A Diffusion Approximation for Temporal-Difference Learning with Linear Features under Markovian Noise

Anonymous Authors[1]

Abstract

Temporal difference (TD) learning with linear function approximation is a core method for policy evaluation. Its classical continuous-time description is an ordinary differential equation (ODE), which captures the asymptotic mean dynamics but neglects stochastic fluctuations determining the error floor. We introduce a stochastic differential equation (SDE) approximation for linear TD(0) under Markovian noise. The resulting model distinguishes the contraction dynamics governed by the projected Bellman operator from the influence of Markovian sampling. As consequences, we complement classical results with a covariance dynamics, a local Ornstein-Uhlenbeck description, an explicit estimate on the mixing time influence on convergence, and a new range of admissible stepsizes.

1. Introduction and related work

TD learning is a central algorithm for policy evaluation in reinforcement learning (Sutton (1988); Sutton & Barto (2018)), with a long line of theoretical analyses.

With linear features and a fixed policy, the asymptotic behavior of TD-learning for policy evaluation is described as a deterministic dynamical system by the linear ODE

$$\dot{\theta} = b - A\theta, \tag{1}$$

which is at the heart of classical stochastic-approximation analysis of TD (Tsitsiklis & Van Roy (1997); Kushner & Yin (2003); Borkar (2008); Mou et al. (2024); Samsonov et al. (2025)). The ODE approach proves where TD should go, but, by construction, it cannot capture high order effects.

Finite-time analyses keep the discrete recursion and give non-asymptotic error bounds under i.i.d. or Markovian data (Korda & Prashanth (2015); Bhandari et al. (2018); Srikant & Ying (2019); Mitra (2024); Lee & Orabona (2025); Mou et al. (2024)). Such bounds typically have the form "decaying transient plus an $O(\alpha)$ variance floor". Recent advances also provide refined statistical and inference guarantees for linear stochastic approximation under Markovian noise (Samsonov et al. (2025)).

Our aim is complementary: we introduce a *non-asymptotic continuous-time* model that explains the leading stochastic term producing this floor. SDE approximations are well-established tools for noisy optimization algorithms, including stochastic gradient methods (Mandt et al. (2017); Li et al. (2017)). Under the name of stochastic modified equations, these provide continuous-time weak approximations, i.e. approximations of the distribution of sample paths, instead of the sample paths themselves. Through the lenses of the SDE, we manage to complement the current panorama of available stepsizes under which linear TD is guaranteed to converge and to recover recent results on non-asymptotic central limit theorems for Markovian recursions (Srikant (2026)).

We analyze the constant stepsize case, from which the classical Robbins–Monro regime follows by considering vanishing diffusion. For stepsize α, the ODE captures the $O(1)$ drift of TD, while random fluctuations around that drift are of order $\sqrt{\alpha}$. A more natural object at this scale is therefore an SDE:

$$\begin{aligned} \mathrm{d}\Theta_t &= (b - A\Theta_t)\,\mathrm{d}t + \sqrt{\alpha}\,B(\Theta_t)\,\mathrm{d}W_t, \\ B(\theta)B(\theta)^\top &= \Gamma(\theta). \end{aligned} \tag{2}$$

The change is not merely notational. Comparing to the classical ODE (1), the new term B is present. This is the term encoding the information about the noise and whose study allows to answer questions such as: How does Markovian sampling influence TD? Which feature map or policy produces less long-run TD noise? What is the variance dynamics? Moreover, the ODE path is encoded in the drift coefficient of the SDE, so that the ODE emerges from taking the expectation of SDE.

Contributions. While prior work characterizes either asymptotic behavior via ODEs or finite-time bounds via dis-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

crete analysis, our approach provides a unified continuous-time stochastic model capturing both drift and fluctuations at finite time, adding interpretability to the known finite-time results. Technically, we use the Markov-chain Poisson equation to decompose correlated TD noise into a martingale term plus a telescoping coboundary. This identifies the diffusion covariance as the long-run covariance of the TD noise. We also construct an affine factor $B(\theta)$ of this covariance. Finally, we derive stability and local Gaussian estimates that turn the diffusion approximation into a diagnostic.

2. Problem setting and assumptions

Markov decision process. We consider policy evaluation for a fixed policy. The policy induces a finite-state irreducible and aperiodic Markov chain $(S_k)_{k\geq 0}$ with stationary distribution μ. Rewards may be random; we assume that R_{k+1} is conditionally independent given (S_k, S_{k+1}), with law depending only on the transition (S_k, S_{k+1}). Thus, the observation driving the TD update $Z_k = (S_k, S_{k+1}, R_{k+1})$ is itself a Markov chain. We denote by P_Z^m its m-step transition kernel, i.e. $P_Z^m(z, A) = \mathbb{P}(Z_{k+m} \in A \mid Z_k = z)$. Its mixing behavior is inherited from the state chain (S_k), since the reward component does not introduce additional temporal dependence beyond (S_k, S_{k+1}).

Temporal difference learning. Let $\phi : \mathcal{S} \to \mathbb{R}^d$ be a feature map and $V_\theta(s) = \phi(s)^\top \theta$. TD(0) with constant stepsize α is

$$\theta_{k+1} = \theta_k + \alpha\big(R_{k+1} + \gamma\phi(S_{k+1})^\top\theta_k - \phi(S_k)^\top\theta_k\big)\phi(S_k). \tag{3}$$

For $z = (s, s', r)$, write

$$\hat{b}(z) = r\,\phi(s), \qquad \hat{A}(z) = \phi(s)(\phi(s) - \gamma\phi(s'))^\top,$$

and define

$$H(\theta, z) = \hat{b}(z) - \hat{A}(z)\theta, \qquad h(\theta) = b - A\theta,$$
$$b = \mathbb{E}_\pi[\hat{b}(Z)], \qquad A = \mathbb{E}_\pi[\hat{A}(Z)],$$

where π denotes the stationary law of Z_k. The TD fixed point is $\theta^\star = A^{-1}b$ when A is nonsingular.

Assumptions. **(A1)** The chain (S_k) is finite, irreducible, and aperiodic, with mixing profile $\varrho(m) := \sup_z \|P_Z^m(z,\cdot) - \pi\|_{\mathrm{TV}}$. **(A2)** Features and rewards are bounded: $\|\phi(s)\| \leq K_\phi$ and $|r| \leq K_R$. **(A3)** A is positive stable, i.e. every eigenvalue of A has positive real part; equivalently, the ODE matrix $-A$ is Hurwitz. In the standard discounted on-policy setting, positive stability follows from full column rank of the feature matrix in $L^2(\mu)$ (Tsitsiklis & Van Roy, 1997). Assumption (A3) is the same stability condition that makes the ODE converge to $\theta^\star$.

3. From TD discrete recursion to an SDE

The goal of this section is to construct a continuous-time stochastic model that captures both the mean dynamics and the leading-order fluctuations of the TD recursion. While classical ODE methods describe only the average behavior, they fail to account for the stochastic effects that determine the steady-state error. To address this limitation, we introduce for the first time a diffusion approximation in the form of a stochastic differential equation (SDE).

Specifically, we derive an order-1 weak approximation of the TD iterates in the sense of Li et al. (2017), namely:

$$\max_{0\leq k\leq T/\alpha} |\mathbb{E}\varphi(\theta_k) - \mathbb{E}\varphi(\Theta_{k\alpha})| \leq C\alpha^1, \qquad \varphi \text{ sufficiently regular.}$$

Remarkably, *sufficiently regular* is not at all a restrictive condition. For example, all polynomials are included, thus one is able to compare all moments. We refer to Section C for details.

Main challenges. Deriving such a model presents *three main challenges*. First, the update noise in TD is Markovian and therefore temporally correlated, preventing a direct application of standard diffusion approximations developed for i.i.d. noise. Second, the diffusion term must capture the cumulative effect of these correlations. Third, the SDE must remain well-posed even when this covariance is singular, necessitating the construction of a suitable matrix square root.

