THE EFFECT OF TEMPORAL RESOLUTION IN OFFLINE TEMPORAL DIFFERENCE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal Difference (TD) algorithms are the most widely employed methods in Reinforcement Learning. Notably, previous theoretical analysis on these algorithms consider the sampling time as fixed a priori, while it has been shown that the temporal resolution can impact data efficiency (Burns et al., 2023). In this work, we provide an analysis of the performance of mean-path semi-gradient TD(0) for offline value estimation, emphasizing the dependence on the temporal resolution, a factor that indeed proves to be of crucial importance. In particular, by considering the continuous-time stochastic linear quadratic dynamical systems with a fixed data-budget, the behaviour of the Mean Squared Error on value estimation shows an optimal non-trivial value for the time discretization, and that the latter impacts the reliability of the algorithm. We also show that this behavior differs from that of the Monte Carlo algorithm (Zhang et al., 2023). We verify the theoretical characterization in numerical experiments in linear quadratic system instances.

1 INTRODUCTION

Temporal Difference (TD) is a fundamental idea in Reinforcement Learning (RL) based on bootstrapping value estimates from sampled rewards and current predictions, and it has nowadays become the core method for model-free reinforcement learning algorithms. In RL, samples typically come from a sampling procedure which follows discrete time intervals, where the temporal resolution is fixed a-priori for each application. Previous studies have shown that temporal resolution is an important factor in data efficiency (Burns et al., 2023; Zhang et al., 2023) but is often overlooked in RL research. While the convergence and statistical properties of TD have been studied extensively in the literature (Sutton, 1988; Jaakkola et al., 1993; Tsitsiklis & Van Roy, 1997; Bhandari et al., 2018; Lakshminarayanan & Szepesvari, 2018; Asadi et al., 2024), little is known about the effect of temporal discretization on the TD algorithm from both theoretical and applied perspectives.

In this paper, we study the impact of temporal resolution in value estimation using TD. In particular, we look into a specific class of systems, a continuous-time linear stochastic dynamical system with quadratic instantaneous reward (see e.g. Zhang et al. (2023)):

$$\begin{cases} \mathrm{d}x(t) = ax(t)\mathrm{d}t + \sigma \mathrm{d}w(t) \\ V(x(\tau)) = -\mathbb{E}[\int_{\tau}^{\infty} \gamma^{t-\tau} q x^2(t)\mathrm{d}t] \end{cases} \tag{1}$$

where $w(t)$ is a Wiener process. The drift coefficient a is unknown, while the diffusion coefficient σ, the reward weight q and the discount factor $\gamma \in (0, 1)$ are assumed to be known. The value function $V(\cdot)$ is defined as the expected cumulative discounted reward. Estimating the infinite-horizon value $V(x(\tau))$ corresponds to policy evaluation for a fixed linear policy in the continuous-time Linear Quadratic Regulator (LQR) (Lindquist, 1990; Zhang et al., 2023). Note that the optimal policy for this problem is indeed linear in the state. We analyze the Mean-Squared Error (MSE) of the value estimate from a widely used TD algorithm, semi-gradient TD(0) (Sutton & Barto, 2018), in the offline setting, in order to understand how finite-sample properties change with respect to the temporal resolution. By leveraging the fact that for this specific type of system, we can compute the n-th moment of the state in closed form, for any n, we provide a characterization of the MSE and identify a trade-off modulated by temporal resolution. Fig. 1 illustrates the trade-off through a numerical experiment, where we plot the learning curve of an offline mean-path semi-gradient TD(0) algorithm (Bhandari et al., 2018), under two different initializations (see Appendix A.1). The

