rowland et al., 2018; Dabney et al., 2018b). For quantile representations, Rowland et al. (2024) established asymptotic convergence of QTD, while Cheng et al. (2026a) established asymptotic normality and studied statistical inference for QTD. Rowland et al. (2023) studied the statistical benefits of quantile approximations for value estimation. Finite-sample analyses are more developed for categorical and nonparametric distributional methods (Peng et al., 2024; 2025). QTD resembles nonsmooth quantile regression, but its evolving Bellman target prevents a direct application of fixed-objective nonsmooth SGD analyses (Shamir & Zhang, 2013; Harvey et al., 2019). Supplementary background and related work are provided in Appendix A.

Paper organization. Section 2 defines the model, the QTD target, and the algorithm. Section 3 states the general finite-sample result and its consequences for the three step-size schedules. Section 4 explains the two-stage proof, and Section 5 provides numerical verification. Supplemental background and related work, complete proofs, and additional experiments are all deferred to the appendix.

2 Problem Setup

2.1 Distributional policy evaluation

Consider a discounted Markov decision process $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}_R, P, \gamma \rangle$. Here $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $s, s' \in \mathcal{S}$ denote states, and $a \in \mathcal{A}$ denotes an action. Given (s, a), $\mathcal{P}_R(\cdot \mid s, a)$ is the reward distribution, supported on $[0, 1]$, and $P(\cdot \mid s, a)$ is the distribution of the next state. $\gamma \in (0, 1)$ is the discount factor. Fix a policy π, where $\pi(a \mid s)$ is the probability of choosing action a in state s. At time t, the policy draws $A_t \sim \pi(\cdot \mid S_t)$, the environment draws $R_t \sim \mathcal{P}_R(\cdot \mid S_t, A_t)$ and $S_{t+1} \sim P(\cdot \mid S_t, A_t)$. Starting from $S_0 = s$, the return is

$G^\pi(s) = \sum_{t=0}^{\infty} \gamma^t R_t$. We denote its distribution by $\eta^\pi(s) = \mathrm{Law}(G^\pi(s))$ and collect these statewise distributions as $\boldsymbol{\eta}^\pi = (\eta^\pi(s)\colon s \in \mathcal{S})$. Because rewards lie in $[0,1]$, the return lies in $[0,(1-\gamma)^{-1}]$.

For a state-indexed collection of probability measures $\boldsymbol{\eta} = (\eta(s)\colon s \in \mathcal{S})$, let $b_{r,\gamma}(x) = r + \gamma x$. We use $(b_{r,\gamma})_{\#}\nu$ for the distribution of $r + \gamma X$ when $X \sim \nu$. The distributional Bellman operator is

$$(\mathcal{T}^\pi \boldsymbol{\eta})(s) = \sum\nolimits_{a\in\mathcal{A},,s'\in\mathcal{S}} \pi(a \mid s) P(s' \mid s,a) \int_{\mathbb{R}} (b_{r,\gamma})_{\#}\eta(s'), \mathcal{P}_R(\mathrm{d}r \mid s,a). \tag{1}$$

Its unique fixed point is the collection of return distributions $\boldsymbol{\eta}^\pi$.

We next define the finite quantile representation used by QTD. For a probability measure ν on $\mathbb{R}$, let $F_\nu(x) = \nu((-\infty, x])$ be its cumulative distribution function (CDF) and let $F_\nu^{-1}(\tau) = \inf\{x\colon F_\nu(x) \geq \tau\}$ be its quantile function. At the mid-quantile levels $\tau_i = (2i-1)/(2m)$, $i \in [m] \coloneqq \{1,\ldots,m\}$, define $\Pi_m\nu = m^{-1}\sum_{i=1}^m \delta_{F_\nu^{-1}(\tau_i)}$, where δ_x denotes a point mass at x. For a state-indexed collection $\boldsymbol{\eta}$, let $\boldsymbol{\Pi}_m$ apply the projection statewise, so that $(\boldsymbol{\Pi}_m\boldsymbol{\eta})(s) = \Pi_m\eta(s)$.

We next introduce the W_∞ metric, under which the distributional Bellman operator is contractive and the quantile projection is nonexpansive. For two probability measures μ and ν define $W_\infty(\mu,\nu) = \sup_{u\in(0,1)} \left|F_\mu^{-1}(u) - F_\nu^{-1}(u)\right|$. For state-indexed collections $\boldsymbol{\mu} = (\mu(s)\colon s \in \mathcal{S})$ and $\boldsymbol{\nu} = (\nu(s)\colon s \in \mathcal{S})$, write $\overline{W}_\infty(\boldsymbol{\mu},\boldsymbol{\nu}) = \max_{s\in\mathcal{S}} W_\infty(\mu(s),\nu(s))$. The metric W_∞ is the largest horizontal separation between two quantile functions and therefore measures the worst quantile-location error. The additional measure-theoretic details are collected in Appendix B.

Proposition 2.1 (Bellman and projection stability)**.** *For all state-indexed collections of probability measures $\boldsymbol{\eta}$ and $\boldsymbol{\eta}'$, the distributional Bellman operator and the quantile projection satisfy*

$$\overline{W}_\infty(\mathcal{T}^\pi\boldsymbol{\eta}, \mathcal{T}^\pi\boldsymbol{\eta}') \leq \gamma\overline{W}_\infty(\boldsymbol{\eta},\boldsymbol{\eta}'), \qquad \overline{W}_\infty(\boldsymbol{\Pi}_m\boldsymbol{\eta}, \boldsymbol{\Pi}_m\boldsymbol{\eta}') \leq \overline{W}_\infty(\boldsymbol{\eta},\boldsymbol{\eta}').$$

Proposition 2.1 follows from Bellemare et al. (2023, Proposition 4.15 and Lemma 5.25). It implies that $\boldsymbol{\Pi}_m\mathcal{T}^\pi$ is a γ-contraction and therefore has a unique fixed point $\boldsymbol{\eta}_m$. Thus, it identifies the target of QTD and lets us control the distance between distributions before the iterates enter a neighborhood of η_m, where the conditional mean update admits a first-order Taylor expansion. We write $\eta_m(s) = m^{-1}\sum_{i=1}^m \delta_{\theta_m(s,i)}$, with $\theta_m(s,1) \leq \ldots \leq \theta_m(s,m)$, and call the array $\boldsymbol{\theta}_m = (\theta_m(s,i)\colon s \in \mathcal{S}, i \in [m])$ the quantile-projected Bellman fixed point. More generally, an array $\boldsymbol{\theta} = (\theta(s,i)\colon s \in \mathcal{S}, i \in [m]) \in \mathbb{R}^{\mathcal{S}\times[m]}$ represents the statewise distributions $\eta_{\boldsymbol{\theta}}(s) = \frac{1}{m}\sum_{i=1}^m \delta_{\theta(s,i)}$; its entries need not be ordered. In particular, $\boldsymbol{\eta}_m = \boldsymbol{\eta}_{\boldsymbol{\theta}_m}$. The approximation error $\overline{W}_\infty(\boldsymbol{\eta}_m,\boldsymbol{\eta}^\pi)$ vanishes as $m \to \infty$ under our assumptions (Cheng et al., 2026b, Theorem 3.2). And we measure the statistical error by $\|\boldsymbol{\theta} - \boldsymbol{\theta}_m\|_\infty$, which bounds the corresponding W_∞ error: $\overline{W}_\infty(\boldsymbol{\eta}_{\boldsymbol{\theta}},\boldsymbol{\eta}_m) \leq \|\boldsymbol{\theta} - \boldsymbol{\theta}_m\|_\infty$. In the following, all the results for $\|\boldsymbol{\theta} - \boldsymbol{\theta}_m\|_\infty$ hold naturally for $\overline{W}_\infty(\boldsymbol{\eta}_{\boldsymbol{\theta}},\boldsymbol{\eta}_m)$.

2.2 Synchronous QTD

At iteration $t+1$, for each state $s \in \mathcal{S}$, the generative model draws independently across states

$$a^{(t+1,s)} \sim \pi(\cdot \mid s), \qquad r^{(t+1,s)} \sim \mathcal{P}_R(\cdot \mid s, a^{(t+1,s)}), \qquad s'^{(t+1,s)} \sim P(\cdot \mid s, a^{(t+1,s)}).$$

The superscript $(t+1,s)$ records the iteration and the state being updated. For a parameter array $\boldsymbol{\theta}$ and a threshold $z \in \mathbb{R}$, define the empirical Bellman-target CDF

$$\widehat{F}_{s,\boldsymbol{\theta}}^{(t+1)}(z) = \frac{1}{m}\sum\nolimits_{j=1}^m \mathbf{1}\left\{r^{(t+1,s)} + \gamma\theta\left(s'^{(t+1,s)}, j\right) < z\right\}.$$

Here $\mathbf{1}\{\cdot\}$ denotes the indicator of an event. For each (s,i), the synchronous QTD update is

$$\theta^{(t+1)}(s,i) = \theta^{(t)}(s,i) + \alpha_t\left\{\tau_i - \widehat{F}_{s,\boldsymbol{\theta}^{(t)}}^{(t+1)}\left(\theta^{(t)}(s,i)\right)\right\},$$

where $(\alpha_t)_{t\geq 0}$ are deterministic, positive and nonincreasing step sizes. If too few sampled targets lie below the current coordinate $\theta^{(t)}(s,i)$, the update moves it upward; if too many lie below it, the update moves it downward. Each coordinate is therefore pushed toward its prescribed quantile level τ_i. Rowland et al. (2024) established almost-sure convergence of tabular QTD to the fixed point $\boldsymbol{\theta}_m$; our goal is to quantify this convergence at finite time. Throughout the general analysis, $(\alpha_t)_{t\geq 0}$ is deterministic, positive, and nonincreasing, with $\bar{\alpha} = \alpha_0$. Section 3 states the resulting guarantees for constant, polynomially decreasing, and harmonic schedules.

2.3 REGULARITY ASSUMPTIONS

We impose the following conditions on the reward distributions.

Assumption 1 (Lipschitz reward densities). *For every $(s,a) \in \mathcal{S} \times \mathcal{A}$, the reward distribution admits a density $p_{s,a}$ on $(0,1)$. For some constants $C_0 \geq 1$ and $L \geq 0$, uniformly over $(s,a) \in \mathcal{S} \times \mathcal{A}$, $0 < p_{s,a}(x) \leq C_0$ for all $x \in (0,1)$ and $p_{s,a}$ is L-Lipschitz on $(0,1)$. At the endpoints, either:*

(i) there is a constant $c_0 > 0$ such that $p_{s,a}(x) \geq c_0$ for every $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $x \in (0,1)$; or

(ii) let $\widetilde{p}_{s,a}$ denote the zero extension of $p_{s,a}$ to $\mathbb{R}$: it equals $p_{s,a}$ on $(0,1)$ and zero elsewhere. The function $\widetilde{p}_{s,a}$ is L-Lipschitz on $\mathbb{R}$. Moreover, there is a constant $\kappa \in (0, 1/2]$ such that every $p_{s,a}$ is nondecreasing on $(0,\kappa)$ and nonincreasing on $(1-\kappa, 1)$.

Assumption 1 imposes regularity conditions on the reward densities. The upper bound C_0 prevents too much probability from being concentrated in a very short interval, while the Lipschitz constant L prevents the density from changing abruptly. We allow two behaviors at the endpoints of the reward range. In case (i), the density stays uniformly positive throughout $(0,1)$. In case (ii), it may decrease to zero at either endpoint, as happens for some beta distributions, but it must do so without oscillating. These conditions ensure that, near the target quantiles, location errors induce a non-negligible change in the corresponding CDF values, yielding a corrective QTD update.

Assumption 2 (No boundary quantile). *Let $P^\pi(s' \mid s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) P(s' \mid s, a)$ be the transition probability from s to s' under policy π. For every integer $m \geq 1$, every $s, s' \in \mathcal{S}$ with $P^\pi(s' \mid s) > 0$, and every $i, j \in [m]$, we assume that $\theta_m(s,i) - \gamma\theta_m(s',j) \notin \{0,1\}$.*

To interpret Assumption 2, observe that a reward r aligns the next-state value $\gamma\theta_m(s',j)$ with the current quantile $\theta_m(s,i)$ precisely when $r = \theta_m(s,i) - \gamma\theta_m(s',j)$. The assumption simply says that none of these relevant reward values is exactly 0 or 1, the two endpoints at which the density may change its behavior. This excludes degenerate boundary configurations that would otherwise obstruct the local analysis around the target quantiles.

3 MAIN RESULTS

Our main result gives a high-probability bound for the last QTD iterate under any positive, nonincreasing step-size sequence. The result separates a burn-in requirement for reaching the local region from the subsequent error bound, which consists of the contracted error inherited at entrance and the accumulated sampling fluctuations. We first state this general result and then specialize it to three step size schedules. To state the precise bounds, we need one property of the MDP. Define

$$c_{\mathcal{M}} := \inf_{m \in \mathbb{N},\ s \in \mathcal{S},\ i \in [m],\ 0 < |u| \leq (1-\gamma)^{-1}} \frac{\left| F_{(\mathcal{T}^\pi \boldsymbol{\eta}_m)(s)}(\theta_m(s,i) + u) - \tau_i \right|}{\tau_i(1-\tau_i)|u|}. \tag{2}$$

Under Assumptions 1 and 2, $c_{\mathcal{M}} > 0$ and depends only on the MDP $\mathcal{M}$ (Cheng et al., 2026b, Lemma B.1). It measures how sensitively the Bellman-target CDF changes when a candidate quantile moves away from its target. A larger $c_{\mathcal{M}}$ means a stronger corrective signal in the QTD update.

Theorem 3.1 (General nonincreasing step sizes). *Suppose Assumptions 1 and 2 hold, $\boldsymbol{\theta}^{(0)} \in [0, (1-\gamma)^{-1}]^{\mathcal{S} \times [m]}$. Assume further that the initial, and hence largest, step size satisfies $\alpha_0 C_0 \leq 1$. Fix $\delta \in (0,1)$ and $T \geq 8$, and set $u_T = \lfloor T/4 \rfloor$ and $v_T = \lfloor T/2 \rfloor$. Let $\{u_T, u_T+1, \ldots, v_T - 1\}$ be the entrance block and $\{v_T, v_T+1, \ldots, T\}$ be the local analysis block. Define the accumulated step size and squared step size on the entrance block, and the accumulated step size after that block, by*

$$A_T^{\mathrm{ent}} := \sum_{t=u_T}^{v_T-1} \alpha_t, \qquad V_T^{\mathrm{ent}} := \sum_{t=u_T}^{v_T-1} \alpha_t^2, \qquad A_T^{\mathrm{loc}} := \sum_{t=v_T}^{T-1} \alpha_t.$$

We also define

$$\ell_T(\delta) = \log \frac{8|\mathcal{S}|mT^2}{\delta}, \qquad b_T(\delta) = 2\sqrt{\frac{\alpha_{u_T}\ell_T(\delta)}{c_{\mathcal{M}}(1-\gamma)}} + \frac{\alpha_{u_T}\ell_T(\delta)}{3}.$$

If the following two burn-in conditions hold:

$$c_{\mathrm{g}} A_T^{\mathrm{ent}} \geq D_{\mathrm{g}} + \frac{\beta}{2} V_T^{\mathrm{ent}} + \sqrt{2 V_T^{\mathrm{ent}} \log \frac{2}{\delta}}, \qquad \textit{(Entrance)} \tag{3a}$$

$$b_T(\delta) \le \min\left\{\frac{r_{\mathrm{out}}}{16}, \frac{\mu}{64 L_h}\right\}, \qquad \textit{(Capture)} \tag{3b}$$

where $L_h := L(1+\gamma)^2/2$, $\mu := c_{\mathcal{M}}(1-\gamma)/(4m)$ is the local contraction rate, and r_{out} is the local-region radius. r_{out} together with $c_{\mathrm{g}}, D_{\mathrm{g}}, \beta$ is defined in Appendix C.1. Then, with probability at least $1-\delta$, it holds that

$$\left\|\boldsymbol{\theta}^{(T)} - \boldsymbol{\theta}_m\right\|_\infty \le \frac{r_{\mathrm{out}}}{2} \exp\left\{-\frac{15 c_{\mathcal{M}}(1-\gamma)}{64m} A_T^{\mathrm{loc}}\right\} + \frac{31}{30} b_T(\delta). \tag{4}$$

The first burn-in condition (3a) says that the accumulated progress A_T^{ent} is sufficient to overcome the initial error D_{g} and noise (the V_T^{ent} terms) and reach a neighborhood of the target. The second (3b) says that subsequent sampling fluctuations $b_T(\delta)$ are small enough to keep the iterate there. This division into blocks is only a proof device; QTD runs without a restart or change of step-size rule.

Stage I in Section 4 establishes entrance. Conditional on entrance, Stage II contracts the error remaining at entrance, producing the exponential term in (4), and controls the subsequent accumulated noise, producing $b_T(\delta)$. We next specialize the general result to the three step-size schedules.