Approach. We address these challenges as follows. Using the Poisson equation associated with the underlying Markov chain, we decompose the TD noise into a martingale component and a telescoping coboundary term, which allows us to isolate the effective diffusion covariance $\Gamma(\theta)$. We then construct an affine factorization $B(\theta)$ such that $B(\theta)B(\theta)^\top = \Gamma(\theta)$, ensuring sufficient regularity of the SDE coefficients. This construction is crucial to guarantee well-posedness even when $\Gamma(\theta)$ is singular.

We conclude the section by stating a theorem that formalizes the weak convergence of the TD iterates to the solution of the SDE (2).

Poisson decomposition. Let

$$g_\theta(z) = H(\theta, z) - h(\theta), \qquad \mathbb{E}_\pi[g_\theta(Z)] = 0.$$

For each fixed θ, let u_θ be the centered solution of the Poisson equation

$$u_\theta - P_Z u_\theta = g_\theta. \tag{4}$$

Table 1. Comparison of stepsize conditions in finite-time analyses of TD/linear SA under Markovian noise.

Work	Without projection	Not function of t_{mix}	Not function of horizon T
Korda & Prashanth (2015)	✓	×	×
Bhandari et al. (2018)	×	×	✓
Srikant & Ying (2019)	×	×	×
Mitra (2024)	✓	×	✓
Mou et al. (2024)	✓	×	×
Lee & Orabona (2025)	✓	✓	×
This work	✓	✓	✓

Then the TD noise admits the decomposition

$$g_{\theta_k}(Z_k) = u_{\theta_k}(Z_k) - u_{\theta_k}(Z_{k+1}) + \xi_{k+1}, \qquad \xi_{k+1} = u_{\theta_k}(Z_{k+1}) - P_Z u_{\theta_k}(Z_k), \tag{5}$$

where (ξ_{k+1}) is a martingale difference. The first two terms in (5) form a coboundary. This is the step that converts Markovian correlation into a martingale diffusion without assuming independent samples. For details, refer to Section C.

Effective covariance. The effective covariance field is

$$\Gamma(\theta) = \mathbb{E}_\pi\big[(u_\theta(Z_1) - P_Z u_\theta(Z_0)) \cdot (u_\theta(Z_1) - P_Z u_\theta(Z_0))^\top\big]. \tag{6}$$

Equivalently, if $C_m(\theta) = \mathbb{E}_\pi[g_\theta(Z_0) g_\theta(Z_m)^\top]$, then

$$\Gamma(\theta) = C_0(\theta) + \sum_{m\geq 1} \big(C_m(\theta) + C_m(\theta)^\top\big). \tag{7}$$

If the data were i.i.d., all lag terms would vanish and $\Gamma(\theta)$ would reduce to the one-step covariance. The lag terms in (7) are the mathematical record of temporal dependence. They can either inflate or reduce the diffusion depending on the sign of the correlations while slow mixing makes the sum longer.

Proposition 3.1 (Mixing controls quadratic forms of the diffusion)**.** *Under* (A1)–(A2)*, there are constants $c_0, c_1 > 0$ depending only on the bounded TD components such that, for every $\theta \in \mathbb{R}^d$ and every $M \succeq 0$,*

$$\mathrm{tr}(M\Gamma(\theta)) \leq \mathrm{tr}(M)\, \tau_{\mathrm{corr}} (c_0 + c_1 \|\theta\|)^2, \qquad \tau_{\mathrm{corr}} = 1 + 4 \sum_{m\geq 1} \varrho(m). \tag{8}$$

Moreover, τ_{corr} is equivalent, up to universal constants, to the usual total-variation mixing time $t_{\mathrm{mix}}(1/4)$.

A proof of Proposition 3.1 is given in Section D. We remark that $\tau_{\mathrm{corr}} < \infty$ by (A1).

Worst-direction and total noise. Proposition 3.1 gives, by taking $M = vv^\top$, $\|\Gamma(\theta)\|_{\mathrm{op}} \leq \tau_{\mathrm{corr}}(c_0 + c_1 \|\theta\|)^2$, while $M = I_d$ yields $\mathrm{tr}(\Gamma(\theta)) \leq d\, \tau_{\mathrm{corr}}(c_0 + c_1 \|\theta\|)^2$. The first quantity is the largest directional variance of the TD noise, whereas the second is the total injected variance. Since $\|\Gamma(\theta)\|_{\mathrm{op}} \leq \mathrm{tr}(\Gamma(\theta)) \leq d\, \|\Gamma(\theta)\|_{\mathrm{op}}$, the gap between the two measures how spread out the noise is across parameter directions. This motivates

$$d_{\mathrm{eff}}(\theta) := \frac{\mathrm{tr}(\Gamma(\theta))}{\|\Gamma(\theta)\|_{\mathrm{op}}} \in [1, d], \tag{9}$$

whenever $\Gamma(\theta) \neq 0$. Thus τ_{corr} controls the worst-direction amplitude, while $d_{\mathrm{eff}}(\theta)$ measures how many directions are effectively noisy.

A Lipschitz diffusion coefficient. The effective covariance $\Gamma(\theta)$ aggregates temporally correlated noise and may be singular (see Section E), so its principal square root need not be Lipschitz. or the SDE to serve as a stable and interpretable continuous-time proxy for the TD iterates, its drift and diffusion coefficients should have such regularity. To this end, we construct an affine factor $B(\theta)$.

Proposition 3.2 (Affine factorization with sparsity bound)**.** *Assume* (A1)–(A2)*. Let $E = \{(s, s') : P^\pi(s'|s) > 0\}$. Then there exists an integer $q \leq \min\{d(d+1),\, 2|E|\}$ and an affine map*

$$B : \mathbb{R}^d \to \mathbb{R}^{d\times q}, \qquad B(\theta) = B_0 + \sum_{i=1}^{d} \theta_i B_i, \tag{10}$$

such that $\Gamma(\theta) = B(\theta)B(\theta)^\top$ for all θ. Hence, B is globally Lipschitz and has linear growth.

The explicit construction of such $B(\theta)$, along with its probabilistic interpretation, is given in Section E.

The SDE model for TD.

Theorem 3.3 (TD–SDE weak approximation and fluctuation limit)**.** *Assume* (A1)–(A3)*. Fix $T, R < \infty$ and let $\theta_k^R, \Theta_{k\alpha}^R$ denote the stopped TD recursion and the stopped*

SDE (2) *when they leave the ball* $\{\|\theta\| \le R\}$*, respectively. Then, for every sufficiently regular test function* φ*,*

$$\max_{0\le k\le T/\alpha} \left|\mathbb{E}\varphi(\theta_k^R) - \mathbb{E}\varphi(\Theta_{k\alpha}^R)\right| \le C_{T,R}\,\alpha.$$

Theorem 3.3 makes rigorous the statement "the SDE models TD": the SDE (2) is a weak approximation of the discrete TD recursion, i.e. the law of the SDE sampled at matching time $k\alpha$ is close to the law of the discrete iterate θ_k. A proof of this fact along with a recipe to reproduce it for general discrete recursions can be found in Section C.

3.1. Consequences: stability and Gaussian structure

An immediate payoff of the SDE formulation is that stability and covariance analysis become continuous-time calculations. Let $P = P^\top \succ 0$ solve the Lyapunov equation $A^\top P + PA = I$. The conditioning of this matrix measures how strongly the drift contracts, appearing as the constant in the deterministic part of the estimate below.