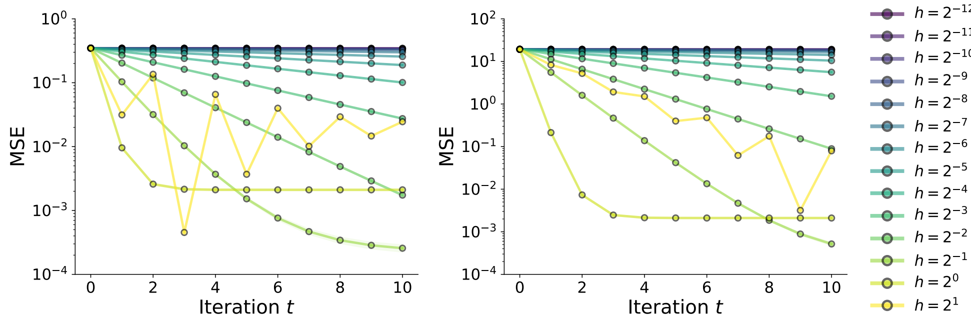

Figure 1: Learning curves of TD(0) show different behavior with respect to temporal resolution h.

result shows that the best MSE is achieved at an intermediate temporal resolution h, highlighting the existence of a non-trivial optimal discretization.

The contributions of our work are as follows. First, we develop a framework for analyzing and understanding the impact of temporal resolution in the offline value estimation accuracy of TD learning. Second, we derive an approximate expression of MSE for the Offline Mean-Path Semi-gradient TD(0), which shows a trade-off with respect to the length of the sampling intervals. We then obtain the expression of the optimal temporal resolution h^* that shows how it scales with the data budget B. These theoretical results allow us to better understand the behavior of TD algorithms with respect to temporal resolution h and data budget B. Lastly, we contrast the trade-off with that of Monte Carlo methods and offer suggestions for choosing temporal resolution in practice. We also conduct numerical experiments to validate the theoretical findings. To our best knowledge, this work represents a first step toward understanding the impact of the temporal resolution in TD methods.

2 Related works

Temporal discretization It is well known that the choice of temporal discretization can affect the performance of various RL algorithms. This literature fall into two main categories. The first one studies temporal abstraction, built on top of a base discretization. Sutton et al. (1999) formalized this in the options framework. Numerous variants have shown improved performance, particularly in video games (Sharma et al., 2017; Lakshminarayanan et al., 2017; Machado et al., 2018; Metelli et al., 2020; Dabney et al., 2021). The other line of work is concerned with the base-level discretization rather than building abstractions (Huang et al., 2019; Huang & Zhu, 2020; Park et al., 2021; Lutter et al., 2022; Farrahi & Mahmood, 2023).

The work with the most relevant problem setting to ours is a recent study by Zhang et al. (2023) which analyzed the impact of temporal discretization on the value estimation performance of Monte Carlo methods. Similar to our setting, their work focused on linear quadratic systems and provided analytical results for both finite horizon and infinite horizon settings. However, Monte Carlo methods operate in a fundamentally different way from temporal difference. It remains an open question whether the trade-off observed in their setting extends to TD learning for continuous-time systems.

Continuous-time RL Our work focuses on continuous-time dynamical systems. Although RL typically assumes a discrete-time framework, several works have applied RL to continuous-time systems (Baird, 1994; Bradtke & Duff, 1994; Doya, 2000; Wang et al., 2020; Basei et al., 2022; Jia & Zhou, 2022b). Jia & Zhou (2022a) provides a unified continuous-time formulation of various TD methods, and proved that the time-discretized version of these algorithms converge to the continuous-time counterpart in the limit of the discretization. However, the behavior and estimation error of the discretized TD algorithms with a non-zero discretization h, over a continuous state space, have yet to be characterized.

Theoretical analysis of TD Theoretical properties of TD methods have been extensively studied in the literature, as mentioned in Section 1. However, we do not revisit them here, since our focus is on understanding how TD value estimation is affected by temporal resolution. Readers interested in a recent overview of TD theory are referred to the related work sections in Tu & Recht (2018) for Least-squares based methods and in Patil et al. (2024) for stochastic-gradient based methods.