Corollary 3.2 (Constant and polynomial step sizes)**.** *Under the assumptions of Theorem 3.1, the following bounds hold whenever its two burn-in conditions (3a), (3b) are satisfied.*

(i) If $\alpha_t = \eta$, then, with probability at least $1-\delta$,

$$\left\|\boldsymbol{\theta}^{(T)} - \boldsymbol{\theta}_m\right\|_\infty \le C \exp\left\{-\frac{c_{\mathcal{M}}(1-\gamma)\eta T}{Cm}\right\} + C\sqrt{\frac{\eta \ell_T(\delta)}{c_{\mathcal{M}}(1-\gamma)}} + C\eta\ell_T(\delta).$$

(ii) If $\alpha_t = c(t+1)^{-a}$ with $a \in (0,1)$ and $cC_0 \le 1$, then, with probability at least $1-\delta$,

$$\left\|\boldsymbol{\theta}^{(T)} - \boldsymbol{\theta}_m\right\|_\infty \le C \exp\left\{-\frac{c_{\mathcal{M}}(1-\gamma)cT^{1-a}}{Cm}\right\} + C\sqrt{\frac{c\ell_T(\delta)}{c_{\mathcal{M}}(1-\gamma)T^a}} + \frac{Cc\ell_T(\delta)}{T^a}. \tag{5}$$

For a constant step size, QTD forgets its initialization geometrically, but every iteration continues to inject noise at the same scale. The iterate therefore eventually fluctuates in a neighborhood whose leading radius scales as $\sqrt{\eta}$. A smaller η shrinks this neighborhood but slows entrance, while the burn-in conditions require $T\eta$ to be sufficiently large and $\eta\ell_T(\delta)$ sufficiently small.

For polynomial decay, the decreasing steps also reduce the stochastic fluctuations, giving the leading last-iterate rate $T^{-a/2}$ up to logarithmic factors. For every fixed $a \in (0,1)$, the burn-in conditions hold for all sufficiently large T. As in the asymptotic analysis of Rowland et al. (2024), square summability of the step sizes is not required. Here we further obtain a finite-time rate: for every $a \in (0,1)$, including $a \le 1/2$, the last-iterate error decays as $T^{-a/2}$ up to logarithmic factors.

The endpoint t^{-1} decay of the harmonic schedule requires a separate dyadic-block argument. We divide time into blocks whose endpoints double, such as $[N, 2N)$, $[2N, 4N)$, and $[4N, 8N)$; these are called *dyadic blocks*. For the harmonic step size, the individual steps become smaller over time, but their sum within each such block remains of constant order. Thus, every block reduces the error inherited from the previous block by a fixed factor, while the sampling noise becomes smaller.

Corollary 3.3 (Harmonic step size)**.** *Under the assumptions of Theorem 3.1, let $\alpha_t = c/(t+t_0)$. There is a universal constant $C_{\mathrm{harm}} > 0$ such that, if $c \ge C_{\mathrm{harm}}\mu^{-1}$ and $t_0 \ge cC_0$, then there is a finite deterministic burn-in $T_{\mathrm{harm}}(\delta)$, defined in (34), for which the following holds. For every $T \ge T_{\mathrm{harm}}(\delta)$, with probability at least $1-\delta$,*

$$\left\|\boldsymbol{\theta}^{(T)} - \boldsymbol{\theta}_m\right\|_\infty \le C\left\{\sqrt{\frac{c\log(C|\mathcal{S}|mT/\delta)}{c_{\mathcal{M}}(1-\gamma)T}} + \frac{c\log(C|\mathcal{S}|mT/\delta)}{T}\right\}. \tag{6}$$

The condition $c \ge C_{\mathrm{harm}}\mu^{-1}$ ensures that each dyadic block makes enough progress to forget the error inherited from the preceding block to keep pace with the decreasing noise, while the offset t_0 keeps the early steps stable. The harmonic schedule recovers the canonical $T^{-1/2}$ statistical rate for

the last iterate, up to logarithmic factors and the lower-order T^{-1} term. For comparison, polynomial decay gives $T^{-a/2}$, while a constant step size reaches a noise floor of order $\sqrt{\eta}$. Choosing $c \asymp \mu^{-1}$ and treating all other problem parameters as fixed, the bound scales as $\sqrt{m/T}$ up to logarithmic factors in m and T. This matches the leading polynomial dependence on m and sample size of the model-based estimator (Cheng et al., 2026b).

4 Proof Sketch

The proof uses two stages because different information is available at different distances from the target. Far from $\boldsymbol{\theta}_m$, a first order Taylor approximation need not be accurate, but CDF monotonicity still shows that the mean update points towards the target. Stage I uses this global directional information to reach a neighborhood of $\boldsymbol{\theta}_m$. Once the iterate is close, Stage II uses a local Taylor expansion, leading to a linearized update map that is entrywise nonnegative and contractive. This keeps the iterate local and gives the sharp stochastic rate. The stages correspond to the two conditions in Theorem 3.1 respectively.

In the following, we first write QTD in vector form, which separates its mean update from the sampling noise. At iteration t, the synchronous batch contains one sampled (a, r, s') triplet for every state, denoted by $\mathcal{Z}_t = \{(a^{(t,s)}, r^{(t,s)}, s'^{(t,s)}) : s \in \mathcal{S}\}$. Set $\mathcal{F}_t = \sigma(\boldsymbol{\theta}^{(0)}, \mathcal{Z}_1, \dots, \mathcal{Z}_t)$, and let $F_{s,a}$ be the CDF of the reward distribution $\mathcal{P}_R(\cdot \mid s, a)$. Thus, $\mathcal{F}_t$ records the initialization and all samples observed through time t. The mean and sample update fields are

$$\begin{aligned} h_{s,i}(\boldsymbol{\theta}) &= \tfrac{1}{m} \sum\nolimits_{a\in\mathcal{A},\, s'\in\mathcal{S}} \sum\nolimits_{j=1}^{m} \pi(a \mid s) P(s' \mid s, a) F_{s,a}\big(\theta(s,i) - \gamma\theta(s', j)\big) - \tau_i, \\ \widehat{h}_{s,i}^{(t+1)}(\boldsymbol{\theta}) &= \tfrac{1}{m} \sum\nolimits_{j=1}^{m} \mathbf{1}\big\{r^{(t+1,s)} + \gamma\theta(s'^{(t+1,s)}, j) < \theta(s,i)\big\} - \tau_i. \end{aligned} \tag{7}$$

With $\boldsymbol{\xi}^{(t+1)} = \widehat{\boldsymbol{h}}^{(t+1)}(\boldsymbol{\theta}^{(t)}) - \boldsymbol{h}(\boldsymbol{\theta}^{(t)})$, the vector QTD recursion is

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \alpha_t\{\boldsymbol{h}(\boldsymbol{\theta}^{(t)}) + \boldsymbol{\xi}^{(t+1)}\}, \qquad \mathbb{E}[\boldsymbol{\xi}^{(t+1)} \mid \mathcal{F}_t] = \mathbf{0}. \tag{8}$$

Here $\boldsymbol{h}(\boldsymbol{\theta})$ is the average update direction and $\boldsymbol{h}(\boldsymbol{\theta}_m) = 0$, while $\boldsymbol{\xi}^{(t+1)}$ is the centered error of the direction computed from the current sample batch. Define $\boldsymbol{e}^{(t)} = \boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}_m$, then $\boldsymbol{h}(\boldsymbol{\theta}^{(t)}) = \boldsymbol{h}(\boldsymbol{\theta}_m + \boldsymbol{e}^{(t)})$ and the recursion can be studied directly in terms of the error vector $\boldsymbol{e}^{(t)}$.

4.1 Stage I: the global error decreases until the local region is reached

Outside the local region, Bellman structure still provides a global directional signal, though being non-smooth. In particular, if a signed coordinate $\varsigma e_{s,i} = \|\boldsymbol{e}\|_\infty$ for some $\varsigma \in \{-1, 1\}$, then

$$\varsigma\Big(\{\theta(s,i) - \gamma\theta(s',j)\} - \{\theta_m(s,i) - \gamma\theta_m(s',j)\}\Big) = \varsigma\{e_{s,i} - \gamma e_{s',j}\} \geq (1-\gamma)\|\boldsymbol{e}\|_\infty.$$

CDF monotonicity converts this gap into a drift towards the target $\boldsymbol{\theta}_m$. The maximizing coordinate, however, may change from one iteration to the next, so tracking any single coordinate does not yield a stable potential. We therefore aggregate all signed coordinate errors through the smooth maximum

$$\Phi_\beta(\boldsymbol{e}) = \beta^{-1} \log \sum\nolimits_{s\in\mathcal{S},\ i\in[m]} \left(e^{\beta e_{s,i}} + e^{-\beta e_{s,i}}\right).$$

For large β, Φ_β closely approximates $\|\boldsymbol{e}\|_\infty$ while remaining smooth for a one-step Taylor expansion. Lemma 4.1 makes the preceding inward-drift intuition quantitative.

Lemma 4.1 (Global drift of the smooth maximum)**.** *There exist some positive constants r_{out}, c_{g} and β (specified in Appendix C.1). For every $\boldsymbol{e}, \boldsymbol{u} \in \mathbb{R}^{\mathcal{S}\times[m]}$,*

$$\|\boldsymbol{e}\|_\infty \leq \Phi_\beta(\boldsymbol{e}) \leq \|\boldsymbol{e}\|_\infty + \frac{\log(2|\mathcal{S}|m)}{\beta}, \quad \|\nabla\Phi_\beta(\boldsymbol{e})\|_1 \leq 1, \quad \boldsymbol{u}^\top\nabla^2\Phi_\beta(\boldsymbol{e})\boldsymbol{u} \leq \beta\|\boldsymbol{u}\|_\infty^2.$$

Moreover, if $\boldsymbol{e} = \boldsymbol{\theta} - \boldsymbol{\theta}_m$ and $\|\boldsymbol{e}\|_\infty \geq r_{\mathrm{out}}/2$, then $\langle\nabla\Phi_\beta(\boldsymbol{e}), \boldsymbol{h}(\boldsymbol{\theta})\rangle \geq c_{\mathrm{g}} > 0$.

In Lemma 4.1, the first three bounds make Φ_β a smooth proxy for the maximum error, while the final inequality $\langle\nabla\Phi_\beta(\boldsymbol{e}), \boldsymbol{h}(\boldsymbol{\theta})\rangle \geq c_{\mathrm{g}}$ ensures a uniform first-order decrease outside the local ball.

As the expected QTD update is $-\alpha \boldsymbol{h}(\boldsymbol{\theta})$, a positive lower bound on $\langle \nabla \Phi_\beta(\boldsymbol{e}), \boldsymbol{h}(\boldsymbol{\theta}) \rangle$ ensures that this update decreases Φ_β to first order. We then show that this decrease dominates the accumulated sampling noise. Fix $0 \le u < v$ and define

$$A_{u,v} = \sum_{t=u}^{v-1} \alpha_t, \qquad V_{u,v} = \sum_{t=u}^{v-1} \alpha_t^2, \qquad \tau_{u,v} = \inf\{u \le t \le v : \|\boldsymbol{e}^{(t)}\|_\infty \le r_{\mathrm{out}}/2\}.$$

with $\inf \varnothing = \infty$. Here, $A_{u,v}$ measures the deterministic progress available during the step block $[u, v)$, whereas $V_{u,v}$ is the variance scale of its accumulated sampling noise. For $u \le t < v$ with $t < \tau_{u,v}$, set $Z_{t+1} = -\langle \nabla \Phi_\beta(\boldsymbol{e}^{(t)}), \boldsymbol{\xi}^{(t+1)} \rangle$. Then $\mathbb{E}[Z_{t+1} \mid \mathcal{F}_t] = 0$, $|Z_{t+1}| \le 1$, and Taylor's expansion with Lemma 4.1 gives

$$\Phi_\beta(\boldsymbol{e}^{(t+1)}) \le \Phi_\beta(\boldsymbol{e}^{(t)}) - c_{\mathrm{g}} \alpha_t + \alpha_t Z_{t+1} + \beta \alpha_t^2 / 2.$$

This is the desired global decay: the first negative term is *the inward progress*, while the last two terms are the sampling fluctuation and the second-order cost. If no entrance occurs by time v, we have $\Phi_\beta(\boldsymbol{e}^{(v)}) \ge \|\boldsymbol{e}^{(v)}\|_\infty > r_{\mathrm{out}}/2$, summing over the block forces $\sum_{t=u}^{v-1} \alpha_t Z_{t+1} \ge c_{\mathrm{g}} A_{u,v} - \big(\Phi_\beta(\boldsymbol{e}^{(u)}) - r_{\mathrm{out}}/2\big) - \frac{\beta}{2} V_{u,v}$. Azuma–Hoeffding thus gives the following entrance bound.

Lemma 4.2 (Late entrance)**.** *For any pair $0 \le u < v$,*

$$\mathbb{P}(\tau_{u,v} = \infty) \le \exp\big\{-\big[c_{\mathrm{g}} A_{u,v} - D_{\mathrm{g}} - (\beta/2) V_{u,v}\big]_+^2 / (2 V_{u,v})\big\},$$

where $[x]_+ := \max\{x, 0\}$ denotes the positive part of x, $D_{\mathrm{g}} > 0$ is defined in Appendix C.1.

This bound makes the role of Stage I explicit. The term $c_{\mathrm{g}} A_{u,v}$ is the inward progress supplied by the global drift, D_{g} bounds the initial distance in the smooth potential, and $\beta V_{u,v}/2$ is the second-order cost of using that potential. Once we have $c_{\mathrm{g}} A_{u,v} \ge D_{\mathrm{g}} + \frac{\beta}{2} V_{u,v} + \sqrt{2 V_{u,v} \log(1/\delta)}$, the iterate enters the local ball by time v with probability at least $1 - \delta$. This is the moving-block point emphasized after Corollary 3.2: no global summability condition on squared steps is needed.

4.2 Stage II: Local errors contract and recent noise determines the rate

Local linear structure. After entrance, a fixed drift towards the target is no longer enough to describe the rate: we must determine how quickly an earlier error is forgotten and how the noise from different iterations accumulates. This is where a local linearization of the mean field becomes useful. Write $d_{s,i} = p_{(\mathcal{T}^\pi \boldsymbol{\eta}_m)(s)}(\theta_m(s,i))$ for the Bellman-target density at a target quantile and set $\boldsymbol{D}_m = \mathrm{diag}_{s,i}(d_{s,i})$. We also introduce a nonnegative matrix $\boldsymbol{B}_m$, which records how successor-quantile perturbations change the current update and is given in Appendix C.4 due to space limits. The Jacobian of the mean field at $\boldsymbol{\theta}_m$ then becomes $\boldsymbol{G}_m = \boldsymbol{D}_m - \gamma \boldsymbol{B}_m$.

Lemma 4.3 (Local linear structure)**.** *$\boldsymbol{h}$ is continuously differentiable on the r_{out}-neighborhood of $\boldsymbol{\theta}_m$, with $\nabla \boldsymbol{h}(\boldsymbol{\theta}_m) = \boldsymbol{G}_m$ and $\boldsymbol{B}_m \mathbf{1} = \boldsymbol{D}_m \mathbf{1}$. If $0 \le \alpha C_0 \le 1$, where α denotes any step size, then $\boldsymbol{I} - \alpha \boldsymbol{G}_m \ge \mathbf{0}$ and $\|\boldsymbol{I} - \alpha \boldsymbol{G}_m\|_\infty \le 1 - \alpha\mu$, where $\mu := c_{\mathcal{M}}(1-\gamma)/(4m)$ is the local contraction rate. Moreover, whenever $\|\boldsymbol{e}\|_\infty \le r_{\mathrm{out}}$, $\boldsymbol{h}(\boldsymbol{\theta}_m + \boldsymbol{e}) = \boldsymbol{G}_m \boldsymbol{e} + \boldsymbol{R}(\boldsymbol{e})$, where $\|\boldsymbol{R}(\boldsymbol{e})\|_\infty \le L(1+\gamma)^2 \|\boldsymbol{e}\|_\infty^2 / 2$.*

Lemma 4.3 shows how the Jacobian $\boldsymbol{G}_m$ controls the local dynamics. The entrywise inequality $\boldsymbol{I} - \alpha \boldsymbol{G}_m \ge \mathbf{0}$ makes the propagation weights nonnegative, and the norm bound makes their total influence contract. The quadratic remainder $\|\boldsymbol{R}(\boldsymbol{e})\|_\infty$ is the small difference between the exact nonlinear update and this linear approximation.