Theorem 3.4 (SDE stability and finite-time error)**.** *Assume* (A1)–(A3) *and let* B *be the affine factor in* (10)*. For all sufficiently small* α*, the SDE* (2) *has a unique global strong solution. Moreover, for* $E_t = \Theta_t - \theta^\star$*,*

$$\mathbb{E}\,\|E_t\|^2 \le C_A e^{-\rho_A t}\,\mathbb{E}\,\|E_0\|^2 + C_A\,\alpha\,\tau_{\mathrm{corr}}\,\bar{d}_{\mathrm{eff}}, \quad (11)$$

where $\bar{d}_{\mathrm{eff}} := \sup_{\theta:\Gamma(\theta)\neq 0} \frac{\mathrm{tr}(\Gamma(\theta))}{\|\Gamma(\theta)\|_{\mathrm{op}}} \le d$*.*

One may take $\rho_A \propto 1/\lambda_{\max}(P)$*, while* C_A *depends on the condition number of* P*, the linear-growth constant of* B*, and the norm of the bounded TD components.*

This recovers the usual constant-stepsize TD picture. The added information is the interpretation of the error floor. Finite-time TD analyses prove upper bounds with comparable ingredients, the SDE explains why those ingredients appear. A proof of Theorem 3.4 is presented in Section F.

Remark 3.5 (Removing localization a posteriori). The stopping in Theorem 3.3 is a technical localization device used to obtain uniform weak-error estimates on bounded sets. It is not part of the final SDE model. The localization can be removed *a posteriori*; see Section G for a rigorous proof and a brief discussion of the impact this has on the standard analysis pipeline.

Insights into stepsize choice. The stepsize parameter α in Theorem 3.4 is the same constant stepsize as in the TD recursion. Hence, the SDE stability bound directly translates into the TD regime: for sufficiently small stepsize α, the iterates remain stable and concentrate in an $O(\alpha)$ neighborhood of the TD fixed point. Table 1 places this condition in context with representative finite-time guarantees for linear TD under Markovian noise. The table should be read with some qualifications. These are discussed in Section H.

Finally, we state a comprehensive result on the noise structure of TD. We prove finite-time estimates on how far the SDE of TD is far from being an Ornstein-Uhlenbeck process and provide an explicit evolution equation of the covariance.

Theorem 3.6 (Local Gaussian and covariance dynamics)**.** *Assume* (A1)–(A2)*. Let* $X_t^\alpha = (\Theta_t - \theta^\star)/\sqrt{\alpha}$*. On every finite interval* $[0,T]$*, there exists a constant* $C_T < \infty$*, independent of* $\alpha \in (0,1]$*, such that*

$$\sup_{0\le t\le T} \|X_t^\alpha - G_t\|_{L^2} \le C_T^{1/2}\sqrt{\alpha}. \quad (12)$$

where G *is the Ornstein–Uhlenbeck process*

$$\mathrm{d}G_t = -AG_t\,\mathrm{d}t + B(\theta^\star)\,\mathrm{d}W_t, \qquad G_0 = X_0^\alpha. \quad (13)$$

Hence the leading covariance $\Sigma_t = \mathrm{Cov}(G_t)$ *satisfies*

$$\dot{\Sigma}_t = -A\Sigma_t - \Sigma_t A^\top + \Gamma(\theta^\star). \quad (14)$$

If (A3) *holds,* Σ_t *converges to the unique solution of* $A\Sigma + \Sigma A^\top = \Gamma(\theta^\star)$*.*

Theorem 3.6, whose proof is given in Section I, shows that near the fixed point TD errors are locally Gaussian. In particular, the covariance equation identifies both the size and the geometry of the constant-stepsize error floor. At stationarity,

$$\mathrm{Var}\big(v^\top(\Theta_\infty - \theta^\star)\big) \approx \alpha\, v^\top \Sigma v,$$

so different directions can have different residual variances. Moreover, if P solves $A^\top P + PA = I$, then $\mathbb{E}\,\|\Theta_\infty - \theta^\star\|^2 \approx \alpha\,\mathrm{tr}(P\Gamma(\theta^\star))$. Thus $\Gamma(\theta^\star)$ identifies noisy directions, while P weights them by how slowly they are damped by the drift. Equivalently, the residual variance is governed by the alignment between Markovian noise and weakly contracting directions.

4. Scope and conclusion

The present note introduces for the first time an SDE modeling framework for TD(0) with linear function approximation under Markovian noise, which generalizes the classical ODE description and adds interpretability to discrete finite-time bounds expliciting stability, covariance dynamics, and distributional properties otherwise not accessible. This frameworks is shown able also to refine algorithm design by introducing new admissible stepsizes. On the technical contributions are the identification of the long-run covariance $\Gamma(\theta)$, the quantification of its dependence on mixing, and the construction the affine factor $B(\theta)$.

Limitations and extensions. The present version assumes bounded rewards/features. This makes error bounds more straightforward but generalization to bounded second moments are at reach. Further natural extensions include analysis other reinforcement learning algorithms and general stochastic approximations.

Impact Statement

This paper presents theoretical work aimed at advancing the foundations of machine learning and stochastic approximation. It introduces tools for analyzing the stochastic dynamics of reinforcement-learning algorithms and does not involve datasets, deployed systems, or direct user-facing applications. Its likely impact is methodological: it may help researchers understand variance, stability, and covariance structure in TD and related stochastic approximation algorithms. We do not identify direct negative societal impacts beyond those associated with the downstream systems to which such algorithms may be applied.

References

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on Learning Theory (COLT)*, 2018.

Borkar, V. S. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

Glynn, P. W. and Meyn, S. P. A liapunov bound for solutions of the poisson equation. *The Annals of Probability*, 24(2): 916–931, 1996.

Korda, N. and Prashanth, L. A. On td(0) with function approximation: Concentration bounds and a centered variant. In *International Conference on Machine Learning (ICML)*, 2015.

Kushner, H. J. and Yin, G. G. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2 edition, 2003.

Lee, W.-C. and Orabona, F. A finite-time analysis of td learning with linear function approximation without projections or strong convexity. *arXiv preprint arXiv:2506.01052*, 2025.

Li, Q., Tai, C., and E, W. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2101–2110, 2017.

Mandt, S., Hoffman, M. D., and Blei, D. M. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Meyn, S. P. and Tweedie, R. L. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2 edition, 2009.

Mitra, A. A simple finite-time analysis of td learning with linear function approximation. *arXiv preprint arXiv:2403.02476*, 2024.

Mou, W., Pananjady, A., Wainwright, M. J., Bartlett, P. L., and Srikant, R. Optimal and instance-dependent guarantees for markovian linear stochastic approximation. *Mathematical Statistics and Learning*, 7:41–153, 2024.

Nagaraj, D., Wu, X., Bresler, G., Jain, P., and Netrapalli, P. Least squares regression with markovian data: Fundamental limits and algorithms. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16666–16676, 2020.

Samsonov, S., Sheshukova, M., Moulines, E., and Naumov, A. Statistical inference for linear stochastic approximation with markovian noise. *arXiv preprint arXiv:2505.19102*, 2025.

Srikant. Rates of convergence in the central limit theorem for markov chains, with an application to td learning. *arXiv preprint arXiv:2401.15719*, 2026.

Srikant, R. and Ying, L. Finite-time error bounds for linear stochastic approximation and td learning. In *Conference on Learning Theory (COLT)*, 2019.

Sutton, R. S. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

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

Tsitsiklis, J. N. and Van Roy, B. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

A. Notation, assumptions, and appendix roadmap

Symbol	Meaning
$Z_k = (S_k, S_{k+1}, R_{k+1})$	observation driving the TD update
$\hat{b}(z), \hat{A}(z)$	random affine components of one TD update
$H(\theta, z) = \hat{b}(z) - \hat{A}(z)\theta$	TD increment
$h(\theta) = b - A\theta$	stationary mean field
$g_\theta(z) = H(\theta, z) - h(\theta)$	centered Markovian noise
u_θ	centered Poisson solution, $u_\theta - P_Z u_\theta = g_\theta$
ξ_{k+1}	martingale increment from the Poisson decomposition
$\Gamma(\theta)$	long-run covariance of the TD noise
$B(\theta)$	affine factor satisfying $B(\theta)B(\theta)^\top = \Gamma(\theta)$
$\varrho(m)$	total-variation mixing profile of Z_k
τ_{corr}	$1 + 4\sum_{m\geq 1} \Delta(m)$

The appendix is organized as follows. Section B collects boundedness and Poisson-equation facts. Section C derives the diffusion approximation of the discrete TD recursion. Section D proves the mixing bound from Proposition 3.1. Section E constructs a Lipschitz covariance factor. Section F proves SDE stability. Section G shows how the localization introduced in Theorem 3.3 is removed by means of Theorem 3.4. Section H elaborates on the new insights about stepsize guarantees for convergence of TD. Section I proves finite-time Gaussian estimates and covariance dynamics of TD noise stated in Theorem 3.6.

B. Preliminaries

Lemma B.1 (Bounded TD components)**.** *Under* (A2)*,*

$$\sup_z \left\|\hat{b}(z)\right\| \leq K_R K_\phi, \qquad \sup_z \left\|\hat{A}(z)\right\|_{\mathrm{op}} \leq (1+\gamma)K_\phi^2.$$

Consequently, there are constants $c_0, c_1, L_g < \infty$ such that

$$\|g(\theta, z)\| \leq c_0 + c_1 \|\theta\|, \qquad \|g(\theta, z) - g(\theta', z)\| \leq L_g \|\theta - \theta'\|.$$

Proof. The first bound follows from $\|r(s, s')\phi(s)\| \leq K_R K_\phi$. For the matrix term,

$$\left\|\phi(s)(\phi(s) - \gamma\phi(s'))^\top\right\|_{\mathrm{op}} \leq \|\phi(s)\| \, \|\phi(s) - \gamma\phi(s')\| \leq (1+\gamma)K_\phi^2.$$

Since

$$g(\theta, z) = (\hat{b}(z) - b) - (\hat{A}(z) - A)\theta,$$

the Lipschitz and linear-growth bounds follow by taking suprema over the finite state space. □

Recall, that P_Z is the transition kernel of the Markov chain (Z_k), i.e.

$$P_Z(z, A) = \mathbb{P}(Z_{k+1} \in A \mid Z_k = z).$$

For finite irreducible and aperiodic chains, the mixing profile $\varrho(m) := \sup_z \|P_Z^m(z, \cdot) - \pi\|_{\mathrm{TV}}$ decays geometrically. Standard Poisson-equation results for Markov chains (Meyn & Tweedie, 2009; Glynn & Meyn, 1996) therefore give the following explicit form.

Lemma B.2 (Poisson solution)**.** *Assume* (A1)–(A2)*. For each fixed θ, the centered Poisson equation*

$$u_\theta - P_Z u_\theta = g_\theta, \qquad \mathbb{E}_\pi[u_\theta] = 0,$$

has the unique solution

$$u_\theta(z) = \sum_{m\geq 0} P_Z^m g_\theta(z).$$

Moreover, u_θ is affine in θ and there are constants $C_u, L_u < \infty$ such that

$$\|u_\theta\|_\infty \le C_u(c_0 + c_1\|\theta\|), \qquad \|u_\theta - u_{\theta'}\|_\infty \le L_u\|\theta - \theta'\|.$$

Proof. Since $\mathbb{E}_\pi[g_\theta] = 0$,

$$P_Z^m g_\theta(z) = \int g_\theta(z')\big(P_Z^m(z, dz') - \pi(dz')\big),$$

and hence

$$\|P_Z^m g_\theta\|_\infty \le 2\,\varrho(m)\,\|g_\theta\|_\infty.$$

Geometric decay of $\varrho(m)$ implies absolute convergence in sup norm and gives the stated bound. Linearity of the equation and the affine form of g_θ imply that u_θ is affine. Applying the same bound to $g_\theta - g_{\theta'}$ gives the Lipschitz estimate. Centering follows from stationarity of π and termwise integration. □

C. How the TD–SDE approximation is obtained

This section has two purposes. First, it explains the general recipe for obtaining a diffusion approximation from a discrete stochastic recursion. Second, it applies this recipe to linear TD under Markovian noise. The point is not only to justify Theorem 3.3, but also to make explicit the construction that turns a noisy recursion into a continuous-time SDE.

C.1. A short recipe: from a recursion to an SDE

We briefly recall the principle underlying stochastic modified equations and adapt it to our setting. The key idea is that a *global weak approximation* follows from a *local one-step moment matching*.

Consider a recursion of the form

$$\theta_{k+1} = \theta_k + \bar{\Delta}(\theta_k, Z_k),$$

and a candidate SDE

$$\mathrm{d}\Theta_t = h(\Theta_t)\,\mathrm{d}t + \sqrt{\alpha}\,B(\Theta_t)\,\mathrm{d}W_t.$$

Fix a point $\theta \in \mathbb{R}^d$. Let

$$\bar{\Delta}(\theta) := \theta_1 - \theta$$

denote the one-step increment of the recursion started at θ, and let

$$\Delta(\theta) := \Theta_\alpha - \theta$$

be the increment of the SDE over a time interval of length α.

The SDE is chosen so that these two increments match at the level of their leading moments. Concretely, one computes:

$$\mathbb{E}[\bar{\Delta}(\theta)] \quad \text{and} \quad \mathbb{E}[\bar{\Delta}(\theta)\bar{\Delta}(\theta)^\top],$$

and selects h and B so that the SDE increment satisfies

$$\mathbb{E}[\Delta(\theta)] = \alpha h(\theta) + O(\alpha^2),$$
$$\mathbb{E}[\Delta(\theta)\Delta(\theta)^\top] = \alpha^2 h(\theta)h(\theta)^\top + \alpha^2 B(\theta)B(\theta)^\top + O(\alpha^3).$$

If these leading terms match those of the recursion, and higher moments are controlled, we then obtain an order-1 weak approximation by results in Li et al. (2017). In particular, for any *continuous* test function φ of *at most polynomial growth*,

$$\max_{0 \le k \le T/\alpha} |\mathbb{E}\varphi(\theta_k) - \mathbb{E}\varphi(\Theta_{k\alpha})| \le C\alpha.$$

What changes under Markovian noise. When the noise is i.i.d., the covariance $B(\theta)B(\theta)^\top$ is simply the one-step covariance of $\bar{\Delta}(\theta)$. Under Markovian sampling, this is no longer correct: temporal correlations accumulate across time. The correct object is the *long-run covariance* of the noise process. Identifying this covariance is the main additional step in the TD setting.

C.2. Applying the recipe to TD

The TD recursion can be written as

$$\theta_{k+1} = \theta_k + \alpha H(\theta_k, Z_k) = \theta_k + \alpha h(\theta_k) + \alpha g_{\theta_k}(Z_k),$$

where

$$h(\theta) = b - A\theta, \qquad g_\theta(z) = H(\theta, z) - h(\theta), \qquad \mathbb{E}_\pi[g_\theta(Z)] = 0.$$

The deterministic part immediately identifies the candidate drift: $h(\theta) = b - A\theta$.

If the variables Z_k were independent, the diffusion covariance would simply be

$$\mathbb{E}_\pi[g_\theta(Z) g_\theta(Z)^\top].$$

However, in TD the observations are generated by a Markov chain, so the noise terms are correlated. The one-step covariance misses the cumulative contribution of these correlations. The main task is therefore to identify the correct effective covariance.