To track the accumulated contraction effect, we define the linear propagation matrices $\boldsymbol{\Phi}_{r,k}$ by

$$\boldsymbol{\Phi}_{r,k} = \prod_{j=r}^{k-1} (\boldsymbol{I} - \alpha_j \boldsymbol{G}_m), \qquad \boldsymbol{\Phi}_{r,k} \ge \mathbf{0}, \qquad \|\boldsymbol{\Phi}_{r,k}\|_\infty \le e^{-\mu \sum_{j=r}^{k-1} \alpha_j}.$$

The two bounds serve different purposes. Nonnegativity allows each row to be used as nonnegative weights below, while contraction makes the influence of older errors decay over time. The matrix $\boldsymbol{\Phi}_{r,k}$ therefore tracks how a perturbation introduced at time r contributes to the error at time k.

Local error recursion. Let n be an entrance time with $\|\boldsymbol{e}^{(n)}\|_\infty \le r_{\mathrm{out}}/2$, and let σ be the first subsequent exit from the r_{out}-ball. Before σ, the update rule (8) as well as Lemma 4.3 gives

$$\boldsymbol{e}^{(t+1)} = (\boldsymbol{I} - \alpha_t \boldsymbol{G}_m) \boldsymbol{e}^{(t)} - \alpha_t \boldsymbol{\xi}^{(t+1)} - \alpha_t \boldsymbol{R}(\boldsymbol{e}^{(t)}). \tag{9}$$

For $n < k \le \sigma$, unrolling this recursion gives

$$\boldsymbol{e}^{(k)} = \underbrace{\boldsymbol{\Phi}_{n,k}\boldsymbol{e}^{(n)}}_{\text{entrance error}} - \underbrace{\textstyle\sum_{t=n}^{k-1}\alpha_t\boldsymbol{\Phi}_{t+1,k}\boldsymbol{\xi}^{(t+1)}}_{\text{sampling noise}} - \underbrace{\textstyle\sum_{t=n}^{k-1}\alpha_t\boldsymbol{\Phi}_{t+1,k}\boldsymbol{R}(\boldsymbol{e}^{(t)})}_{\text{nonlinear remainder}}. \tag{10}$$

Thus, (10) summarizes Stage II in one line: the error present at entrance is damped (due to the bound for $\|\boldsymbol{\Phi}_{r,k}\|_\infty$), new sampling noise is propagated through the same dynamics, and curvature contributes a quadratic correction. Writing $L_h = L(1+\gamma)^2/2$, contraction bounds the first term by $e^{-\mu\sum_{j=n}^{k-1}\alpha_j}\left\|\boldsymbol{e}^{(n)}\right\|_\infty$. The third is an exponentially weighted sum of quadratic local errors because $\|\boldsymbol{R}(\boldsymbol{e})\|_\infty \le L_h\|\boldsymbol{e}\|_\infty^2$, so it can be absorbed locally. The main difficulty is the sampling-noise term.

Controlling the sampling-noise term. Fix a coordinate (s,i) and times $n \le r < k \le T$. For $r \le t < k$, define $\boldsymbol{v}_t^\top = \boldsymbol{e}_{(s,i)}^\top\boldsymbol{\Phi}_{t+1,k}$ and $u_t = \boldsymbol{v}_t^\top\mathbf{1}$. Here, $\boldsymbol{e}_{(s,i)}$ selects the output coordinate, $\boldsymbol{v}_t$ describes how sampling noise at iteration t propagates to coordinate (s,i) at time k, and u_t is its total weight. Intuitively, bounding $\mathbb{E}[(\boldsymbol{v}_t^\top\boldsymbol{\xi}^{(t+1)})^2 \mid \mathcal{F}_t]$ through u_t directly would lead to a loose analysis. Instead, we relate it to the actual one-step reduction $u_t - u_{t-1}$, whose sum over t telescopes. The next lemma introduces the variance–drift matching property, and establishes this relation by showing that the variance and weight reduction share the same density-weighted factor.

Lemma 4.4 (Variance–drift matching)**.** *Suppose* $\left\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}_m\right\|_\infty \le r_{\text{out}}$. *For any nonnegative* $\boldsymbol{v} \ge 0$,

$$\boldsymbol{v}^\top\mathbf{1} - \boldsymbol{v}^\top(\boldsymbol{I} - \alpha_t\boldsymbol{G}_m)\mathbf{1} = (1-\gamma)\alpha_t\boldsymbol{v}^\top\boldsymbol{D}_m\mathbf{1}, \tag{11a}$$

$$\mathbb{E}[(\boldsymbol{v}^\top\boldsymbol{\xi}^{(t+1)})^2 \mid \mathcal{F}_t] \le 2c_{\mathcal{M}}^{-1}(\boldsymbol{v}^\top\mathbf{1})(\boldsymbol{v}^\top\boldsymbol{D}_m\mathbf{1}). \tag{11b}$$

The common factor $\boldsymbol{v}^\top\boldsymbol{D}_m\mathbf{1}$ is the key. It determines both the one-step loss of propagation weight in (11a) and the conditional noise variance in (11b). Consequently, a direction $\boldsymbol{v}_t$ with smaller restoring drift (i.e., $u_t - u_{t-1}$) would contribute less conditional stochastic variation in $\boldsymbol{v}_t^\top\boldsymbol{\xi}^{(t+1)}$ given $\mathcal{F}_t$. Indeed, applying the lemma with $\boldsymbol{v} = \boldsymbol{v}_t$ and using step-size monotonicity gives

$$\sum_{t=r}^{k-1}\alpha_t^2\mathbb{E}[(\boldsymbol{v}_t^\top\boldsymbol{\xi}^{(t+1)})^2 \mid \mathcal{F}_t] \le \big(2\alpha_r/[c_{\mathcal{M}}(1-\gamma)]\big)\sum_{t=r}^{k-1}u_t(u_t - u_{t-1}) \le 2\alpha_r/[c_{\mathcal{M}}(1-\gamma)],$$

where the last inequality follows from $0 \le u_t \le 1$ and the telescoping of $u_t - u_{t-1}$. Thus, the accumulated variance is controlled by the actual loss of propagation weight.

The argument above controls one coordinate (s,i) and one time interval $[r,k)$. Lemma D.3 in the appendix and a union bound give, conditionally on $\mathcal{F}_n$ with probability at least $1-\delta$,

$$\max_{n\le r<k\le T}\Big\|\sum_{t=r}^{k-1}\alpha_t\boldsymbol{\Phi}_{t+1,k}\mathbf{1}\{t<\sigma\}\boldsymbol{\xi}^{(t+1)}\Big\|_\infty \le b_{n,T}(\delta). \tag{12}$$

The appendix gives the explicit logarithmic factors in $b_{n,T}(\delta)$, whose leading scale is $\sqrt{\alpha_n/[c_{\mathcal{M}}(1-\gamma)]}$, with a lower-order term linear in α_n. The indicator halts the sum at a possible exit, ensuring that the local assumptions hold. For $k \le \sigma$, it equals one throughout the sum, which is then exactly the sampling-noise term in (10).

Staying inside the local ball. We now show that the uniform noise bound prevents the iterate from leaving the local region. Stage I reaches the ball of radius $r_{\text{out}}/2$, while the local expansion remains valid up to radius r_{out}. On the event (12), if the first exit occurred at some $\sigma \le T$, then (10) with $k = \sigma$, together with contraction and the quadratic remainder bound in Lemma 4.3, would imply $\|\boldsymbol{e}^{(\sigma)}\|_\infty < r_{\text{out}}$, contradicting the definition of σ. The next lemma formalizes this argument.

Lemma 4.5 (Local capture)**.** *Let* $n \le T$ *be an* $(\mathcal{F}_t)$*-stopping time satisfying* $\left\|\boldsymbol{\theta}^{(n)} - \boldsymbol{\theta}_m\right\|_\infty \le r_{\text{out}}/2$. *If* $b_{n,T}(\delta) \le \min\{r_{\text{out}}/16, \mu/(64L_h)\}$, *where* $L_h = L(1+\gamma)^2/2$ *and* $\mu/L_h = +\infty$ *when* $L_h = 0$, *then, conditionally on* $\mathcal{F}_n$, *with probability at least* $1-\delta$, *the iterates remain in the* r_{out}*-neighborhood through time* T, *and*

$$\|\boldsymbol{\theta}^{(T)} - \boldsymbol{\theta}_m\|_\infty \le \|\boldsymbol{\theta}^{(n)} - \boldsymbol{\theta}_m\|_\infty e^{-(15/16)\mu\sum_{t=n}^{T-1}\alpha_t} + b_{n,T}(\delta) + (32L_h/(15\mu))b_{n,T}^2(\delta). \tag{13}$$

This completes Stage II. The three terms on the right-hand side of (13) are the contracted entrance error, the accumulated sampling noise $b_{n,T}(\delta)$, and the quadratic nonlinear correction of order $b_{n,T}^2(\delta)$, respectively. Once $b_{n,T}(\delta)$ is sufficiently small, the nonlinear correction is lower order, and the first-exit argument ensures that the iterates remain in the local region.

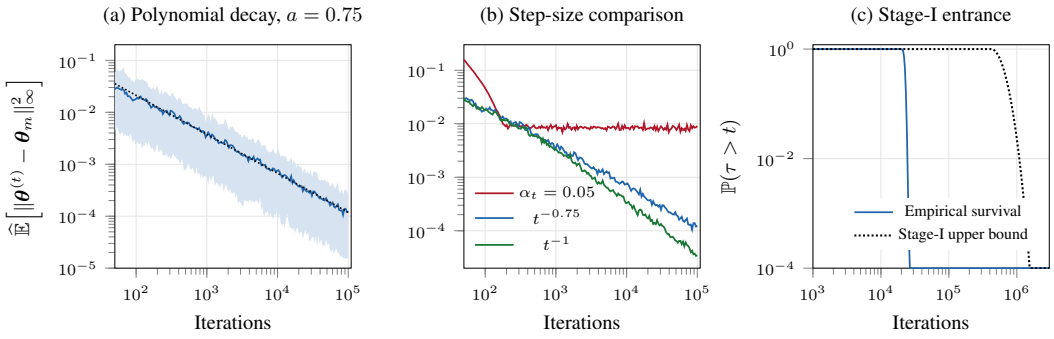

Figure 1: Numerical verification of the finite-time theory. (a) Mean squared sup-norm error for $a = 0.75$, with the 10th–90th percentile band and a fixed-slope reference. (b) Comparison of the three step-size classes. (c) Empirical Stage-I non-entrance probability and the bound from Lemma 4.2.

4.3 Completing the proof of Theorem 3.1

Now, we are ready for the final proof. Set $u = \lfloor T/4 \rfloor$ and $v = \lfloor T/2 \rfloor$, where the first quarter is a warm-up, the next quarter is the entrance window, and the final half is the local window. The entrance condition (3a) gives an entrance time $n \in [u, v]$ with failure probability at most $\delta/2$. Step-size monotonicity gives $b_{n,T}(\delta/2) \le b_T(\delta)$, so the capture condition (3b) allows us to apply Lemma 4.5 from time n onward. Step-size monotonicity further bounds the accumulated local step size from below by $\sum_{t=n}^{T-1} \alpha_t \ge A_T^{\mathrm{loc}}$, and the quadratic remainder is absorbed into the noise term. Consequently, for some constant $C > 0$, we have $\left\|\boldsymbol{e}^{(T)}\right\|_\infty \le C \exp\{-C^{-1}\mu A_T^{\mathrm{loc}}\} + C b_T(\delta)$. Finally, a union bound yields (4) with probability at least $1 - \delta$, completing the proof. The explicit constants and remaining details, as well as the proofs of the corollaries, are given in Appendix C.7.

5 Numerical Verification

Figure 1 provides three numerical checks of our finite-time theory. Panels (a)–(b) use a one-state, one-action MDP with uniform rewards, $\gamma = 0.5$, $m = 7$, and 200 independent trajectories up to $T = 10^5$. For $\alpha_t = 4/(t + 20)^{0.75}$, Panel (a) compares the mean squared error with a reference line whose slope is fixed at the predicted value -0.75; the two curves are nearly parallel. Panel (b) compares constant, polynomial, and harmonic step sizes. The constant schedule reaches a noise floor, whereas the decreasing schedules continue to improve, with the harmonic schedule performing best at long horizons, consistently with its $T^{-1/2}$ raw-error rate. Panel (c) checks the Stage-I entrance bound in Lemma 4.2. The empirical non-entrance probability stays below the initialization-specific upper bound; the gap reflects the conservativeness of the theoretical bounds. Appendix E gives the full protocol, additional polynomial exponents, and reproducibility details.

6 Conclusion

We have established a global high-probability last-iterate bound for synchronous tabular QTD under general nonincreasing step sizes. Constant steps forget the initialization geometrically before reaching a noise floor, polynomial decay gives the rate $T^{-a/2}$ up to logarithmic factors, and a tuned harmonic schedule recovers the canonical $T^{-1/2}$ rate. In particular, the tuned harmonic schedule yields a W_∞ error scaling as $\sqrt{m/T}$ up to logarithmic factors in m and T, matching the model-based statistical rate. The proof cleanly separates global and local roles: order monotonicity and Bellman contraction localize the iterates, after which a positive, contractive linearization and variance–drift matching yield the sharp stochastic fluctuation.

AI USE STATEMENT

Generative AI was used only for language editing and structural reorganization, not for mathematical claims or proofs. The authors verified the resulting text and take responsibility for it.

REPRODUCIBILITY STATEMENT

Sections 2–4 state the algorithm, assumptions, results, and proof strategy; the appendix gives complete proofs. Code, seed, checkpoint data, and reference-line fits for Section 5 are included with the source.

REFERENCES

R Raj Bahadur. A note on quantiles in large samples. *The Annals of Mathematical Statistics*, 37(3): 577–580, 1966. doi: 10.1214/aoms/1177699450.

Gabriel Barth-Maron, Matthew W Hoffman, David Budden, Will Dabney, Dan Horgan, T. B. Dhruva, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. Distributed distributional deterministic policy gradients. In *International Conference on Learning Representations*, 2018.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pp. 449–458. PMLR, 2017.

Marc G Bellemare, Salvatore Candido, Pablo Samuel Castro, Jun Gong, Marlos C Machado, Subhodeep Moitra, Sameera S Ponda, and Ziyu Wang. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836):77–82, 2020.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. Adaptive Computation and Machine Learning. The MIT Press, 2023. doi: 10.7551/mitpress/14207.001.0001.

Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.

Cristian Bodnar, Adrian Li, Karol Hausman, Peter Pastor, and Mrinal Kalakrishnan. Quantile Qt-Opt for risk-aware vision-based robotic grasping. In *Proceedings of Robotics: Science and Systems*, 2020. doi: 10.15607/RSS.2020.XVI.075.

Likai Chen, Georg Keilbar, and Wei Biao Wu. Recursive quantile estimation: Non-asymptotic confidence bounds. *Journal of Machine Learning Research*, 24(91):1–25, 2023.

Zijie Cheng, Yang Peng, and Zhihua Zhang. Online inference for quantile temporal difference learning in distributional reinforcement learning. *arXiv preprint arXiv:2608.12973*, 2026a.

Zijie Cheng, Yang Peng, and Zhihua Zhang. Statistical efficiency and inference of quantile distributional reinforcement learning. *arXiv preprint arXiv:2607.08444*, 2026b.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pp. 1096–1105. PMLR, 2018a.

Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1):2892–2901, 2018b. doi: 10.1609/aaai.v32i1.11791.

Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pp. 1579–1613. PMLR, 2019.

Jack Kiefer. On Bahadur's representation of sample quantiles. *The Annals of Mathematical Statistics*, 38(5):1323–1342, 1967. doi: 10.1214/aoms/1177698690.

Keith Knight. Limiting distributions for L_1 regression estimators under general conditions. *The Annals of Statistics*, 26(2):755–770, 1998. doi: 10.1214/aos/1028144858.

Roger Koenker. *Quantile Regression*, volume 38 of *Econometric Society Monographs*. Cambridge University Press, 2005. doi: 10.1017/CBO9780511754098.

Roger Koenker and Gilbert Bassett, Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. doi: 10.2307/1913643.

Harold J Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*. Springer, 2003.

Yang Peng, Liangyu Zhang, and Zhihua Zhang. Statistical efficiency of distributional temporal difference learning. *Advances in Neural Information Processing Systems*, 37:24724–24761, 2024.

Yang Peng, Kaicheng Jin, Liangyu Zhang, and Zhihua Zhang. A finite sample analysis of distributional TD learning with linear function approximation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5687.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. doi: 10.1137/0330046.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. doi: 10.1214/aoms/1177729586.

Mark Rowland, Marc Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 29–37. PMLR, 2018.

Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning*, pp. 5528–5536. PMLR, 2019.

Mark Rowland, Yunhao Tang, Clare Lyle, Rémi Munos, Marc G Bellemare, and Will Dabney. The statistical benefits of quantile temporal-difference learning for value estimation. In *International Conference on Machine Learning*, pp. 29210–29231. PMLR, 2023.

Mark Rowland, Rémi Munos, Mohammad Gheshlaghi Azar, Yunhao Tang, Georg Ostrovski, Anna Harutyunyan, Karl Tuyls, Marc G Bellemare, and Will Dabney. An analysis of quantile temporal-difference learning. *Journal of Machine Learning Research*, 25(163):1–47, 2024. URL https://jmlr.org/papers/v25/23-0154.html.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, pp. 71–79. PMLR, 2013.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Luke Tierney. A space-efficient recursive procedure for estimating a quantile of an unknown distribution. *SIAM Journal on Scientific and Statistical Computing*, 4(4):706–711, 1983. doi: 10.1137/0904048.

Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. Fully parameterized quantile function for distributional reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

Fan Zhou, Jianing Wang, and Xingdong Feng. Non-crossing quantile regression for distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 33, 2020.

A Supplementary Background and Related Work

Nonsmooth SGD and QTD. For a fixed scalar response Y, a τ-quantile minimizes the expected pinball loss

$$q \longmapsto \mathbb{E}\big[\rho_\tau(Y-q)\big], \qquad \rho_\tau(u) = u\big(\tau - \mathbf{1}\{u<0\}\big).$$

A sample subgradient with respect to q is $\mathbf{1}\{Y<q\}-\tau$; under our continuous-density assumptions, averaging gives the CDF mismatch $F_Y(q)-\tau$. QTD steps in the negative of this direction. Nonsmooth SGD theory provides averaged- and last-iterate guarantees for fixed convex objectives under several regimes (Shamir & Zhang, 2013; Harvey et al., 2019).

The important difference is that QTD does not repeatedly sample from one fixed response distribution. To make this distinction explicit, freeze a quantile array $\boldsymbol{\vartheta}$ and, for state s, let

$$Y_{\boldsymbol{\vartheta}}(s) = R + \gamma\vartheta(S', J), \qquad J \sim \mathrm{Unif}([m]),$$

where the action, reward, and next state are sampled according to the policy and MDP, and J is independent of this transition. For fixed $\boldsymbol{\vartheta}$, averaging over j is the conditional expectation over J, so the mean increment at coordinate (s,i) is the negative of a subgradient in q of

$$L_{s,i}(q;\boldsymbol{\vartheta}) = \mathbb{E}\big[\rho_{\tau_i}(Y_{\boldsymbol{\vartheta}}(s)-q)\big].$$

The cited nonsmooth-SGD analyses repeatedly use noisy subgradients of one fixed convex objective. QTD instead takes this negative-subgradient step at $q=\theta(s,i)$ while simultaneously setting $\boldsymbol{\vartheta}=\boldsymbol{\theta}$ and updating the full array. It varies the prediction coordinate while treating the target as fixed, even though the target distribution was generated by the current array. This semi-gradient recursion is generally not SGD on one fixed convex objective, so fixed-objective bounds do not directly control distance to the coupled quantile-projected Bellman fixed point. Our global localization and local stability arguments handle this feedback.

Classical quantile estimation and regression. Quantiles admit two complementary descriptions. Minimizers of expected pinball loss satisfy $F(q^-) \le \tau \le F(q)$; at a regular quantile with a continuous, strictly increasing CDF, this reduces to $F(q)-\tau=0$. The loss view underlies regression quantiles and modern quantile regression (Koenker & Bassett, 1978; Koenker, 2005). The estimating-equation view underlies the local expansions used in quantile asymptotics. Bahadur and Kiefer representations quantify how empirical CDF fluctuations translate into sample-quantile errors (Bahadur, 1966; Kiefer, 1967). Knight's identity (Knight, 1998) is a standard tool for expanding the nonsmooth quantile-regression objective: it separates a pinball-loss difference into a linear score and an integral remainder.

The same conversion is visible in our local analysis. A positive density near a target quantile turns a small location error into a proportional CDF error, which supplies the restoring drift of the QTD recursion. Unlike ordinary quantile regression, however, that CDF is a Bellman-target CDF assembled from all successor-state quantiles. The local derivative is therefore a coupled matrix rather than a collection of independent scalar slopes.

Recursive quantile estimation and stochastic approximation. The classical Robbins–Monro procedure estimates the solution of a mean equation from noisy observations and provides the basic stochastic approximation template for online quantile estimation (Robbins & Monro, 1951). Recursive procedures tailored to quantiles include the space-efficient estimator of Tierney (1983); more recent work gives nonasymptotic confidence bounds for recursive quantile estimates (Chen et al., 2023). General stochastic approximation theory studies stability, limiting dynamics, and averaged iterates under broad conditions (Kushner & Yin, 2003; Benaïm et al., 2005; Polyak & Juditsky, 1992). These results provide scalar and asymptotic analogues for QTD. Ordinary recursive quantile estimation samples repeatedly from a fixed response distribution, whereas QTD changes its Bellman-target distribution with the current array. We additionally need global finite-time control of this coupled recursion and a high-probability last-iterate rate. Our two stages separate these tasks: global drift establishes entrance, and variance-sensitive local analysis gives the rate afterward.

Most directly, Rowland et al. (2024) established almost-sure asymptotic convergence of tabular QTD, while Cheng et al. (2026a) developed functional limit theorems and online inference for QTD. Our question is complementary: we seek a global high-probability finite-time bound for the last iterate from arbitrary initialization and identify the variance-sensitive mechanism that governs its local fluctuations.

Quantile representations in distributional reinforcement learning. Quantile regression distributional RL introduced a fixed grid of learned quantile locations as an alternative to categorical value distributions (Dabney et al., 2018b). Subsequent work developed implicit quantile networks, fully parameterized quantile functions, and non-crossing architectures for more flexible representations (Dabney et al., 2018a; Yang et al., 2019; Zhou et al., 2020). The broader link between distributional statistics, finite collections of learned statistics, and imputation strategies, which reconstruct a distribution from those statistics, was studied by Rowland et al. (2019). Much of this line emphasizes richer representations and function-approximation algorithms. Our fixed-grid tabular setting lets us isolate how Bellman coupling, global localization, and variance–drift matching interact in finite time. Complementary finite-sample analyses cover nonparametric and categorical distributional TD, including linear function approximation (Peng et al., 2024; 2025); these representations have update structures different from the coupled quantile-score recursion studied here.

For the same quantile-projected target, Cheng et al. (2026b, Theorems 3.1–3.2) give a model-based $\widetilde{O}(\sqrt{m/n})\,\overline{W}_\infty$ estimation bound and a modulus-of-continuity approximation bound to the exact return law. Their estimator forms an empirical model from n independent samples per state–action pair, or $n|\mathcal{S}||\mathcal{A}|$ samples in total. Our synchronous QTD uses one fresh transition per state per iteration, or $T|\mathcal{S}|$ samples over T iterations. With gain parameter $c \asymp \mu^{-1}$, under the burn-in condition, Corollary 3.3 matches their leading polynomial dependence on quantile resolution and the respective sample counts in the statistical rate.

More explicitly, their approximation bound is

$$\overline{W}_\infty(\boldsymbol{\eta}_m, \boldsymbol{\eta}^\pi) \le \frac{1}{1-\gamma} \max_{s\in\mathcal{S}} \omega_s\left(\frac{1}{2m}\right), \qquad \omega_s(\rho) = \sup_{\substack{u,v\in(0,1)\\ |u-v|\le\rho}} \left|F^{-1}_{\eta^\pi(s)}(u) - F^{-1}_{\eta^\pi(s)}(v)\right|.$$

Under the reward assumptions here, the return quantile functions extend continuously to the endpoints and hence are uniformly continuous. Therefore $\max_s \omega_s(1/(2m)) \to 0$. Combining this existing approximation bound with our statistical rates bounds the total $\overline{W}_\infty$ error to $\boldsymbol{\eta}^\pi$.

B Additional Preliminaries

This appendix records background definitions that are useful for locating the paper within distributional reinforcement learning but are not needed to state the main theorem.

Let $\mathscr{P}$ be the space of probability measures on $\mathbb{R}$. For $p \in [1,\infty)$ and $\mu, \nu \in \mathscr{P}$ with finite pth moments, their p-Wasserstein distance is

$$W_p(\mu,\nu) = \left(\int_0^1 \left|F^{-1}_\mu(u) - F^{-1}_\nu(u)\right|^p \,\mathrm{d}u\right)^{1/p}.$$

For any (possibly extended) metric d on $\mathscr{P}$, its statewise supremum extension is

$$\overline{d}(\boldsymbol{\mu},\boldsymbol{\nu}) = \sup_{s\in\mathcal{S}} d(\mu(s),\nu(s)).$$

The main text uses the case $d = W_\infty$ because both the distributional Bellman operator and the midpoint quantile projection are stable in that metric.

For a measurable map $g : \mathbb{R} \to \mathbb{R}$, the pushforward $g_\#\mu$ is defined by $(g_\#\mu)(A) = \mu(g^{-1}(A))$ for every Borel set A. Therefore, the reward-mixture term in the distributional Bellman operator is interpreted as the measure satisfying

$$\left[\int (b_{r,\gamma})_\#\eta(s')\,\mathcal{P}_R(\mathrm{d}r \mid s,a)\right](A) = \int (b_{r,\gamma})_\#\eta(s')(A)\,\mathcal{P}_R(\mathrm{d}r \mid s,a).$$

This identity makes explicit that the integral in Equation (1) is a mixture of probability measures rather than an ordinary scalar integral.

C Proofs for Sections 3 and 4

C.1 Explicit constants for the main results

This subsection records the proof-level constants suppressed from Sections 3 and 4. First define the distance between the fixed-point density arguments and the endpoints of the reward support by

$$\Delta_m = \min_{\substack{P^\pi(s'|s)>0 \\ i,j\in[m]}} \min\left\{\left|\theta_m(s,i) - \gamma\theta_m(s',j)\right|, \left|\theta_m(s,i) - \gamma\theta_m(s',j) - 1\right|\right\} > 0,$$

where $P^\pi(s' \mid s) = \sum_{a\in\mathcal{A}} \pi(a \mid s)P(s' \mid s, a)$. Set

$$r_{\text{out}} = \begin{cases} \min\left\{\dfrac{\Delta_m}{2(1+\gamma)}, \dfrac{1}{8mC_0(1+\gamma)}, \dfrac{c_{\mathcal{M}}(1-\gamma)}{32mL(1+\gamma)^2}\right\}, & \text{under Assumption 1(i),} \\ \min\left\{\dfrac{1}{8mC_0(1+\gamma)}, \dfrac{c_{\mathcal{M}}(1-\gamma)}{32mL(1+\gamma)^2}\right\}, & \text{under Assumption 1(ii),} \end{cases}$$

where a fraction with denominator $L = 0$ is interpreted as $+\infty$. Define the local contraction and nonlinear-remainder scales by

$$\mu = \frac{c_{\mathcal{M}}(1-\gamma)}{4m}, \qquad L_h = \frac{L(1+\gamma)^2}{2},$$

Since the step sizes are nonincreasing, write $\bar{\alpha} = \sup_{t\geq 0} \alpha_t = \alpha_0$ for their largest value. Define the global-drift constants by

$$\begin{aligned} c_{\text{g}} &= \frac{c_{\mathcal{M}}\tau_1(1-\tau_1)(1-\gamma)r_{\text{out}}}{8}, \\ \beta &= \frac{4}{(1-\gamma)r_{\text{out}}} \log\left[\frac{2|\mathcal{S}|m(1+2c_{\text{g}})}{c_{\text{g}}}\right], \\ D_{\text{g}} &= \frac{1+\bar{\alpha}}{1-\gamma} - \frac{r_{\text{out}}}{2} + \frac{\log(2|\mathcal{S}|m)}{\beta}. \end{aligned}$$

These global-drift constants are used only in the entrance condition. Local capture instead requires $b_T(\delta) \leq \min\{r_{\text{out}}/16, \mu/(64L_h)\}$, where $\mu/L_h = +\infty$ when $L_h = 0$. For use in the proofs below, also write $u_T = \lfloor T/4 \rfloor$ and $v_T = \lfloor T/2 \rfloor$, and define

$$A_T^{\text{ent}} = \sum_{t=u_T}^{v_T-1} \alpha_t, \qquad V_T^{\text{ent}} = \sum_{t=u_T}^{v_T-1} \alpha_t^2, \qquad A_T^{\text{loc}} = \sum_{t=v_T}^{T-1} \alpha_t.$$

Also define

$$\Gamma_T = c_{\text{g}}A_T^{\text{ent}} - D_{\text{g}} - \frac{\beta}{2}V_T^{\text{ent}}, \qquad b_T(\delta) = 2\sqrt{\frac{\alpha_{u_T}\ell_T(\delta)}{c_{\mathcal{M}}(1-\gamma)}} + \frac{\alpha_{u_T}\ell_T(\delta)}{3}.$$

For Corollary 3.3, one admissible universal choice is

$$C_{\text{harm}} = \max\left\{8, \frac{16\log 4}{15\log(4/3)}\right\}.$$

Then $c \geq C_{\text{harm}}/\mu$ ensures both $\exp\{-15\mu c\log(4/3)/16\} \leq 1/4$ and $\mu c/6 \geq 4/3$. The latter inequality is used only in the sharper harmonic entrance argument.

C.2 Proof of Lemma 4.1: Global drift

Proof. Put $V \coloneqq \|\boldsymbol{e}\|_\infty$ for the largest coordinate error. The bounds in the standard log-sum-exp inequalities give

$$\|\boldsymbol{e}\|_\infty \leq \Phi_\beta(\boldsymbol{e}) \leq \|\boldsymbol{e}\|_\infty + \frac{\log(2|\mathcal{S}|m)}{\beta}.$$

For $(s,i) \in \mathcal{S} \times [m]$ and $\varsigma \in \{-1,1\}$, define the softmax weight

$$w_{s,i,\varsigma}(\boldsymbol{e}) \coloneqq \frac{\exp(\beta \varsigma e_{s,i})}{\sum_{(s',j)\in\mathcal{S}\times[m]} \{\exp(\beta e_{s',j}) + \exp(-\beta e_{s',j})\}}.$$

If X is the auxiliary random signed basis vector that equals $\varsigma \boldsymbol{e}_{s,i}$ with probability $w_{s,i,\varsigma}(\boldsymbol{e})$, then

$$\nabla \Phi_\beta(\boldsymbol{e}) = \mathbb{E}[X], \qquad \nabla^2 \Phi_\beta(\boldsymbol{e}) = \beta \mathsf{Cov}(X).$$

Since $\|X\|_1 = 1$ and $\left|\boldsymbol{u}^\top X\right| \le \|\boldsymbol{u}\|_\infty$, it follows that

$$\|\nabla \Phi_\beta(\boldsymbol{e})\|_1 \le 1, \qquad \boldsymbol{u}^\top \nabla^2 \Phi_\beta(\boldsymbol{e}) \boldsymbol{u} = \beta \mathsf{Var}(\boldsymbol{u}^\top X) \le \beta \|\boldsymbol{u}\|_\infty^2 .$$

It remains to prove the drift bound. Let $\mathcal{I}(\boldsymbol{e})$ be the set of signed coordinates whose errors are close to the maximum:

$$\mathcal{I}(\boldsymbol{e}) \coloneqq \left\{ (s,i,\varsigma) : \varsigma e_{s,i} \ge V - \frac{(1-\gamma) r_{\mathrm{out}}}{4} \right\}.$$

If $\varsigma = 1$, then for every (s',j),

$$e_{s,i} - \gamma e_{s',j} \ge (1-\gamma)V - \frac{(1-\gamma) r_{\mathrm{out}}}{4} \ge \frac{(1-\gamma)V}{2}.$$

We denote

$$V^* = \min\left\{ V, \frac{1+\bar{\alpha}}{1-\gamma} \right\}$$

and monotonicity of the reward CDFs and the definition of $c_{\mathcal{M}}$ give

$$\begin{aligned} h_{s,i}(\boldsymbol{\theta}) &\ge F_{(\mathcal{T}^\pi \boldsymbol{\eta}_m)(s)} \left(\theta_m(s,i) + \frac{(1-\gamma)V^*}{2} \right) - \tau_i \\ &\ge \frac{c_{\mathcal{M}} \tau_1 (1-\tau_1)(1-\gamma)V^*}{2} \coloneqq a_m V^* \end{aligned}$$

The same argument with all inequalities reversed shows that $-h_{s,i}(\boldsymbol{\theta}) \ge a_m V^*$ when $\varsigma = -1$.