C.3. Removing Markovian dependence with the Poisson equation

For each fixed θ, let u_θ solve the centered Poisson equation

$$u_\theta - P_Z u_\theta = g_\theta.$$

Then

$$g_{\theta_k}(Z_k) = \xi_{k+1} + \{u_{\theta_k}(Z_k) - u_{\theta_k}(Z_{k+1})\},$$

where

$$\xi_{k+1} = u_{\theta_k}(Z_{k+1}) - P_Z u_{\theta_k}(Z_k), \qquad \mathbb{E}[\xi_{k+1} \mid \mathcal{F}_k] = 0.$$

Thus the Markovian noise is decomposed into two pieces:

$$\text{Markovian noise} = \text{martingale noise} + \text{coboundary}.$$

The coboundary is a telescoping term. Indeed, for $K \le T/\alpha$,

$$\begin{aligned}
&\alpha \sum_{k=0}^{K-1} \{u_{\theta_k}(Z_k) - u_{\theta_k}(Z_{k+1})\} \\
&= \alpha\{u_{\theta_0}(Z_0) - u_{\theta_{K-1}}(Z_K)\} + \alpha \sum_{k=1}^{K-1} \{u_{\theta_k}(Z_k) - u_{\theta_{k-1}}(Z_k)\}.
\end{aligned}$$

On a stopped ball $\{\|\theta\| \le R\}$, u_θ is bounded and Lipschitz in θ, while $\|\theta_{k+1} - \theta_k\| = O(\alpha)$. Hence

$$\sup_{K \le T/\alpha} \left\| \alpha \sum_{k=0}^{K-1} \{u_{\theta_k}(Z_k) - u_{\theta_k}(Z_{k+1})\} \right\| = O(\alpha).$$

Therefore, at the diffusion scale, TD behaves like

$$\theta_{k+1} \approx \theta_k + \alpha h(\theta_k) + \alpha \xi_{k+1}.$$

C.4. The effective covariance

The martingale increment has conditional covariance

$$Q(\theta, z) = \mathbb{E}\big[(u_\theta(Z_1) - P_Z u_\theta(z))(u_\theta(Z_1) - P_Z u_\theta(z))^\top \mid Z_0 = z\big].$$

Averaging over the stationary distribution gives the effective covariance

$$\Gamma(\theta) = \mathbb{E}_\pi[Q(\theta, Z_0)].$$

Equivalently,

$$\Gamma(\theta) = \mathbb{E}_\pi\big[(u_\theta(Z_1) - P_Z u_\theta(Z_0)) \cdot (u_\theta(Z_1) - P_Z u_\theta(Z_0))^\top\big].$$

This is also the long-run covariance of the original centered TD noise:

$$\Gamma(\theta) = C_0(\theta) + \sum_{m\geq 1} \big(C_m(\theta) + C_m(\theta)^\top\big),$$

where

$$C_m(\theta) = \mathbb{E}_\pi[g_\theta(Z_0) g_\theta(Z_m)^\top].$$

This identity is the key structural point. The diffusion covariance is not only the instantaneous variance of TD noise; it also contains all lagged correlations created by the Markov chain.

C.5. Moment matching for TD

The effective one-step TD increment is

$$\bar{\Delta}(\theta) = \alpha h(\theta) + \alpha \xi_{k+1}.$$

Its leading moments are

$$\mathbb{E}[\bar{\Delta}(\theta)] = \alpha h(\theta) + O(\alpha^2),$$

and

$$\mathbb{E}[\bar{\Delta}(\theta)\bar{\Delta}(\theta)^\top] = \alpha^2 h(\theta) h(\theta)^\top + \alpha^2 \Gamma(\theta) + O(\alpha^3).$$

The candidate SDE is therefore

$$\mathrm{d}\Theta_t = h(\Theta_t)\,\mathrm{d}t + \sqrt{\alpha}\, B(\Theta_t)\,\mathrm{d}W_t, \qquad B(\theta)B(\theta)^\top = \Gamma(\theta).$$

Over a time interval of length α, its increment satisfies the same expansions:

$$\mathbb{E}[\Delta(\theta)] = \alpha h(\theta) + O(\alpha^2),$$

and

$$\mathbb{E}[\Delta(\theta)\Delta(\theta)^\top] = \alpha^2 h(\theta) h(\theta)^\top + \alpha^2 \Gamma(\theta) + O(\alpha^3).$$

Thus the SDE matches the leading drift and the leading covariance of the TD recursion. By the first-order moment-matching theorem for stochastic modified equations,

$$\max_{0\leq k\leq T/\alpha} \Big|\mathbb{E}\varphi(\theta_k^{\alpha,R}) - \mathbb{E}\varphi(\Theta_{k\alpha}^{\alpha,R})\Big| \leq C_{T,R}\,\alpha,$$

for all φ continuous of at most polynomial growth.

C.6. Takeaway

The construction follows a simple conceptual recipe.

First, write the recursion as

$$\text{new iterate} = \text{old iterate} + \text{drift} + \text{noise}.$$

Second, identify the correct noise covariance at the diffusion scale. For i.i.d. noise this is the one-step covariance. For Markovian noise it is the long-run covariance, obtained here through the Poisson equation. Third, choose a factor $B(\theta)$ satisfying

$$B(\theta)B(\theta)^\top = \Gamma(\theta).$$

Finally, write the SDE whose one-step moments match those of the recursion:

$$\mathrm{d}\Theta_t = h(\Theta_t)\,\mathrm{d}t + \sqrt{\alpha}\, B(\Theta_t)\,\mathrm{d}W_t.$$

For linear TD under Markovian noise, this procedure yields precisely the SDE in Theorem 3.3.

D. Time correlations and diffusion size

Proof of Proposition 3.1. For a unit vector v, define $f_{\theta,v}(z) = v^\top g_\theta(z)$. It is centered and bounded by $\|g_\theta\|_\infty$. From the long-run covariance representation,

$$v^\top \Gamma(\theta) v = \mathbb{E}_\pi[f_{\theta,v}(Z_0)^2] + 2\sum_{m\geq 1} \mathbb{E}_\pi[f_{\theta,v}(Z_0) f_{\theta,v}(Z_m)].$$

The variance term is at most $\|g_\theta\|_\infty^2$. For $m \geq 1$,

$$\begin{aligned}|\mathbb{E}_\pi[f_{\theta,v}(Z_0) f_{\theta,v}(Z_m)]| &= \left| \int \pi(dz) f_{\theta,v}(z) P_Z^m f_{\theta,v}(z) \right| \\ &\leq 2\|f_{\theta,v}\|_\infty^2 \varrho(m) \leq 2\|g_\theta\|_\infty^2 \varrho(m).\end{aligned}$$

Hence, for every unit vector v,

$$v^\top \Gamma(\theta) v \leq \left(1 + 4\sum_{m\geq 1} \varrho(m)\right) \|g_\theta\|_\infty^2 .$$

By Lemma B.1,

$$v^\top \Gamma(\theta) v \leq \tau_{\text{corr}} (c_0 + c_1 \|\theta\|)^2.$$

Now let $M \succeq 0$ and write its spectral decomposition as

$$M = \sum_{j=1}^d \lambda_j v_j v_j^\top, \qquad \lambda_j \geq 0, \quad \|v_j\| = 1.$$

Since $\Gamma(\theta)$ is positive semidefinite,

$$\text{tr}(M\Gamma(\theta)) = \sum_{j=1}^d \lambda_j v_j^\top \Gamma(\theta) v_j.$$

Using the previous bound for each v_j gives

$$\text{tr}(M\Gamma(\theta)) \leq \sum_{j=1}^d \lambda_j \tau_{\text{corr}} (c_0 + c_1\|\theta\|)^2 = \text{tr}(M)\, \tau_{\text{corr}} (c_0 + c_1\|\theta\|)^2.$$

It remains to relate τ_{corr} to the usual mixing time. Let $t_{\text{mix}}(\varepsilon) = \min\{m : \varrho(m) \leq \varepsilon\}$ and choose $\varepsilon = 1/4$. Standard submultiplicativity of the pairwise total-variation distance gives

$$\sum_{m\geq 1} \varrho(m) \leq 2 t_{\text{mix}}(1/4),$$

and hence $\tau_{\text{corr}} \leq 1 + 8 t_{\text{mix}}(1/4)$. Conversely, by the definition of $t_{\text{mix}}(1/4)$, $\varrho(m) > 1/4$ for $m < t_{\text{mix}}(1/4)$, so

$$\tau_{\text{corr}} = 1 + 4\sum_{m\geq 1} \varrho(m) > t_{\text{mix}}(1/4).$$

□

E. A convenient choice of the diffusion factor

The SDE only depends on the covariance through the identity

$$B(\theta) B(\theta)^\top = \Gamma(\theta).$$

Thus $B(\theta)$ does not have to be the principal square root $\Gamma(\theta)^{1/2}$. This distinction is useful because the principal square root can be not regular enough for the SDE formulation when $\Gamma(\theta)$ is singular.

Indeed, even in dimension one, the map $a \mapsto \sqrt{a}$ is not Lipschitz near zero. Hence, without a uniform lower bound

$$\Gamma(\theta) \succeq \lambda I_d$$

on the region of interest, Lipschitz regularity of $\Gamma(\theta)^{1/2}$ is not automatic. Such a lower bound would amount to assuming that the effective TD noise excites every parameter direction. This is a nondegeneracy condition stronger than what we need.

We instead use the square-integrability of rewards to build a different factor $B(\theta)$ that is affine in θ.

Proof of Proposition 3.2. Write

$$g_\theta(z) = g^{(0)}(z) + \sum_{i=1}^{d} \theta_i g^{(i)}(z),$$

where, for $z = (s, s', r)$,

$$g^{(0)}(z) = r\phi(s) - b, \qquad g^{(i)}(z) = (A - \hat{A}(z))e_i.$$

For $i \geq 1$,

$$g^{(i)}(s, s', r) = Ae_i - \phi(s)\big(\phi_i(s) - \gamma\phi_i(s')\big),$$

so $g^{(i)}$ depends only on the transition edge (s, s'), and not on the reward sample.

By linearity of the Poisson equation,

$$u_\theta = u^{(0)} + \sum_{i=1}^{d} \theta_i u^{(i)}, \qquad u^{(i)} = \sum_{m \geq 0} P_Z^m g^{(i)}.$$

For $i \geq 1$, since $g^{(i)}$ is reward-free and the reward variable does not affect future state transitions, every $P_Z^m g^{(i)}$ is reward-free. Therefore $u^{(i)}$ is reward-free for every $i \geq 1$.

The martingale increment has the affine decomposition

$$\xi_\theta = \xi^{(0)} + \sum_{i=1}^{d} \theta_i \xi^{(i)}, \qquad \xi^{(i)} = u^{(i)}(Z_1) - P_Z u^{(i)}(Z_0).$$

For $i \geq 1$, $\xi^{(i)}$ depends only on the state-transition structure. Hence its scalar components are measurable with respect to the reachable edge (S, S'), and therefore lie in a linear space of dimension at most $|E|$.

It remains to account for the $i = 0$ coefficient. Write

$$R = \bar{r}(S, S') + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid S, S'] = 0.$$

The mean-reward part $\bar{r}(S, S')\phi(S)$ is edge-measurable, so it lies in the same edge space of dimension at most $|E|$. The centered reward noise contributes functions of the form

$$\mathbf{1}_{\{(S,S')=(s,s')\}}\varepsilon, \qquad (s, s') \in E,$$

which span at most one additional L^2-mode per reachable edge. Hence the scalar components of all coefficients

$$\{\xi^{(i)} : 0 \leq i \leq d\}$$

belong to a finite-dimensional subspace $\mathcal{H} \subset L^2$ with

$$\dim(\mathcal{H}) \leq 2|E|.$$

On the other hand, since there are $d(d+1)$ scalar components

$$\{(\xi^{(i)})_r : 0 \leq i \leq d,\ 1 \leq r \leq d\},$$

the general Hilbert-space construction gives

$$\dim(\mathcal{H}) \le d(d+1).$$

Thus we may take

$$q = \dim(\mathcal{H}) \le \min\{d(d+1),\, 2|E|\}.$$

Let $e_1, \dots, e_q$ be an orthonormal basis of $\mathcal{H}$, and define $B(\theta) \in \mathbb{R}^{d\times q}$ by

$$B_{r\ell}(\theta) = \langle (\xi_\theta)_r, e_\ell \rangle_{L^2}, \qquad 1 \le r \le d, \quad 1 \le \ell \le q.$$

Since ξ_θ is affine in θ, every entry of $B(\theta)$ is affine in θ. Hence B is globally Lipschitz and has linear growth.

Finally, because each $(\xi_\theta)_r \in \mathcal{H}$, Parseval's identity gives

$$\begin{aligned}(B(\theta)B(\theta)^\top)_{rs} &= \sum_{\ell=1}^{q} \langle (\xi_\theta)_r, e_\ell \rangle_{L^2} \langle (\xi_\theta)_s, e_\ell \rangle_{L^2} \\ &= \langle (\xi_\theta)_r, (\xi_\theta)_s \rangle_{L^2} \\ &= \mathbb{E}\big[(\xi_\theta)_r (\xi_\theta)_s\big] = \Gamma(\theta)_{rs}.\end{aligned}$$

Therefore

$$B(\theta)B(\theta)^\top = \Gamma(\theta),$$

yielding the desired conclusion. □

Remark E.1 (Edge count, sparsity, and reward-noise structure). Consider

$$E = \{(s,s') :\ P^\pi(s'|s) > 0\}, \qquad \kappa = \max_s |\{s' : P^\pi(s'|s) > 0\}|.$$

Since each state has at most κ outgoing transitions and there are n states, we always have

$$|E| \le n\kappa.$$

Equality holds when every state has exactly κ outgoing edges; otherwise $|E|$ can be strictly smaller.

The refined bound

$$q \le 2|E|$$

captures the fact that the effective noise dimension is governed by the *transition graph* rather than the ambient dimension d. However, the precise contribution of the reward depends on how reward randomness is structured.

State-dependent vs. edge-dependent noise. There are two canonical regimes:

- **State-dependent noise:**

  $$R = \bar{r}(S,S') + \varepsilon_S, \qquad \varepsilon_S \text{ depends only on } S.$$

  In this case, the centered noise contributes at most one independent direction per state. Hence the noise dimension is controlled by n, and one can obtain a sharper bound of order $n\kappa$ or even n depending on structure.

- **Edge-dependent noise:**

  $$R = \bar{r}(S,S') + \varepsilon_{S,S'}, \qquad \varepsilon_{S,S'} \text{ depends on the edge.}$$

  Here each transition (s,s') may carry its own independent noise mode. This yields up to one additional L^2-direction per edge, leading to the bound $q \le 2|E|$.

Illustration. Consider a state s with three possible successors s_1, s_2, s_3. The two regimes are illustrated below.

State-dependent noise.

Edge	Mean reward $\bar{r}(s, s')$	Noise
(s, s_1)	5	$\mathcal{N}(0, 1)$
(s, s_2)	2	$\mathcal{N}(0, 1)$
(s, s_3)	-1	$\mathcal{N}(0, 1)$

Edge-dependent noise.

Edge	Mean reward $\bar{r}(s, s')$	Noise
(s, s_1)	5	$\mathcal{N}(0, 1)$
(s, s_2)	2	0
(s, s_3)	-1	Cauchy

In the state-dependent case, all transitions share the same noise source, so the number of independent noise directions does not grow with the number of outgoing edges. In contrast, in the edge-dependent case, each transition can introduce a distinct noise mode, and the effective dimension scales with $|E|$.