For a signed coordinate outside $\mathcal{I}(\boldsymbol{e})$, its exponential weight is at most $\exp\{-\beta(1-\gamma) r_{\mathrm{out}}/4\}$ times the weight of a maximizing signed coordinate. Hence

$$w \coloneqq \sum_{(s,i,\varsigma) \notin \mathcal{I}(\boldsymbol{e})} w_{s,i,\varsigma}(\boldsymbol{e}) \le 2|\mathcal{S}| m \exp\left\{ -\frac{\beta(1-\gamma) r_{\mathrm{out}}}{4} \right\} = \frac{c_{\mathrm{g}}}{1 + 2c_{\mathrm{g}}}.$$

Every signed mean-field coordinate has magnitude at most one. The coordinates in $\mathcal{I}(\boldsymbol{e})$ contribute at least $a_m V^*$ per unit softmax weight, whereas all remaining coordinates contribute at least -1. Moreover, since $a_m V^* \ge 2c_{\mathrm{g}}$,

$$w \le \frac{a_m V^*}{2(1 + a_m V^*)}.$$

Therefore,

$$\langle \nabla \Phi_\beta(\boldsymbol{e}), \boldsymbol{h}(\boldsymbol{\theta}) \rangle \ge a_m V^* (1-w) - w \ge \frac{a_m V^*}{2}, \tag{14}$$

Since $a_m V^* \ge 2c_{\mathrm{g}}$, we have actually proved a stronger version of the drift bound stated in Lemma 4.1. □

C.3 Proof of Lemma 4.2: Late entrance

Proof. Let $\boldsymbol{e}^{(t)}$ denote the error at time t, and fix the block endpoints $u < v$ from Lemma 4.2. Set

$$\boldsymbol{e}^{(t)} \coloneqq \boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}_m.$$

For integers $r < q$, write $A_{r,q}$ and $V_{r,q}$ for the accumulated step size and accumulated squared step size, respectively:

$$A_{r,q} \coloneqq \sum_{t=r}^{q-1} \alpha_t, \qquad V_{r,q} \coloneqq \sum_{t=r}^{q-1} \alpha_t^2.$$

For the global comparison argument, use the drift scale c_{g} and smoothing parameter β defined in Appendix C.1. We use the smooth maximum

$$\Phi_\beta(\boldsymbol{e}) \coloneqq \frac{1}{\beta} \log \sum_{s\in\mathcal{S},\, i\in[m]} \left(e^{\beta e_{s,i}} + e^{-\beta e_{s,i}}\right).$$

Lemma 4.1 gives

$$\|\boldsymbol{e}\|_\infty \le \Phi_\beta(\boldsymbol{e}) \le \|\boldsymbol{e}\|_\infty + \frac{\log(2|\mathcal{S}|m)}{\beta}, \tag{15}$$

$$\|\nabla\Phi_\beta(\boldsymbol{e})\|_1 \le 1, \qquad \boldsymbol{u}^\top \nabla^2\Phi_\beta(\boldsymbol{e})\boldsymbol{u} \le \beta \|\boldsymbol{u}\|_\infty^2, \tag{16}$$

and, whenever $\|\boldsymbol{\theta} - \boldsymbol{\theta}_m\|_\infty \ge r_{\mathrm{out}}/2$,

$$\langle \nabla\Phi_\beta(\boldsymbol{\theta} - \boldsymbol{\theta}_m), \boldsymbol{h}(\boldsymbol{\theta})\rangle \ge c_{\mathrm{g}}. \tag{17}$$

For the sharper conditional form used in the Stage-I numerical diagnostic, define the realized budget at the left endpoint of the block by

$$D_{u,\beta} \coloneqq \Phi_\beta(\boldsymbol{e}^{(u)}) - \frac{r_{\mathrm{out}}}{2}.$$

This quantity is $\mathcal{F}_u$-measurable.

For $t \ge u$, define the centered one-step fluctuation

$$Z_{t+1} \coloneqq -\left\langle \nabla\Phi_\beta(\boldsymbol{e}^{(t)}), \boldsymbol{\xi}^{(t+1)} \right\rangle.$$

Then $\mathbb{E}[Z_{t+1} \mid \mathcal{F}_t] = 0$ and $|Z_{t+1}| \le 1$. Indeed, every coordinate of $\boldsymbol{\xi}^{(t+1)}$ is the difference of two numbers in $[0, 1]$. Moreover,

$$\boldsymbol{h}(\boldsymbol{\theta}^{(t)}) + \boldsymbol{\xi}^{(t+1)} = \widehat{\boldsymbol{h}}^{(t+1)}(\boldsymbol{\theta}^{(t)}), \qquad \left\|\widehat{\boldsymbol{h}}^{(t+1)}(\boldsymbol{\theta}^{(t)})\right\|_\infty \le 1.$$

Thus, on $\{t < \tau_{u,v}\}$, Taylor's theorem, the QTD recursion, and Equations (16)–(17) imply

$$\Phi_\beta(\boldsymbol{e}^{(t+1)}) \le \Phi_\beta(\boldsymbol{e}^{(t)}) - c_{\mathrm{g}}\alpha_t + \alpha_t Z_{t+1} + \frac{\beta}{2}\alpha_t^2. \tag{18}$$

Define the corresponding stopped process by

$$M_k \coloneqq \sum_{t=u}^{k-1} \alpha_t \mathbf{1}\{t < \tau_{u,v}\} Z_{t+1}, \qquad u \le k \le v,$$

which is a martingale because $\mathbf{1}\{t < \tau_{u,v}\}$ is $\mathcal{F}_t$-measurable; its increments are bounded by α_t. On $\{\tau_{u,v} = \infty\}$, summing Equation (18) from u to $v-1$ and using $\Phi_\beta(\boldsymbol{e}^{(v)}) > r_{\mathrm{out}}/2$ yields

$$M_v \ge c_{\mathrm{g}} A_{u,v} - D_{u,\beta} - \frac{\beta}{2} V_{u,v}.$$

Conditionally on $\mathcal{F}_u$, the Azuma–Hoeffding inequality therefore gives

$$\mathbb{P}(\tau_{u,v} = \infty \mid \mathcal{F}_u) \le \exp\left\{ -\frac{\left(c_{\mathrm{g}} A_{u,v} - D_{u,\beta} - \frac{\beta}{2} V_{u,v}\right)_+^2}{2V_{u,v}} \right\}. \tag{19}$$

By Lemma D.1 and Lemma 4.1, we know that $D_{u,\beta} \le D_{\mathrm{g}}$. Therefore, the right-hand side of Equation (19) is at most the same expression with $D_{u,\beta}$ replaced by D_{g}. Taking expectations with respect to $\mathcal{F}_u$ therefore yields

$$\mathbb{P}(\tau_{u,v} = \infty) \le \exp\left\{ -\frac{\left(c_{\mathrm{g}} A_{u,v} - D_{\mathrm{g}} - \frac{\beta}{2} V_{u,v}\right)_+^2}{2V_{u,v}} \right\}.$$

This is the claimed initialization-uniform entrance bound. In particular, the entrance condition (3a) of Theorem 3.1 gives directly

$$c_{\mathrm{g}} A_T^{\mathrm{ent}} - D_{\mathrm{g}} - \frac{\beta}{2} V_T^{\mathrm{ent}} \geq \sqrt{2 V_T^{\mathrm{ent}} \log \frac{2}{\delta}}.$$

Thus, for $(u, v) = (u_T, v_T)$, the right-hand side is at most $\delta/2$.

For completeness, Stage I also admits a fixed-m refinement. Define

$$c_{\mathcal{M},m} := \inf_{\substack{s \in \mathcal{S},\ i \in [m] \\ 0 < |z| \leq (1-\gamma)^{-1}}} \left| \frac{F_{(\mathcal{T}^\pi \boldsymbol{\eta}_m)(s)}(\theta_m(s,i) + z) - \tau_i}{\tau_i(1-\tau_i) z} \right|. \tag{20}$$

Then $c_{\mathcal{M},m} \geq c_{\mathcal{M}}$. The preceding entrance proof only invokes the identifiability inequality at the current value of m, so the same conditional bound remains valid when the Stage-I constants r_{out}, c_{g}, and β are computed with $c_{\mathcal{M},m}$. The main theorem retains the smaller, uniform constant $c_{\mathcal{M}}$ to use one MDP-dependent sensitivity constant across quantile resolutions. □

C.4 Proof of Lemma 4.3: Local linear structure

For clarity, the matrix in the local derivative $\boldsymbol{G}_m = \boldsymbol{D}_m - \gamma \boldsymbol{B}_m$ has entries

$$(\boldsymbol{B}_m)_{(s,i),(s',j)} = \frac{1}{m} \sum_{a \in \mathcal{A}} \pi(a \mid s) P(s' \mid s, a) p_{s,a}\big(\theta_m(s,i) - \gamma \theta_m(s',j)\big).$$

Here $\boldsymbol{D}_m = \mathrm{diag}_{s,i}(d_{s,i})$ and $d_{s,i} = p_{(\mathcal{T}^\pi \boldsymbol{\eta}_m)(s)}(\theta_m(s,i))$.

Proof. The boundary argument in the proof of Lemma D.2 shows that, throughout the r_{out}-ball, every density argument either remains in $(0,1)$ or remains in one of the two exterior components. Under condition (ii) of Assumption 1, the zero extension is Lipschitz across the endpoints as well. Hence $\boldsymbol{h}$ is continuously differentiable on this ball. Differentiating Equation (7) at $\boldsymbol{\theta}_m$ gives

$$\nabla \boldsymbol{h}(\boldsymbol{\theta}_m) = \boldsymbol{D}_m - \gamma \boldsymbol{B}_m = \boldsymbol{G}_m.$$

Summing a row of $\boldsymbol{B}_m$ over (s', j) and using Equation (38) gives $\boldsymbol{B}_m \mathbf{1} = \boldsymbol{D}_m \mathbf{1}$. This proves the first part of Lemma 4.3.

Equation (41) gives $d_{s,i} \geq c_{\mathcal{M}}/(4m)$. Since $d_{s,i} \leq C_0$, the condition $0 \leq \alpha C_0 \leq 1$ implies $\boldsymbol{I} - \alpha \boldsymbol{G}_m \geq \mathbf{0}$. Its row sums satisfy

$$(\boldsymbol{I} - \alpha \boldsymbol{G}_m)\mathbf{1} = \mathbf{1} - \alpha(1-\gamma)\boldsymbol{D}_m \mathbf{1} \leq \left(1 - \frac{\alpha c_{\mathcal{M}}(1-\gamma)}{4m}\right) \mathbf{1}.$$

For a nonnegative matrix, the largest row sum is its induced ℓ_∞ norm. Consequently,

$$\boldsymbol{I} - \alpha \boldsymbol{G}_m \geq \mathbf{0}, \qquad \|\boldsymbol{I} - \alpha \boldsymbol{G}_m\|_\infty \leq 1 - \frac{\alpha c_{\mathcal{M}}(1-\gamma)}{4m} = 1 - \alpha\mu,$$

which proves the second part of Lemma 4.3. For completeness, the same argument also gives the additional structural fact that $\boldsymbol{G}_m$ is a nonsingular M-matrix. Choose any $\alpha \in (0, C_0^{-1}]$. The preceding bound gives $\|\boldsymbol{I} - \alpha \boldsymbol{G}_m\|_\infty < 1$, so the Neumann series converges and

$$\boldsymbol{G}_m^{-1} = \alpha \sum_{k=0}^{\infty} (\boldsymbol{I} - \alpha \boldsymbol{G}_m)^k \geq \mathbf{0}.$$

Moreover, $\boldsymbol{G}_m = \boldsymbol{D}_m - \gamma \boldsymbol{B}_m$ has nonpositive off-diagonal entries. Thus it is a Z-matrix with a nonnegative inverse, equivalently a nonsingular M-matrix.

Finally, fix one density argument x at the fixed point and let $w = e_{s,i} - \gamma e_{s',j}$ be its scalar perturbation. The L-Lipschitz property gives

$$|F_{s,a}(x + w) - F_{s,a}(x) - p_{s,a}(x) w| \leq \frac{L}{2} w^2.$$

Averaging this inequality and using $|w| \leq (1+\gamma) \|\boldsymbol{e}\|_\infty$ proves the final claim of Lemma 4.3. □

C.5 Proof of Lemma 4.4: Variance–drift matching

Proof. Since $\boldsymbol{B}_m\mathbf{1} = \boldsymbol{D}_m\mathbf{1}$,

$$\boldsymbol{v}^\top\mathbf{1} - \boldsymbol{v}^\top(\boldsymbol{I} - \alpha_t\boldsymbol{G}_m)\mathbf{1} = \alpha_t\boldsymbol{v}^\top\boldsymbol{G}_m\mathbf{1} = (1-\gamma)\alpha_t\boldsymbol{v}^\top\boldsymbol{D}_m\mathbf{1},$$

which proves Equation (11a).

For the variance bound, let $Y_{s,i}^{(t+1)}$ be the random empirical-CDF value used to update the current coordinate:

$$Y_{s,i}^{(t+1)} := \frac{1}{m}\sum_{j=1}^{m}\mathbf{1}\left\{r^{(t+1,s)} + \gamma\theta^{(t)}\big(s'^{(t+1,s)}, j\big) < \theta^{(t)}(s,i)\right\}.$$

Its conditional mean is denoted by $q_{s,i}^{(t)} := \mathbb{E}[Y_{s,i}^{(t+1)} \mid \mathcal{F}_t]$. Since $Y_{s,i}^{(t+1)} \in [0,1]$, its conditional variance is at most $q_{s,i}^{(t)}(1-q_{s,i}^{(t)})$. Assumption 1 implies that the density entering $q_{s,i}^{(t)}$ is bounded by C_0, so the definition of r_{out} gives

$$\begin{aligned}\left|q_{s,i}^{(t)} - \tau_i\right| &\le C_0(1+\gamma)\left\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}_m\right\|_\infty \\ &\le \frac{1}{8m} \le \frac{1}{2}\tau_i(1-\tau_i).\end{aligned}$$

Consequently,

$$\begin{aligned}\mathbb{E}[(\xi_{s,i}^{(t+1)})^2 \mid \mathcal{F}_t] &\le q_{s,i}^{(t)}(1-q_{s,i}^{(t)}) \\ &= \tau_i(1-\tau_i) + (q_{s,i}^{(t)} - \tau_i)(1-2\tau_i) - (q_{s,i}^{(t)} - \tau_i)^2 \\ &\le \frac{3}{2}\tau_i(1-\tau_i) \le 2c_{\mathcal{M}}^{-1}d_{s,i},\end{aligned} \tag{21}$$

where the last inequality follows from Equation (41). For every $\boldsymbol{v} \ge \mathbf{0}$, Cauchy–Schwarz inequality gives

$$(\boldsymbol{v}^\top\boldsymbol{\xi}^{(t+1)})^2 \le (\boldsymbol{v}^\top\mathbf{1})\sum_{s,i} v_{s,i}(\xi_{s,i}^{(t+1)})^2.$$

Taking conditional expectations and applying Equation (21) proves Equation (11b). □

C.6 Proof of Lemma 4.5: Local capture

Proof. Condition on $\mathcal{F}_n$. Let $\boldsymbol{e}^{(t)}$ be the local error and let σ be the first exit time from the r_{out}-ball:

$$\boldsymbol{e}^{(t)} := \boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}_m, \qquad \sigma := \inf\left\{t \ge n : \left\|\boldsymbol{e}^{(t)}\right\|_\infty > r_{\text{out}}\right\}.$$

The constants μ and L_h below are, respectively, the local contraction scale and the coefficient of the quadratic remainder:

$$\mu := \frac{c_{\mathcal{M}}(1-\gamma)}{4m}, \qquad L_h := \frac{L(1+\gamma)^2}{2}.$$

For $r \le k$, let $\boldsymbol{\Phi}_{r,k}$ denote the linearized propagator from time r to time k:

$$\boldsymbol{\Phi}_{r,k} := \prod_{j=r}^{k-1}(\boldsymbol{I} - \alpha_j\boldsymbol{G}_m).$$

The bold symbol $\boldsymbol{\Phi}_{r,k}$ is a matrix propagator and is distinct from the scalar smooth maximum Φ_β used in the entrance argument. Under the standing assumptions, $\bar{\alpha}C_0 \le 1$ and $\alpha_t \le \bar{\alpha}$. Moreover, $\mu \le c_{\mathcal{M}}/(4m) \le d_{s,i} \le C_0$, where we used Equation (41). Hence $0 \le \mu\alpha_t \le 1$. The bounds proved in Lemma 4.3 give

$$\boldsymbol{I} - \alpha_t\boldsymbol{G}_m \ge \mathbf{0}, \qquad \|\boldsymbol{I} - \alpha_t\boldsymbol{G}_m\|_\infty \le 1 - \mu\alpha_t.$$

For $t < \sigma$, Equation (9) gives

$$\boldsymbol{e}^{(t+1)} = (\boldsymbol{I} - \alpha_t\boldsymbol{G}_m)\boldsymbol{e}^{(t)} - \alpha_t\boldsymbol{\xi}^{(t+1)} - \alpha_t\boldsymbol{R}(\boldsymbol{e}^{(t)}).$$

Consequently, for every $n < k \le \sigma$,

$$\boldsymbol{e}^{(k)} = \boldsymbol{\Phi}_{n,k}\boldsymbol{e}^{(n)} - \sum_{t=n}^{k-1} \alpha_t \boldsymbol{\Phi}_{t+1,k}\boldsymbol{\xi}^{(t+1)} - \sum_{t=n}^{k-1} \alpha_t \boldsymbol{\Phi}_{t+1,k}\boldsymbol{R}(\boldsymbol{e}^{(t)}).$$

Define the local logarithmic complexity $\ell_T^{\mathrm{loc}}(\delta)$ and the resulting uniform noise envelope $b_{n,T}(\delta)$ by

$$\ell_T^{\mathrm{loc}}(\delta) := \log \frac{4|\mathcal{S}|mT^2}{\delta},$$
$$b_{n,T}(\delta) := 2\sqrt{\frac{\alpha_n \ell_T^{\mathrm{loc}}(\delta)}{c_{\mathcal{M}}(1-\gamma)}} + \frac{\alpha_n \ell_T^{\mathrm{loc}}(\delta)}{3}.$$

Lemma D.3 implies that, with conditional probability at least $1-\delta$,

$$\max_{n \le r < k \le T} \left\| \sum_{t=r}^{k-1} \alpha_t \boldsymbol{\Phi}_{t+1,k} \mathbf{1}\{t < \sigma\} \boldsymbol{\xi}^{(t+1)} \right\|_\infty \le b_{n,T}(\delta). \tag{22}$$

The definition of r_{out} gives $L_h r_{\mathrm{out}}/\mu \le 1/16$, and the other two smallness conditions in Lemma D.4 are exactly the assumed bounds on $b_{n,T}(\delta)$.