Probabilistic interpretation of the Hilbertian construction. The Hilbert-space construction gives an intrinsic interpretation. The Brownian directions correspond to orthonormal L^2 modes spanning all scalar components of the martingale increment ξ_θ. Writing

$$W_t = (W_t^1, \dots, W_t^q),$$

where the components are independent one-dimensional Brownian motions, the diffusion term becomes

$$B(\theta)\,\mathrm{d}W_t = \sum_{\ell=1}^{q} B_{\cdot\ell}(\theta)\,\mathrm{d}W_t^\ell.$$

Thus each column of $B(\theta)$ describes the loading of the TD martingale noise onto one orthogonal noise mode e_ℓ. The SDE noise is therefore a finite superposition of independent Brownian perturbations associated with the finite-dimensional L^2 span generated by

$$\big\{(\xi^{(i)})_r : \ 0 \le i \le d, \ 1 \le r \le d\big\}.$$

F. Proof of SDE stability

Proof of Theorem 3.4. By Proposition 3.2, the diffusion coefficient B is affine. Hence it is globally Lipschitz and has linear growth. Since the drift

$$h(\theta) = b - A\theta$$

is also globally Lipschitz and has linear growth, the SDE

$$\mathrm{d}\Theta_t = (b - A\Theta_t)\,\mathrm{d}t + \sqrt{\alpha}\,B(\Theta_t)\,\mathrm{d}W_t$$

admits a unique global strong solution for every $\alpha > 0$.

It remains to prove the stability estimate. Let

$$\theta^\star = A^{-1}b, \qquad E_t = \Theta_t - \theta^\star.$$

Then

$$\mathrm{d}E_t = -AE_t\,\mathrm{d}t + \sqrt{\alpha}\,B(\theta^\star + E_t)\,\mathrm{d}W_t.$$

Since A is positive stable, there exists a unique symmetric positive definite matrix P solving

$$A^\top P + PA = I.$$

Define the Lyapunov function

$$V(e) = e^\top Pe.$$

Let $\lambda_{\min}(P)$ and $\lambda_{\max}(P)$ denote the extremal eigenvalues of P. Then

$$\lambda_{\min}(P)\|e\|^2 \le V(e) \le \lambda_{\max}(P)\|e\|^2.$$

Applying Itô's formula to $V(E_t)$ gives

$$\begin{aligned}\mathcal{L}V(e) &= -2e^\top PAe + \alpha\ \mathrm{tr}\big(PB(\theta^\star+e)B(\theta^\star+e)^\top\big)\\ &= -e^\top(A^\top P + PA)e + \alpha\ \mathrm{tr}(P\Gamma(\theta^\star+e))\\ &= -\|e\|^2 + \alpha\ \mathrm{tr}(P\Gamma(\theta^\star+e))\,.\end{aligned}$$

By Proposition 3.1 with $M = vv^\top$,

$$\|\Gamma(\theta)\|_{\mathrm{op}} \le \tau_{\mathrm{corr}}(c_0 + c_1\|\theta\|)^2.$$

With the effective noise dimension in (9), with the convention $d_{\mathrm{eff}}(\theta) = 0$ if $\Gamma(\theta) = 0$, we have

$$\mathrm{tr}(\Gamma(\theta)) = d_{\mathrm{eff}}(\theta)\,\|\Gamma(\theta)\|_{\mathrm{op}}\,.$$

Therefore,

$$\begin{aligned}\mathrm{tr}(P\Gamma(\theta^\star+e)) &\le \|P\|_{\mathrm{op}}\ \mathrm{tr}(\Gamma(\theta^\star+e))\\ &= \|P\|_{\mathrm{op}}\ d_{\mathrm{eff}}(\theta^\star+e)\,\|\Gamma(\theta^\star+e)\|_{\mathrm{op}}\\ &\le \lambda_{\max}(P)\,\bar{d}_{\mathrm{eff}}\,\tau_{\mathrm{corr}}\big(c_0 + c_1\|\theta^\star+e\|\big)^2,\end{aligned}$$

where

$$\bar{d}_{\mathrm{eff}} := \sup_{\theta:\Gamma(\theta)\neq 0} d_{\mathrm{eff}}(\theta) \le d.$$

Using

$$(c_0 + c_1\|\theta^\star+e\|)^2 \le 2(c_0 + c_1\|\theta^\star\|)^2 + 2c_1^2\|e\|^2\,,$$

we obtain

$$\mathcal{L}V(e) \le -\|e\|^2 + \alpha\,\tau_{\mathrm{corr}}\,\bar{d}_{\mathrm{eff}}\left(K_0 + K_1\|e\|^2\right),$$

where one may take

$$K_0 = 2\lambda_{\max}(P)(c_0 + c_1\|\theta^\star\|)^2, \qquad K_1 = 2\lambda_{\max}(P)c_1^2.$$

Choose $\alpha_0 > 0$ small enough that

$$1 - \alpha\tau_{\mathrm{corr}}\bar{d}_{\mathrm{eff}}K_1 \ge \frac{1}{2}, \qquad \forall\alpha\in(0,\alpha_0).$$

Then

$$\mathcal{L}V(e) \le -\frac{1}{2}\|e\|^2 + \alpha\,\tau_{\mathrm{corr}}\,\bar{d}_{\mathrm{eff}}K_0.$$

Using $\|e\|^2 \ge V(e)/\lambda_{\max}(P)$, this implies

$$\mathcal{L}V(e) \le -\rho_A V(e) + \alpha\,\tau_{\mathrm{corr}}K_0, \qquad \rho_A := \frac{1}{2\lambda_{\max}(P)}.$$

Taking expectations in Itô's formula and applying Gronwall's inequality yields

$$\mathbb{E}V(E_t) \le e^{-\rho_A t}\,\mathbb{E}V(E_0) + \frac{\alpha\,\tau_{\mathrm{corr}}\,\bar{d}_{\mathrm{eff}}K_0}{\rho_A}\left(1 - e^{-\rho_A t}\right).$$

Finally, converting back from V to the Euclidean norm gives

$$\begin{aligned}\mathbb{E}\|E_t\|^2 &\le \frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}e^{-\rho_A t}\,\mathbb{E}\|E_0\|^2 + \frac{\alpha\,\tau_{\mathrm{corr}}\,\bar{d}_{\mathrm{eff}}K_0}{\rho_A\lambda_{\min}(P)}\\ &= \kappa(P)\,e^{-\rho_A t}\,\mathbb{E}\|E_0\|^2 + 2\kappa(P)\,\alpha\tau_{\mathrm{corr}}\bar{d}_{\mathrm{eff}}\,K_0,\end{aligned}$$

where $\kappa(P) := \frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}$ is the conditioning number of the (symmetric) matrix P. Renaming constants gives

$$\mathbb{E}\,\|E_t\|^2 \le C_A e^{-\rho_A t}\mathbb{E}\,\|E_0\|^2 + C_A\alpha\,\tau_{\mathrm{corr}}\,\bar{d}_{\mathrm{eff}}.$$

where C_A depends on the Lyapunov geometry of A, in particular on the condition number of P, and on the linear-growth constants of B.

The same estimate also gives

$$\sup_{t\geq 0}\mathbb{E}\|\Theta_t\|^2 < \infty,$$

because $\Theta_t = E_t + \theta^\star$. Hence the unique solution to (2) has uniformly bounded second moment. □

G. Removal of localization

The weak approximation theorem is stated for stopped processes. This section records the standard argument showing that, once global stability of the SDE is known, the localization can be removed on every finite time interval.

Corollary G.1 (Removal of localization)**.** *Let $(\Theta_t^{(R)})_{t\geq 0}$ denote the unique strong solution of the SDE obtained by truncating the coefficients of (2) outside the ball $\{\|\theta\| \leq R\}$, and let $(\Theta_t)_{t\geq 0}$ denote the unique global strong solution of (2). Then for every $T > 0$,*

$$\lim_{R\to\infty}\mathbb{P}\Big(\sup_{0\leq t\leq T}\|\Theta_t^{(R)} - \Theta_t\| = 0\Big) = 1.$$

Equivalently,

$$\sup_{0\leq t\leq T}\|\Theta_t^{(R)} - \Theta_t\| \xrightarrow{\mathbb{P}} 0, \qquad R\to\infty.$$

Consequently, any weak limit identified for the localized processes coincides with the weak limit of the original process on every finite time interval.