We now work on the event in Equation (22). Applying Lemma D.4 gives

$$\left\|\boldsymbol{e}^{(T)}\right\|_\infty \le \left\|\boldsymbol{e}^{(n)}\right\|_\infty \exp\left\{-\frac{15\mu}{16}\sum_{t=n}^{T-1}\alpha_t\right\} + b_{n,T}(\delta) + \frac{32L_h}{15\mu} b_{n,T}^2(\delta).$$

The same bootstrap also proves that the exit time satisfies $\sigma > T$. This is the claimed local-capture statement and completes the proof. □

C.7 Proofs for Section 3

Proof of Theorem 3.1. Apply Lemma 4.2 with $u = u_T$ and $v = v_T$. The entrance condition gives $\Gamma_T \ge \sqrt{2V_T^{\mathrm{ent}} \log(2/\delta)}$, so

$$\mathbb{P}(\tau_{u_T,v_T} = \infty) \le \frac{\delta}{2}.$$

On the complementary event, set $n = \tau_{u_T,v_T}$. This is an $(\mathcal{F}_t)$-stopping time, $u_T \le n \le v_T$, and $\left\|\boldsymbol{\theta}^{(n)} - \boldsymbol{\theta}_m\right\|_\infty \le r_{\mathrm{out}}/2$. Because the step sizes are nonincreasing,

$$\alpha_n \le \alpha_{u_T}, \qquad \sum_{t=n}^{T-1}\alpha_t \ge \sum_{t=v_T}^{T-1}\alpha_t = A_T^{\mathrm{loc}}.$$

Moreover, the local envelope in Lemma 4.5 satisfies $b_{n,T}(\delta/2) \le b_T(\delta)$, because $\ell_T^{\mathrm{loc}}(\delta/2) = \ell_T(\delta)$ and $\alpha_n \le \alpha_{u_T}$. The capture condition (3b) therefore lets us apply that lemma. To make the random starting time explicit, for each $t \in \{u_T, \dots, v_T\}$ let $\mathcal{E}_t$ denote the local-capture conclusion when the lemma is started at time t, and define $\mathcal{E}_{\mathrm{loc}}$ on $\{n < \infty\}$ by $\mathcal{E}_{\mathrm{loc}} = \mathcal{E}_n$. The conditional conclusion of the lemma gives

$$\mathbb{P}(\mathcal{E}_t^c \mid \mathcal{F}_t)\mathbf{1}\{n = t\} \le \frac{\delta}{2}\mathbf{1}\{n = t\}.$$

Taking expectations and summing over the disjoint events $\{n = t\}$ shows

$$\mathbb{P}(\mathcal{E}_{\mathrm{loc}}^c,\ n < \infty) \le \sum_{t=u_T}^{v_T} \mathbb{E}\left[\mathbf{1}\{n = t\}\mathbb{P}(\mathcal{E}_t^c \mid \mathcal{F}_t)\right] \le \frac{\delta}{2}.$$

On $\mathcal{E}_{\mathrm{loc}}$, the lemma yields

$$\left\|\boldsymbol{\theta}^{(T)} - \boldsymbol{\theta}_m\right\|_\infty \le \frac{r_{\mathrm{out}}}{2}\exp\left\{-\frac{15\mu}{16}A_T^{\mathrm{loc}}\right\} + b_T(\delta) + \frac{32L_h}{15\mu}b_T^2(\delta).$$

The assumed inequality $b_T(\delta) \le \mu/(64L_h)$ bounds the quadratic remainder when $L_h > 0$ because

$$\frac{32L_h}{15\mu} b_T^2(\delta) \le \frac{32}{15 \cdot 64} b_T(\delta) = \frac{1}{30} b_T(\delta).$$

The remainder vanishes when $L_h = 0$. Thus, on the same event,

$$\left\| \boldsymbol{\theta}^{(T)} - \boldsymbol{\theta}_m \right\|_\infty \le \frac{r_{\text{out}}}{2} \exp\left\{ -\frac{15\mu}{16} A_T^{\text{loc}} \right\} + \frac{31}{30} b_T(\delta). \tag{23}$$

A union bound gives this inequality with probability at least $1 - \delta$. This is exactly Equation (4). □

Proof of Corollary 3.2. For the constant schedule, $\alpha_{u_T} = \eta$ and $A_T^{\text{loc}} \ge \eta T/3$ for $T \ge 8$. Substitution into Equation (23) gives the sharper appendix bound

$$\left\| \boldsymbol{\theta}^{(T)} - \boldsymbol{\theta}_m \right\|_\infty \le \frac{r_{\text{out}}}{2} e^{-5\mu\eta T/16} + \frac{31}{15} \sqrt{\frac{\eta \ell_T(\delta)}{c_{\mathcal{M}}(1-\gamma)}} + \frac{31}{90} \eta \ell_T(\delta),$$

which implies the first claim.

Now let $\alpha_t = c(t+1)^{-a}$ with $a \in (0,1)$. For $T \ge 8$, monotonicity and the block lengths give the explicit bounds

$$\alpha_{u_T} \le 4cT^{-a}, \qquad A_T^{\text{loc}} \ge \frac{c}{2} T^{1-a}.$$

Substituting these inequalities into Equation (23) yields

$$\left\| \boldsymbol{\theta}^{(T)} - \boldsymbol{\theta}_m \right\|_\infty \le \frac{r_{\text{out}}}{2} \exp\left\{ -\frac{15\mu c}{32} T^{1-a} \right\} + \frac{62}{15} \sqrt{\frac{c\ell_T(\delta)}{c_{\mathcal{M}}(1-\gamma)T^a}} + \frac{62c\ell_T(\delta)}{45T^a}.$$

Because $\mu = c_{\mathcal{M}}(1-\gamma)/(4m)$, this explicit inequality proves Equation (5).

For completeness, the burn-in conditions are eventually satisfied for every fixed δ. Indeed, on the entrance block,

$$\frac{c}{8} T^{1-a} \le A_T^{\text{ent}} \le 2cT^{1-a}, \qquad V_T^{\text{ent}} \le 8c^2 T^{1-2a}.$$

Consequently, $V_T^{\text{ent}}/A_T^{\text{ent}} \le 64cT^{-a}$, $(A_T^{\text{ent}})^2/V_T^{\text{ent}} \ge T/512$, and $b_T(\delta) \to 0$. Indeed, because $A_T^{\text{ent}} \to \infty$ and $V_T^{\text{ent}}/A_T^{\text{ent}} \to 0$, for all sufficiently large T we have

$$D_{\text{g}} \le \frac{c_{\text{g}}}{4} A_T^{\text{ent}}, \qquad \frac{\beta}{2} V_T^{\text{ent}} \le \frac{c_{\text{g}}}{4} A_T^{\text{ent}}.$$

Thus $\Gamma_T \ge c_{\text{g}} A_T^{\text{ent}}/2$, and

$$\frac{\Gamma_T^2}{2V_T^{\text{ent}}} \ge \frac{c_{\text{g}}^2 (A_T^{\text{ent}})^2}{8V_T^{\text{ent}}} \longrightarrow \infty.$$

This verifies the entrance condition for every fixed δ, while $b_T(\delta) \to 0$ verifies the local condition. Hence the argument covers the full range $a \in (0,1)$. □

Dependence of the sufficient burn-in on quantile resolution. Under Assumption 1(ii), keep $c_{\mathcal{M}}, C_0, L, \gamma$ and the polynomial-schedule parameters a, c fixed. The explicit constants satisfy

$$r_{\text{out}} \asymp m^{-1}, \qquad \mu \asymp m^{-1}, \qquad c_{\text{g}} \asymp m^{-2}, \qquad \beta = \widetilde{\Theta}(m), \qquad D_{\text{g}} = O(1).$$

For $\alpha_t = c(t+1)^{-a}$, the entrance-block sums have orders $A_T^{\text{ent}} \asymp T^{1-a}$ and $V_T^{\text{ent}} \asymp T^{1-2a}$. Separately dominating D_{g}, the potential-curvature cost $\beta V_T^{\text{ent}}/2$, and the martingale deviation in Equation (3a) gives, up to logarithmic factors,

$$T^{1-a} \gtrsim m^2, \qquad T^a \gtrsim m^3, \qquad T \gtrsim m^4.$$

The capture condition itself only requires $T^a \gtrsim m^2$, up to logarithmic factors. Consequently, a sufficient burn-in length can be chosen of order

$$\widetilde{O}\left(\max\{ m^{2/(1-a)},\ m^{3/a},\ m^4 \} \right).$$

Here logarithmic factors may depend on the failure probability δ and the number of coordinates $|\mathcal{S}|m$, and the implicit constants may depend on the fixed schedule and problem parameters. The same polynomial scaling holds under Assumption 1(i) if $\Delta_m \gtrsim m^{-1}$. If this boundary margin is smaller, it also enters the local radius and the global-drift constants, so the displayed scaling need not apply.

Proof of Corollary 3.3. Write $\alpha_t = c/(t+t_0)$. The offset condition gives $\bar{\alpha}C_0 = (c/t_0)C_0 \le 1$, and the gain condition gives

$$q \coloneqq \exp\{-15\mu c \log(4/3)/16\} \le \frac{1}{4}.$$

We sharpen the entrance step by retaining the dependence of the global drift on the current error magnitude. Write

$$\boldsymbol{e}^{(t)} \coloneqq \boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}_m.$$

Set

$$\lambda_{\mathrm{h}} \coloneqq \frac{\mu}{6}, \qquad p \coloneqq \lambda_{\mathrm{h}} c.$$

By the choice of C_{harm} in Appendix C.1, $p \ge 4/3 > 1$.

Also set

$$B_\Phi \coloneqq \frac{1+\bar{\alpha}}{1-\gamma} + \frac{\log(2|\mathcal{S}|m)}{\beta}.$$

Lemma D.1 and the smooth-max equivalence imply $\Phi_\beta(\boldsymbol{e}^{(t)}) \le B_\Phi$ for every t.

Choose an integer $u \ge \max\{1, t_0\}$ and define

$$v = \left\lceil (u+t_0)^{4/3} - t_0 \right\rceil, \qquad N = 2v.$$

We will increase u below until a finite list of deterministic inequalities is satisfied. Let

$$\tau = \inf\left\{ u \le t \le v : \left\| \boldsymbol{e}^{(t)} \right\|_\infty \le \frac{r_{\mathrm{out}}}{2} \right\},$$

with the convention $\inf \varnothing = \infty$, and define

$$Z_{t+1} = -\left\langle \nabla \Phi_\beta(\boldsymbol{e}^{(t)}), \boldsymbol{\xi}^{(t+1)} \right\rangle.$$

Then $\mathbb{E}[Z_{t+1} \mid \mathcal{F}_t] = 0$ and $|Z_{t+1}| \le 1$. On $\{t < \tau\}$, Equation (14) yields

$$\langle \nabla \Phi_\beta(\boldsymbol{e}), \boldsymbol{h}(\boldsymbol{\theta}) \rangle \ge \frac{\mu}{4} \|\boldsymbol{e}\|_\infty,$$

and the definition of β gives

$$\frac{\log(2|\mathcal{S}|m)}{\beta} \le \frac{(1-\gamma) r_{\mathrm{out}}}{4} \le \frac{V}{2}.$$

Thus $\Phi_\beta(\boldsymbol{e}) \le 3V/2$ and

$$\langle \nabla \Phi_\beta(\boldsymbol{e}), \boldsymbol{h}(\boldsymbol{\theta}) \rangle \ge \frac{\mu}{6} \Phi_\beta(\boldsymbol{e}).$$

This inequality and the same Taylor expansion as in the proof of Lemma 4.2 give

$$\Phi_\beta(\boldsymbol{e}^{(t+1)}) \le (1 - \lambda_{\mathrm{h}} \alpha_t) \Phi_\beta(\boldsymbol{e}^{(t)}) + \alpha_t Z_{t+1} + \frac{\beta}{2} \alpha_t^2. \tag{24}$$

For $u \le t < v$, define the deterministic backward weights

$$\omega_{t,v} \coloneqq \prod_{j=t+1}^{v-1} (1 - \lambda_{\mathrm{h}} \alpha_j), \qquad \omega_{u-1,v} \coloneqq \prod_{j=u}^{v-1} (1 - \lambda_{\mathrm{h}} \alpha_j).$$

Since $0 \le \lambda_{\mathrm{h}} \alpha_j \le 1$, $0 \le \omega_{t,v} \le 1$. Integral comparison gives

$$\begin{aligned} \omega_{u-1,v} &\le \exp\left\{ -\lambda_{\mathrm{h}} \sum_{t=u}^{v-1} \alpha_t \right\} \\ &\le \left(\frac{u+t_0}{v+t_0} \right)^p \le (u+t_0)^{-p/3}. \end{aligned} \tag{25}$$

We also need a weighted squared-step estimate. For $k \in \{u, \dots, v\}$, set

$$H_k \coloneqq \sum_{t=u}^{k-1} \alpha_t^2 \prod_{j=t+1}^{k-1} (1 - \lambda_{\mathrm{h}} \alpha_j).$$

Then $H_u = 0$ and, writing $x = k + t_0$,

$$H_{k+1} = \left(1 - \frac{p}{x}\right) H_k + \frac{c^2}{x^2}.$$

An induction therefore yields

$$H_k \le \frac{c^2}{(p-1)(k+t_0)}, \qquad u \le k \le v. \tag{26}$$

Indeed, if the bound holds at k, then

$$H_{k+1} \le \frac{c^2}{p-1} \frac{x-1}{x^2} \le \frac{c^2}{(p-1)(x+1)}.$$

Because $0 \le \omega_{t,v} \le 1$, Equation (26) also gives

$$\sum_{t=u}^{v-1} \alpha_t^2 \omega_{t,v}^2 \le \frac{c^2}{(p-1)(v+t_0)}.$$

Now consider the stopped martingale

$$M_v \coloneqq \sum_{t=u}^{v-1} \alpha_t \omega_{t,v} \mathbf{1}\{t < \tau\} Z_{t+1}.$$

Azuma–Hoeffding and the preceding squared-weight bound imply

$$\mathbb{P}\left(M_v > c \sqrt{\frac{2 \log(4/\delta)}{(p-1)(v+t_0)}} \right) \le \frac{\delta}{4}. \tag{27}$$

On $\{\tau = \infty\}$, unrolling Equation (24) gives

$$\begin{aligned} \Phi_\beta(\boldsymbol{e}^{(v)}) \le & B_\Phi \omega_{u-1,v} + M_v + \frac{\beta}{2} \sum_{t=u}^{v-1} \alpha_t^2 \omega_{t,v} \\ \le & B_\Phi (u + t_0)^{-p/3} + M_v + \frac{\beta c^2}{2(p-1)(v+t_0)}. \end{aligned}$$

Increase u, if necessary, until

$$\begin{aligned} B_\Phi (u + t_0)^{-p/3} &\le \frac{r_{\mathrm{out}}}{8}, \\ \frac{\beta c^2}{2(p-1)(v+t_0)} &\le \frac{r_{\mathrm{out}}}{8}, \\ c \sqrt{\frac{2 \log(4/\delta)}{(p-1)(v+t_0)}} &\le \frac{r_{\mathrm{out}}}{8}. \end{aligned} \tag{28}$$

All three conditions hold for all sufficiently large u, because $v + t_0 \ge (u + t_0)^{4/3}$.

On the event complementary to that in Equation (27), the event $\{\tau = \infty\}$ would imply

$$\Phi_\beta(\boldsymbol{e}^{(v)}) \le \frac{3 r_{\mathrm{out}}}{8} < \frac{r_{\mathrm{out}}}{2},$$

contradicting

$$\Phi_\beta(\boldsymbol{e}^{(v)}) \ge \left\| \boldsymbol{e}^{(v)} \right\|_\infty > \frac{r_{\mathrm{out}}}{2}.$$

Hence $\mathbb{P}(\tau = \infty) \le \delta/4$.

On the complementary event, $n = \tau \in [u, v]$ is a stopping time satisfying

$$\left\|\boldsymbol{\theta}^{(n)} - \boldsymbol{\theta}_m\right\|_\infty \le \frac{r_{\text{out}}}{2}.$$

The interval from entrance to N leaves enough time for a fixed local reduction: since $v \ge u \ge t_0$,

$$\sum_{t=v}^{N-1} \alpha_t \ge c \log \frac{2v + t_0}{v + t_0} \ge c \log(4/3).$$

Moreover, $\alpha_n \le c/(u + t_0)$, so the initial local envelope satisfies

$$b_{n,N}(\delta/4) \le \bar{b}_u^{\text{init}} := 2\sqrt{\frac{c}{c_{\mathcal{M}}(1-\gamma)(u+t_0)} \log \frac{16|\mathcal{S}|mN^2}{\delta}} + \frac{c}{3(u+t_0)} \log \frac{16|\mathcal{S}|mN^2}{\delta}.$$

Since $N = O((u + t_0)^{4/3})$, we have $\bar{b}_u^{\text{init}} \to 0$. Increase u, if necessary, until

$$\bar{b}_u^{\text{init}} \le \min\left\{\frac{r_{\text{out}}}{16}, \frac{\mu}{64 L_h}\right\},$$

where the second entry is $+\infty$ when $L_h = 0$. Lemma 4.5, in the exact form supplied by Lemma D.4, now gives

$$\begin{aligned}\left\|\boldsymbol{\theta}^{(N)} - \boldsymbol{\theta}_m\right\|_\infty &\le q\frac{r_{\text{out}}}{2} + b_{n,N}(\delta/4) + \frac{32 L_h}{15\mu} b_{n,N}^2(\delta/4) \\ &\le \frac{r_{\text{out}}}{8} + \frac{31 r_{\text{out}}}{480} < \frac{r_{\text{out}}}{2}.\end{aligned}$$

The quadratic term is at most $b_{n,N}(\delta/4)/30$ when $L_h > 0$ and vanishes when $L_h = 0$. The same lemma keeps the trajectory inside the r_{out}-ball through time N. Its failure probability, conditionally on $\mathcal{F}_n$, is at most $\delta/4$. Summing these conditional bounds over the disjoint events $\{n = t\}$ gives an unconditional failure probability at most $\delta/4$, just as in the proof of Theorem 3.1. Thus entrance and initial capture together cost at most $\delta/2$.

It remains to propagate the rate beyond N. Let $N_j = 2^j N$ and assign failure probabilities $\delta_j = \delta/2^{j+3}$, so that $\sum_{j \ge 0} \delta_j = \delta/4$. On the full block $[N_j, N_{j+1}]$, define the exact local envelope

$$b_j = 2\sqrt{\frac{\alpha_{N_j}}{c_{\mathcal{M}}(1-\gamma)} \log \frac{4|\mathcal{S}|mN_{j+1}^2}{\delta_j}} + \frac{\alpha_{N_j}}{3} \log \frac{4|\mathcal{S}|mN_{j+1}^2}{\delta_j}.$$

Since $N_j \ge N \ge t_0$, integral comparison and monotonicity give

$$\sum_{t=N_j}^{N_{j+1}-1} \alpha_t \ge c \log(4/3), \qquad \alpha_{N_j} \le \frac{c}{N_j}.$$

Also, $2^j = N_j/N \le N_j$, and hence

$$\log \frac{4|\mathcal{S}|mN_{j+1}^2}{\delta_j} = \log \frac{2^{j+7}|\mathcal{S}|mN_j^2}{\delta} \le L_j, \qquad L_j := \log \frac{128|\mathcal{S}|mN_j^3}{\delta}.$$

Introduce the square-root and linear parts

$$S_j := \sqrt{\frac{cL_j}{c_{\mathcal{M}}(1-\gamma)N_j}}, \qquad U_j := \frac{cL_j}{N_j}, \qquad R_j := S_j + U_j.$$

Then

$$b_j \le 2S_j + \frac{1}{3}U_j \le 2R_j. \tag{29}$$

Because L_j/N_j decreases with j (here $L_0 \geq \log 128 > 3$), both S_j and U_j decrease. Increase u, and consequently N, once more if necessary, until

$$2S_0 + \frac{1}{3}U_0 \leq \min\left\{\frac{r_{\text{out}}}{16}, \frac{\mu}{64L_h}\right\}.$$

All b_j then satisfy the local-capture conditions. Every requirement above holds for all sufficiently large u: all three entrance terms in Equation (28) tend to zero, and both local noise envelopes vanish. We may therefore take the smallest integer $u \geq \max\{1, t_0\}$ satisfying this finite list of deterministic inequalities. This makes the initial time N constructive and deterministic.

Set $E_j = \left\|\boldsymbol{\theta}^{(N_j)} - \boldsymbol{\theta}_m\right\|_\infty$. If $E_j \leq r_{\text{out}}/2$, the exact local-capture estimate on block j gives, outside an event of conditional probability at most δ_j,

$$E_{j+1} \leq qE_j + b_j + \frac{32L_h}{15\mu}b_j^2 \leq qE_j + \frac{31}{30}b_j.$$

Since $q \leq 1/4$ and $b_j \leq r_{\text{out}}/16$, this bound also gives $E_{j+1} < r_{\text{out}}/2$. Thus every successive application of the local lemma is legitimate on the preceding success events. Conditioning on $\mathcal{F}_{N_j}$ and summing over the blocks bounds their total failure probability by $\sum_{j\geq 0}\delta_j$; no independence is needed. Iteration yields

$$E_j \leq q^j E_0 + \frac{31}{30}\sum_{\ell=0}^{j-1} q^{j-1-\ell}b_\ell. \tag{30}$$

For $0 \leq \ell < j$, monotonicity of L_j gives

$$S_\ell \leq 2^{(j-\ell)/2}S_j, \qquad U_\ell \leq 2^{j-\ell}U_j.$$

The linear term is kept separate because it decays as N_j^{-1}, whereas the square-root term decays as $N_j^{-1/2}$, up to logarithms. With $r = j - 1 - \ell$ and $q \leq 1/4$, we obtain

$$\begin{aligned}\sum_{\ell=0}^{j-1} q^{j-1-\ell}b_\ell &\leq 2\sqrt{2}S_j\sum_{r=0}^{j-1}(q\sqrt{2})^r + \frac{2}{3}U_j\sum_{r=0}^{j-1}(2q)^r \\ &\leq \frac{2\sqrt{2}}{1-\sqrt{2}/4}S_j + \frac{4}{3}U_j.\end{aligned}$$

Define the numerical constant

$$K_* := \frac{31\sqrt{2}}{15(1-\sqrt{2}/4)}.$$

Since $K_* > 62/45$, Equation (30) implies $E_j \leq q^j E_0 + K_* R_j$. The inherited error can eventually be absorbed into R_j: indeed, $R_j \geq 2^{-j/2}S_0$ and hence $q^j r_{\text{out}}/(2R_j) \to 0$. Let j_0 be the smallest nonnegative integer for which $q^{j_0}r_{\text{out}}/2 \leq R_{j_0}$. This inequality remains true thereafter because $R_{j+1} \geq R_j/2$ and $q \leq 1/4$. Consequently, for every $j \geq j_0$,

$$E_j \leq (1 + K_*)R_j. \tag{31}$$

Now fix an integer $T \geq N_{j_0}$ and choose J such that $N_J \leq T < N_{J+1}$. If $T > N_J$, apply the local lemma on the partial block $[N_J, T]$ with failure probability $\delta/4$. Since $T < 2N_J$,

$$\log\frac{16|\mathcal{S}|mT^2}{\delta} \leq \log\frac{64|\mathcal{S}|mN_J^2}{\delta} \leq L_J,$$

so its envelope satisfies $b_{\text{part}} \leq 2S_J + U_J/3 \leq 2R_J$ and the local-capture smallness conditions. Using the exact bootstrap estimate without requiring contraction on this partial block gives

$$\begin{aligned}\left\|\boldsymbol{\theta}^{(T)} - \boldsymbol{\theta}_m\right\|_\infty &\leq E_J + b_{\text{part}} + \frac{32L_h}{15\mu}b_{\text{part}}^2 \\ &\leq \left(1 + K_* + \frac{31}{15}\right)R_J.\end{aligned}$$

The same bound follows directly from Equation (31) when $T = N_J$.

Set

$$L_T \coloneqq \log \frac{128|\mathcal{S}|mT^3}{\delta}, \qquad K_{\mathrm{harm}} \coloneqq K_* + \frac{46}{15}.$$

Since $N_J \le T < 2N_J$ and $L_J \le L_T$, the preceding bound implies

$$\left\|\boldsymbol{\theta}^{(T)} - \boldsymbol{\theta}_m\right\|_\infty \le K_{\mathrm{harm}} \left\{ \sqrt{\frac{2cL_T}{c_{\mathcal{M}}(1-\gamma)T}} + \frac{2cL_T}{T} \right\}. \tag{32}$$

For $C_{2,*} = 6K_{\mathrm{harm}}$, we have $L_T \le 3\log(C_{2,*}|\mathcal{S}|mT/\delta)$, and Equation (32) implies Equation (6) with $C = C_{2,*}$.

The entrance and initial-capture events cost at most $\delta/2$, all full dyadic blocks cost at most $\delta/4$, and the final partial block costs at most $\delta/4$. Thus the bound holds with probability at least $1-\delta$ for each fixed $T \ge T_{\mathrm{harm}}(\delta)$, where $T_{\mathrm{harm}}(\delta) = N_{j_0}$ is finite and deterministic. Only the local contraction condition $c \ge C_{\mathrm{harm}}/\mu$ is required of the gain coefficient. □

Dependence of the sufficient burn-in on quantile resolution. Under Assumption 1(ii), keep $c_{\mathcal{M}}, C_0, L, \gamma$ fixed and take the natural harmonic tuning $c \asymp \mu^{-1} \asymp m$ together with $t_0 \asymp c$. Then

$$r_{\mathrm{out}} \asymp m^{-1}, \qquad \mu \asymp m^{-1}, \qquad \beta = \widetilde{\Theta}(m), \qquad B_\Phi = O(1), \qquad p = \lambda_{\mathrm{h}} c = \Theta(1),$$

with $p \ge 4/3$ under the gain condition in Corollary 3.3.

Recall that the scale-sensitive entrance construction takes

$$v = \left\lceil (u+t_0)^{4/3} - t_0 \right\rceil, \qquad N = 2v.$$

The three conditions in Equation (28) can be satisfied, up to logarithmic factors, by taking

$$u \gtrsim m^3.$$

Indeed, the initial-error term is then already smaller than r_{out}, while the potential-curvature and martingale terms require, respectively,

$$v \gtrsim m^4 \qquad \text{and} \qquad v \gtrsim m^4$$

up to logarithmic factors. The initial local-capture condition $\bar{b}_u^{\mathrm{init}} \le \min\{r_{\mathrm{out}}/16, \mu/(64L_h)\}$ also requires only $u \gtrsim m^3$ up to logarithmic factors. Consequently, entrance and initial capture can be completed by a deterministic time

$$N = \widetilde{O}(m^4). \tag{33}$$

It remains to absorb the error inherited at time N into the stochastic scale of the subsequent dyadic blocks. Recall that $N_j = 2^j N$ and

$$R_j = S_j + U_j, \qquad S_j = \sqrt{\frac{cL_j}{c_{\mathcal{M}}(1-\gamma)N_j}}, \qquad U_j = \frac{cL_j}{N_j}.$$

Under the preceding choice of N,

$$S_0 = \widetilde{\Omega}(m^{-3/2}), \qquad R_j \ge 2^{-j/2} S_0 = \widetilde{\Omega}\left(m^{-3/2} 2^{-j/2}\right).$$

Since $E_0 \le r_{\mathrm{out}}/2 = O(m^{-1})$, the defining condition $q^{j_0} r_{\mathrm{out}}/2 \le R_{j_0}$ is satisfied once

$$(q\sqrt{2})^{j_0} \lesssim m^{-1/2}$$

up to logarithmic factors. Because $q \le 1/4$, one may therefore take

$$j_0 \le \frac{1}{3}\log_2 m + O(\log\log m).$$

Combining this with Equation (33) gives the sufficient harmonic burn-in bound

$$T_{\mathrm{harm}}(\delta) = N_{j_0} = \widetilde{O}(m^{13/3}), \tag{34}$$

which is also polynomial in m. The same polynomial scaling holds under Assumption 1(i) if $\Delta_m \gtrsim m^{-1}$. If the boundary margin is smaller, it also enters r_{out} and the preceding scaling need not apply.

D TECHNICAL LEMMAS

D.1 GLOBAL BOUNDEDNESS

Lemma D.1. *Let B be the largest possible discounted return and let Δ be the largest extra excursion caused by one update:*

$$B \coloneqq (1-\gamma)^{-1}, \qquad \Delta \coloneqq \bar{\alpha}(1-\gamma)^{-1}.$$

If $\boldsymbol{\theta}^{(0)} \in [0, B]^{\mathcal{S}\times[m]}$, then, almost surely,

$$\boldsymbol{\theta}^{(t)} \in [-\Delta, B+\Delta]^{\mathcal{S}\times[m]} \qquad \text{for every } t \geq 0. \tag{35}$$

Consequently,

$$\left\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}_m\right\|_\infty \leq \frac{1+\bar{\alpha}}{1-\gamma} \qquad \text{for every } t \geq 0. \tag{36}$$

Proof of Lemma D.1. We prove Equation (35) by induction. Suppose all coordinates of $\boldsymbol{\theta}^{(t)}$ belong to $[-\Delta, B+\Delta]$. If $\theta^{(t)}(s,i) < -\gamma\Delta$, then every sampled quantity $r^{(t+1,s)} + \gamma\theta^{(t)}(s'^{(t+1,s)}, j)$ is at least $-\gamma\Delta$, so the update of this coordinate is nonnegative. Otherwise, since $\alpha_t \leq \bar{\alpha}$,

$$\theta^{(t+1)}(s,i) \geq -\gamma\Delta - \bar{\alpha} = -\Delta.$$

Similarly, every such sampled quantity is at most $1+\gamma(B+\Delta) = B+\gamma\Delta$. If $\theta^{(t)}(s,i) > B+\gamma\Delta$, its update is nonpositive; otherwise,

$$\theta^{(t+1)}(s,i) \leq B + \gamma\Delta + \bar{\alpha} = B + \Delta.$$

This proves the invariant box. Since $\boldsymbol{\theta}_m \in [0, B]^{\mathcal{S}\times[m]}$, Equation (36) follows. □

D.2 LOCAL DENSITY STABILITY

The following lemma supplies the density lower bound used in the local linearization.