Proof. Define the exit time of the full solution

$$\tau_R := \inf\{t \geq 0 : \|\Theta_t\| \geq R\}.$$

By construction of the localized coefficients and pathwise uniqueness of strong solutions,

$$\Theta_t^{(R)} = \Theta_t \qquad \text{for all } t \leq \tau_R.$$

Therefore,

$$\sup_{0\leq t\leq T}\|\Theta_t^{(R)} - \Theta_t\| = 0 \qquad \text{on the event } \{\tau_R > T\}.$$

Hence

$$\begin{aligned}\mathbb{P}\Big(\sup_{0\leq t\leq T}\|\Theta_t^{(R)} - \Theta_t\| = 0\Big) &\geq \mathbb{P}(\tau_R > T)\\ &= 1 - \mathbb{P}\Big(\sup_{0\leq t\leq T}\|\Theta_t\| \geq R\Big).\end{aligned}$$

The final probability converges to zero as $R\to\infty$ because Theorem 3.4 gives global existence and finite moments on every finite time interval. This proves the first claim.

The convergence in probability follows immediately: for every $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}\Big(\sup_{0\leq t\leq T}\|\Theta_t^{(R)} - \Theta_t\| > \varepsilon\Big) &\leq \mathbb{P}(\tau_R \leq T)\\ &= \mathbb{P}\Big(\sup_{0\leq t\leq T}\|\Theta_t\| \geq R\Big) \longrightarrow 0.\end{aligned}$$

The final statement follows because two processes whose sup-norm distance on $[0,T]$ converges to zero in probability have the same weak limits. □

Localization argument and the reversal of the standard analysis pipeline. The localization used above is only a proof device, standard in stochastic analysis. We are not assuming that the TD iterates are bounded as a modeling condition, nor are we modifying the algorithm by imposing boundedness. In particular, this is different from projected linear TD, where the recursion itself is changed by applying a projection step. Here the localization is introduced only temporarily to control estimates, and is then removed *a posteriori* using the stability of the full SDE. Actually, this should be regarded as ***a feature of the novel framework we are presenting***. Both ODE methods and finite-time analysis need to prove some form of iterates non-explosion before starting the analysis of the algorithms recursion. The SDE approach reverses the order: first consider a process in which the iterates are artificially made bounded, then study the stability of the obtained SDE. It will be precisely this stability that, along with convergence of the algorithm, reveals that iterates were bounded in the first place and therefore the stopping device was never active.

H. Additional discussion on stepsize comparisons

The comparison in Table 1 should be interpreted as a qualitative summary rather than a strict ordering of results. Existing works consider different stepsize schedules, including constant, decaying, and Robbins–Monro stepsizes. Guarantees obtained for sufficiently small constant stepsizes typically imply the corresponding decaying-stepsize behavior by allowing the stepsize to decrease over time. Thus, the table focuses on whether a result identifies a regime in which the admissible stepsize is uniform with respect to the horizon and does not explicitly require projection.

A second qualification concerns the dependence on mixing. Even when a stepsize condition is not written as an explicit function of the mixing time, it can still depend on the Markov chain through other problem-dependent constants. Different policies induce different transition kernels, and hence different temporal correlations and mixing behavior, as emphasized by Nagaraj et al. (2020). In this sense, "not a function of t_{mix}" means that the stated admissible regime does not require inserting an explicit upper bound on the mixing time into the stepsize choice; it does not mean that the dynamics are independent of mixing.

Our result should also be contrasted with the curvature-free slow-regime analysis of Lee & Orabona (2025). With their terminology, their stepsize condition is *curvature-free in the slow regime*, whereas our SDE stability estimate is *curvature-dependent* through the Lyapunov geometry of A, encoded by the solution P of

$$A^\top P + PA = I.$$

This dependence is natural for our purpose. The fact we recover a curvature-dependent result is due to our choice of explicitly tracking contraction and stochastic fluctuations to show how the SDE framework is able to clarify finite-time results. Nothing prevents the SDE framework to yield curvature-free results.

I. Proof of Theorem 3.6

Proof. We split the proof into four steps.

Step 1: rescaled dynamics. Since $h(\theta^\star) = 0$, dividing the SDE (2) by $\sqrt{\alpha}$ gives

$$\mathrm{d}X_t^\alpha = -AX_t^\alpha \,\mathrm{d}t + B(\theta^\star + \sqrt{\alpha}X_t^\alpha)\,\mathrm{d}W_t. \tag{15}$$

Step 2: comparison with the frozen Ornstein–Uhlenbeck process. Let G solve

$$\mathrm{d}G_t = -AG_t\,\mathrm{d}t + B(\theta^\star)\,\mathrm{d}W_t, \qquad G_0 = X_0^\alpha.$$

Set

$$D_t := X_t^\alpha - G_t.$$

Subtracting the dynamics of G from (15),

$$\mathrm{d}D_t = -AD_t\,\mathrm{d}t + \big(B(\theta^\star + \sqrt{\alpha}X_t^\alpha) - B(\theta^\star)\big)\,\mathrm{d}W_t, \qquad D_0 = 0. \tag{16}$$

By Proposition 3.2, B is affine, hence globally Lipschitz. Thus, for some $L_B < \infty$,

$$\big\|B(\theta^\star + \sqrt{\alpha}x) - B(\theta^\star)\big\|_F \le L_B\sqrt{\alpha}\,\|x\|. \tag{17}$$

Step 3: finite-time L^2 bound. Applying Itô's formula to $\|D_t\|^2$ gives

$$\begin{aligned}\mathrm{d}\,\|D_t\|^2 &= -2\langle D_t, AD_t\rangle\,\mathrm{d}t \\ &\quad + \left\|B(\theta^\star + \sqrt{\alpha}X_t^\alpha) - B(\theta^\star)\right\|_F^2\,\mathrm{d}t + \mathrm{d}M_t,\end{aligned}$$

where M_t is a martingale. Since A is fixed, there is $c_A < \infty$ such that

$$-2\langle x, Ax\rangle \le c_A\,\|x\|^2.$$

Taking expectations and using (17),

$$\mathbb{E}\,\|D_t\|^2 \le c_A \int_0^t \mathbb{E}\,\|D_s\|^2\,\mathrm{d}s + L_B^2\alpha\int_0^t \mathbb{E}\,\|X_s^\alpha\|^2\,\mathrm{d}s.$$

On every finite interval $[0, T]$, the SDE has finite second moments because its coefficients are globally Lipschitz with linear growth. Hence, for each $T < \infty$,

$$\sup_{0\le s\le T}\mathbb{E}\,\|X_s^\alpha\|^2 < \infty.$$

Consequently, for $0 \le t \le T$,

$$\mathbb{E}\,\|D_t\|^2 \le c_A\int_0^t \mathbb{E}\,\|D_s\|^2\,\mathrm{d}s + C_T\alpha.$$

By Grönwall's lemma,

$$\sup_{0\le t\le T}\mathbb{E}\,\|D_t\|^2 \le C_T\alpha.$$

Equivalently,

$$\sup_{0\le t\le T}\|X_t^\alpha - G_t\|_{L^2} \le C_T^{1/2}\sqrt{\alpha}.$$

Step 4: covariance dynamics. Since G is linear,

$$G_t = e^{-At}G_0 + \int_0^t e^{-A(t-s)}B(\theta^\star)\,\mathrm{d}W_s.$$

Thus G_t is Gaussian whenever G_0 is Gaussian, and its covariance

$$\Sigma_t := \mathrm{Cov}(G_t)$$

satisfies

$$\dot\Sigma_t = -A\Sigma_t - \Sigma_t A^\top + \Gamma(\theta^\star),$$

because $B(\theta^\star)B(\theta^\star)^\top = \Gamma(\theta^\star)$.

If (A3) holds, then $-A$ is Hurwitz. Therefore the Lyapunov equation

$$A\Sigma + \Sigma A^\top = \Gamma(\theta^\star)$$

has a unique solution, and the covariance flow converges to it:

$$\Sigma_t \to \Sigma.$$

□