Lemma D.2. *For every $\boldsymbol{\theta}$ satisfying $\|\boldsymbol{\theta} - \boldsymbol{\theta}_m\|_\infty \leq r_{\text{out}}$ and every z satisfying $|z - \theta_m(s,i)| \leq r_{\text{out}}$, we have*

$$p_{(\mathcal{T}^\pi \boldsymbol{\eta}_{\boldsymbol{\theta}})(s)}(z) \geq \frac{d_{s,i}}{2} \geq \frac{c_{\mathcal{M}}}{2}\tau_i(1-\tau_i). \tag{37}$$

Proof of Lemma D.2. For every $z \in \mathbb{R}$ and $\boldsymbol{\theta} \in \mathbb{R}^{\mathcal{S}\times[m]}$,

$$p_{(\mathcal{T}^\pi \boldsymbol{\eta}_{\boldsymbol{\theta}})(s)}(z) = \frac{1}{m}\sum_{a\in\mathcal{A},\, s'\in\mathcal{S}}\sum_{j=1}^{m} \pi(a \mid s)P(s' \mid s,a)p_{s,a}\big(z - \gamma\theta(s',j)\big). \tag{38}$$

If $\|\boldsymbol{\theta} - \boldsymbol{\theta}_m\|_\infty \leq r_{\text{out}}$ and $|z - \theta_m(s,i)| \leq r_{\text{out}}$, then

$$|z - \gamma\theta(s',j) - (\theta_m(s,i) - \gamma\theta_m(s',j))| \leq (1+\gamma)r_{\text{out}}. \tag{39}$$

Under condition (i) of Assumption 1, the definition of r_{out} ensures that the two arguments in Equation (39) belong to the same connected component of $\mathbb{R}\setminus\{0,1\}$. The density is L-Lipschitz on the interior component and is zero on the two exterior components. Under condition (ii), $p_{s,a}$ is L-Lipschitz on all of $\mathbb{R}$. Thus, in both cases,

$$\begin{aligned} &\left|p_{(\mathcal{T}^\pi \boldsymbol{\eta}_{\boldsymbol{\theta}})(s)}(z) - p_{(\mathcal{T}^\pi \boldsymbol{\eta}_m)(s)}\big(\theta_m(s,i)\big)\right| \\ \leq &L(1+\gamma)r_{\text{out}} \leq \frac{c_{\mathcal{M}}}{32m} \leq \frac{d_{s,i}}{8}. \end{aligned} \tag{40}$$

Here the last inequality follows by taking $u \to 0$ in the definition of $c_{\mathcal{M}}$:

$$d_{s,i} \geq c_{\mathcal{M}}\tau_i(1-\tau_i) \geq \frac{c_{\mathcal{M}}}{4m}. \tag{41}$$

Equations (40) and (41) prove Equation (37). □

D.3 Local Concentration Tools

Lemma D.3. *Suppose $\bar{\alpha}C_0 \le 1$. Let $n \le T$ be an $(\mathcal{F}_t)$-stopping time, and let $\sigma \ge n$ be a stopping time such that $\left\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}_m\right\|_\infty \le r_{\mathrm{out}}$ on $\{t < \sigma\}$. For $r \le k$, set*

$$\boldsymbol{\Phi}_{r,k} \coloneqq \prod_{j=r}^{k-1} (\boldsymbol{I} - \alpha_j \boldsymbol{G}_m).$$

For every $\delta \in (0,1)$, conditionally on $\mathcal{F}_n$, with probability at least $1-\delta$,

$$\max_{(s,i)\in\mathcal{S}\times[m]} \max_{n\le r<k\le T} \left|\sum_{t=r}^{k-1} \alpha_t \boldsymbol{e}_{s,i}^\top \boldsymbol{\Phi}_{t+1,k} \mathbf{1}\{t<\sigma\} \boldsymbol{\xi}^{(t+1)}\right| \le 2\sqrt{\frac{\alpha_n}{c_{\mathcal{M}}(1-\gamma)} \log\frac{2|\mathcal{S}|mT^2}{\delta}} + \frac{\alpha_n}{3}\log\frac{2|\mathcal{S}|mT^2}{\delta}.$$

Proof of Lemma D.3. Fix a coordinate index $q \in \mathcal{S}\times[m]$ and times $r < k$. Let $\boldsymbol{e}_q$ be the corresponding standard basis vector, and put

$$\boldsymbol{v}_t^\top \coloneqq \boldsymbol{e}_q^\top \boldsymbol{\Phi}_{t+1,k}, \qquad u_t \coloneqq \boldsymbol{v}_t^\top \mathbf{1}, \qquad r \le t < k.$$

Here u_t is the total mass of the nonnegative row vector $\boldsymbol{v}_t$. For every j, Lemma 4.3 gives $\boldsymbol{I} - \alpha_j \boldsymbol{G}_m \ge \mathbf{0}$ and $\|\boldsymbol{I} - \alpha_j \boldsymbol{G}_m\|_\infty \le 1 - \mu\alpha_j \le 1$. It follows that $\boldsymbol{v}_t \ge \mathbf{0}$ and $0 \le u_t \le 1$. Define its mass after one additional linearized step by

$$u_t' \coloneqq \boldsymbol{v}_t^\top (\boldsymbol{I} - \alpha_t \boldsymbol{G}_m)\mathbf{1}.$$

Equation (11a) then yields

$$u_t - u_t' = (1-\gamma)\alpha_t \boldsymbol{v}_t^\top \boldsymbol{D}_m \mathbf{1} \ge 0.$$

All propagator factors are polynomials in the same matrix $\boldsymbol{G}_m$ and hence commute. Consequently,

$$u_t' = \boldsymbol{e}_q^\top \boldsymbol{\Phi}_{t,k}\mathbf{1} = u_{t-1} \qquad (r < t < k),$$

with the endpoint identity $u_r' = \boldsymbol{e}_q^\top \boldsymbol{\Phi}_{r,k}\mathbf{1}$. Since $0 \le u_t' \le u_t \le 1$, we obtain

$$\sum_{t=r}^{k-1} u_t(u_t - u_t') \le \sum_{t=r}^{k-1} (u_t - u_t') = u_{k-1} - u_r' \le 1.$$

Moreover, $\mathbf{1}\{t<\sigma\}$ is $\mathcal{F}_t$-measurable, so stopping preserves the martingale-difference property. By Equation (11b) and monotonicity of the step sizes,

$$\sum_{t=r}^{k-1} \alpha_t^2 \mathbb{E}\left[\left(\boldsymbol{v}_t^\top \mathbf{1}\{t<\sigma\}\boldsymbol{\xi}^{(t+1)}\right)^2 \mid \mathcal{F}_t\right] \le \frac{2\alpha_r}{c_{\mathcal{M}}(1-\gamma)} \sum_{t=r}^{k-1} u_t(u_t - u_t') \le \frac{2\alpha_r}{c_{\mathcal{M}}(1-\gamma)}.$$

Moreover,

$$\left|\alpha_t \boldsymbol{v}_t^\top \mathbf{1}\{t<\sigma\}\boldsymbol{\xi}^{(t+1)}\right| \le \alpha_t \boldsymbol{v}_t^\top \mathbf{1} \le \alpha_r.$$

Freedman's inequality and a union bound over at most $|\mathcal{S}|mT^2$ tuples (s,i,r,k) complete the proof. □

Lemma D.4. *Use the local contraction scale μ, quadratic-remainder coefficient L_h, and propagator $\boldsymbol{\Phi}_{r,k}$ introduced in Section C.6. Let n be an $(\mathcal{F}_t)$-stopping time satisfying*

$$\left\|\boldsymbol{\theta}^{(n)} - \boldsymbol{\theta}_m\right\|_\infty \le \frac{r_{\mathrm{out}}}{2},$$

and define

$$\sigma \coloneqq \inf\left\{t \ge n : \left\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}_m\right\|_\infty > r_{\mathrm{out}}\right\},$$

with the convention $\inf \varnothing = \infty$.

Suppose that on an event $\mathcal{E}$*, the propagated linear noise is uniformly bounded by a deterministic envelope* b*:*

$$\max_{n \le r < k \le T} \left\| \sum_{t=r}^{k-1} \alpha_t \mathbf{\Phi}_{t+1,k} \mathbf{1}\{t < \sigma\} \boldsymbol{\xi}^{(t+1)} \right\|_\infty \le b,$$

and

$$b \le \frac{r_{\text{out}}}{16}, \ \frac{L_h r_{\text{out}}}{\mu} \le \frac{1}{16}, \ b \le \frac{\mu}{64 L_h}, \tag{42}$$

where the last bound is interpreted as vacuous when $L_h = 0$*. Then on* $\mathcal{E}$*, we have* $\sigma > T$ *and*

$$\left\| \boldsymbol{\theta}^{(T)} - \boldsymbol{\theta}_m \right\|_\infty \le \left\| \boldsymbol{\theta}^{(n)} - \boldsymbol{\theta}_m \right\|_\infty \exp\left\{ -\frac{15\mu}{16} \sum_{t=n}^{T-1} \alpha_t \right\} + b + \frac{32 L_h}{15\mu} b^2.$$

Proof of Lemma D.4. First define the auxiliary sequence $\boldsymbol{w}^{(k)}$, which contains only the propagated stochastic perturbations, with the same sign as in the error recursion. For $n \le k \le T$, let

$$\boldsymbol{w}^{(n)} := \mathbf{0}, \qquad \boldsymbol{w}^{(k)} := -\sum_{t=n}^{k-1} \alpha_t \mathbf{\Phi}_{t+1,k} \mathbf{1}\{t < \sigma\} \boldsymbol{\xi}^{(t+1)}.$$

Then we know that

$$\left\| \boldsymbol{w}^{(k)} \right\|_\infty \le b$$

and

$$\boldsymbol{w}^{(t+1)} = (\boldsymbol{I} - \alpha_t \boldsymbol{G}_m) \boldsymbol{w}^{(t)} - \alpha_t \mathbf{1}\{t < \sigma\} \boldsymbol{\xi}^{(t+1)}.$$

The residual $y_t := \left\| \boldsymbol{e}^{(t)} - \boldsymbol{w}^{(t)} \right\|_\infty$ therefore tracks the deterministic and nonlinear parts of the error. For $n \le t < T \wedge \sigma$, subtracting the two recursions gives $\boldsymbol{e}^{(t+1)} - \boldsymbol{w}^{(t+1)} = (\boldsymbol{I} - \alpha_t \boldsymbol{G}_m)(\boldsymbol{e}^{(t)} - \boldsymbol{w}^{(t)}) - \alpha_t \boldsymbol{R}(\boldsymbol{e}^{(t)})$, and hence

$$y_{t+1} \le (1 - \mu\alpha_t) y_t + \alpha_t \left\| \boldsymbol{R}(\boldsymbol{e}^{(t)}) \right\|_\infty \le (1 - \mu\alpha_t) y_t + L_h \alpha_t (y_t + b)^2. \tag{43}$$

Now we claim that $y_k \le r_{\text{out}}/2$ for every $n \le k \le T \wedge \sigma$. Suppose $y_t \le r_{\text{out}}/2$. Then

$$\begin{aligned} L_h (y_t + b)^2 &\le 2 L_h y_t^2 + 2 L_h b^2 \\ &\le L_h r_{\text{out}} y_t + 2 L_h b^2 \le \frac{\mu}{16} y_t + 2 L_h b^2. \end{aligned}$$

Moreover, if $L_h > 0$, the two bounds on b in Equation (42) imply

$$2 L_h b^2 \le \frac{\mu b}{32} \le \frac{\mu r_{\text{out}}}{512},$$

while the same bound is immediate when $L_h = 0$. Because $0 \le \mu\alpha_t \le 1$, the coefficient $1 - 15\mu\alpha_t/16$ is nonnegative. Thus, Equation (43) and $y_t \le r_{\text{out}}/2$ yield

$$\begin{aligned} y_{t+1} &\le \left(1 - \frac{15}{16}\mu\alpha_t\right) y_t + 2 L_h \alpha_t b^2 \\ &\le \left(1 - \frac{15}{16}\mu\alpha_t\right) \frac{r_{\text{out}}}{2} + \frac{\mu\alpha_t r_{\text{out}}}{512} \\ &= \frac{r_{\text{out}}}{2} - \frac{239}{512}\mu\alpha_t r_{\text{out}} \le \frac{r_{\text{out}}}{2}. \end{aligned} \tag{44}$$

Because $y_n = \left\| \boldsymbol{e}^{(n)} \right\|_\infty \le r_{\text{out}}/2$, induction proves the claim.

Therefore we know that for every $n \le k \le T \wedge \sigma$,

$$\left\| \boldsymbol{e}^{(k)} \right\|_\infty \le y_k + \left\| \boldsymbol{w}^{(k)} \right\|_\infty \le \frac{r_{\text{out}}}{2} + b \le \frac{9 r_{\text{out}}}{16} < r_{\text{out}}.$$

If $\sigma \le T$, this estimate at $k = \sigma$ contradicts the definition of σ and thus $\sigma > T$.

Now the first inequality in Equation (44) holds on the whole interval. Moreover we have

$$\sum_{t=n}^{T-1} \alpha_t \prod_{\ell=t+1}^{T-1} \left(1 - \frac{15}{16}\mu\alpha_\ell\right) = \frac{1 - \prod_{t=n}^{T-1}\left(1 - \frac{15}{16}\mu\alpha_t\right)}{(15/16)\mu} \le \frac{16}{15\mu}.$$

Iterating the recursion of y_t gives

$$\begin{aligned} y_T &\le y_n \prod_{t=n}^{T-1}\left(1 - \frac{15}{16}\mu\alpha_t\right) + 2L_h b^2 \sum_{t=n}^{T-1} \alpha_t \prod_{\ell=t+1}^{T-1}\left(1 - \frac{15}{16}\mu\alpha_\ell\right) \\ &\le \left\|\boldsymbol{e}^{(n)}\right\|_\infty \exp\left\{-\frac{15\mu}{16}\sum_{t=n}^{T-1}\alpha_t\right\} + \frac{32L_h}{15\mu}b^2. \end{aligned}$$

Therefore we conclude that

$$\left\|\boldsymbol{e}^{(T)}\right\|_\infty \le y_T + b \le \left\|\boldsymbol{e}^{(n)}\right\|_\infty \exp\left\{-\frac{15\mu}{16}\sum_{t=n}^{T-1}\alpha_t\right\} + b + \frac{32L_h}{15\mu}b^2,$$

which completes the proof. □

E Additional Numerical Results

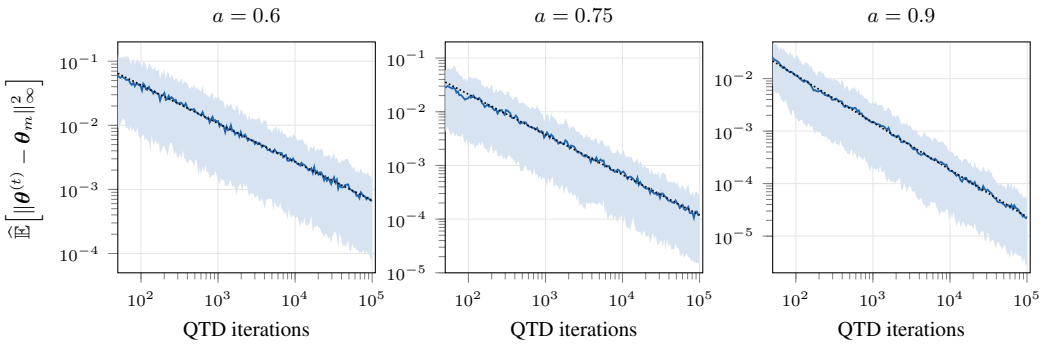

Figure 2: Squared last-iterate error for three polynomial exponents over 200 independent QTD trajectories. Blue lines and shading show the mean and 10th–90th percentiles; dotted lines have slopes fixed at $-a$.

Complete protocol for Figure 1. Panels (a)–(b) use a one-state, one-action MDP with deterministic self transition, $\gamma = 0.5$, and independent $\mathrm{Unif}[0, 1]$ rewards across iterations and replications. We use $m = 7$ quantiles at levels $\tau_i = (2i - 1)/14$, initialize all quantile locations at zero, and run 200 trajectories through $T = 10^5$. The same reward sample is shared by the seven coordinates within one QTD update, as specified in Section 2. We compute the benchmark $\boldsymbol{\theta}_m$ independently by iterating the quantile-projected Bellman operator and solving each quantile by bisection; its maximum CDF residual is 3.4×10^{-16}.

Panel (a) uses $\alpha_t = 4/(t + 20)^{0.75}$. Its dotted line has slope fixed at -0.75 and only its intercept is fitted over $t \in [5 \times 10^3, 10^5]$. Panel (b) compares $\alpha_t = 0.05$, $4/(t + 20)^{0.75}$, and $20/(t + 100)$. The base seed is 20260901, with an offset of 1000003 between schedules; Panel (a) and the polynomial curve in Panel (b) share the same trajectory set. Curves are empirical means of the squared sup-norm error at 181 logarithmically spaced checkpoints. Panel (a) additionally shows the empirical 10th–90th percentile band.

For Panel (c), define $\tau = \inf\{t \ge 0 : \left\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}_m\right\|_\infty \le r_{\mathrm{out}}/2\}$. We use the same singleton MDP construction with one quantile, $\gamma = 0.01$, and $\mathrm{Unif}[0, 1]$ rewards. In this case, $\theta_m = 0.5/(1 - \gamma) =$

0.505051, and we initialize at $\theta^{(0)} = \theta_m + r_{\mathrm{out}}$. The constants computed from the fixed-m refinement in Equation (20) are $c_{\mathcal{M},1} = 2(1-\gamma) = 1.98$, $r_{\mathrm{out}} = 0.123762$, $c_{\mathrm{g}} = 0.007581$, and $\beta = 182.503$. We use $\alpha = 0.75 c_{\mathrm{g}}/\beta = 3.116 \times 10^{-5}$ and seed 20260902. Among 5,000 independent trajectories, none is censored at horizon 2×10^5: the median entrance time is 22,429, the empirical 90th percentile is 23,577, and the maximum is 25,945.

For this deterministic initialization, $D_{0,\beta} = \Phi_\beta(\theta^{(0)} - \theta_m) - r_{\mathrm{out}}/2 = 0.061881$. Panel (c) plots the conditional bound in Equation (19) with this exact realized budget, rather than replacing it by the coarser initialization-uniform value $D_{\mathrm{g}} = 0.952049$. The empirical survival function is evaluated from the 5,000 hitting times. Zero empirical probabilities are shown at 10^{-4} solely for display on the logarithmic axis; the accompanying CSV retains the exact zero values and the unclipped theoretical probabilities.

Additional polynomial exponents. Figure 2 repeats the fixed-slope diagnostic from Figure 1(a) for $a \in \{0.6, 0.75, 0.9\}$. All other settings are unchanged. Each dotted reference has slope fixed at $-a$ and is fitted only in intercept over $t \in [5 \times 10^3, 10^5]$. Across all three exponents, the empirical curves are nearly parallel to their theoretical references after burn-in.