In this work, we focus on a specific algorithm of TD known as the mean-path semi-gradient TD(0), in the offline setting. Semi-gradient TD(0), a standard member of the TD family, updates parameters by following the semi-gradient of the squared TD-error with respect to the parameter (Sutton & Barto, 2018). The mean-path version, introduced by Bhandari et al. (2018), instead follows the mean negative semi-gradient under the stationary distribution. Their finite-sample analysis for mean-path TD did not account for time discretization, nor provided closed-form expressions for estimation quality — both of which are crucial for trade-off analysis. However, this algorithm serves as a good starting point for our analysis. Relatedly, Xiao et al. (2021) analyzed the fixed-point of offline semi-gradient TD(0), under finite state space and overparameterized function approximation, which differs from our setting. And they did not consider time discretization.

3 Problem setting

In this section, we describe the setting where the analysis will be performed, namely, the system, the data, the algorithm, and the objective.

3.1 Continuous-time stochastic linear quadratic system

As discussed in Section 1, the dynamics and the return of the system are given by Eq.1. Without loss of generality, we set the weight of the reward $q = 1$ and assume that the process starts at $x(0) = 0$ (Abbasi-Yadkori et al., 2011; Dean et al., 2020; Zhang et al., 2023). To ensure the value $V \in \mathbb{R}$ is finite, we assume $a < 0$. Using Lemma A.1 from (Zhang et al., 2023), we can derive the closed-form expression for the value V at $x(0)$:

$$V := V(x(0)) = \int_0^\infty \frac{\gamma^t \sigma^2}{2a} \left(1 - e^{2at}\right) \,\mathrm{d}t = \frac{-\sigma^2}{(\ln \gamma)(\ln \gamma + 2a)} \tag{2}$$

We consider a linear function approximation of the value function parameterized by θ: $V_\theta(x) = \phi(x)\,\theta$, where the value is linear in the feature $\phi(x)$. We follow Tu & Recht (2018) and choose the feature as $\phi(x) := x^2 - \frac{\sigma^2}{\ln \gamma}$. Since the value function of a linear quadratic system is quadratic in the state x, it lies exactly in the span of the features. In particular, at the initial state, we have $V_\theta(0) = \phi(0)\theta = -\frac{\sigma^2}{\ln \gamma}\theta$. Equating with Equation 2 gives the true parameter: $\theta^* = \frac{1}{\ln \gamma + 2a}$.

3.2 Offline dataset sampled at time interval h

We work with offline data sampled from the continuous-time dynamics described by Equation 1 at discrete time. The dynamics are sampled N times per trajectory, under a finite data budget B. The data collection procedure is identical to the one in Zhang et al. (2023), where data are sampled through a uniform discretization of the interval $[0, T]$, with $T < \infty$ being the *estimation horizon*, with time increment h. This results in the collection of $N = T/h$ points (which for simplicity is assumed to be an integer) over a single trajectory, at times $t_k := kh$, for $k = 0, \dots, N-1$. Given the data budget B, it is therefore possible to sample from $M = B/N$ different trajectories. At each time instant t_k of each trajectory i, the state $x_i(t_k)$ is observed and the approximate reward incurred in the interval $[t_k, t_k + h]$ is computed as $r_i(t_k) = -hx_i^2(t_k)$. The offline dataset is gathered as $\mathcal{D} = \{(x_i(t_k), r_i(t_k), x_i(t_{k+1})) \mid i = 1, 2, \dots, M \text{ and } k = 0, 1, \dots, N-2\}$.

3.3 Mean-path semi-gradient TD(0) on offline data

The semi-gradient TD(0) algorithm starts with an initial parameter estimate θ_0, which gets updated iteratively toward the true parameter θ^*. At iteration t, it updates the current estimate θ_t according to the sampled triplet containing current state, reward and next state (x, r, x'), by

$\theta_{t+1} = \theta_t + \alpha g_t(\theta_t)$ where α is the learning rate, and $g_t(\theta_t)$ is the negative semi-gradient at iteration t: $g_t(\theta_t) = \left(r + \left(\gamma^h \phi(x') - \phi(x)\right)\theta_t\right)\phi(x)$, where γ^h is the effective discount factor in the discretized system. In this work, we consider instead an *offline* version of the mean-path TD introduced by Bhandari et al. (2018), whose update rule involves the mean negative semi-gradient over some distribution rather than the stochastic gradient. In the offline setting, the mean negative semi-gradient is computed over the empirical distribution induced by the whole dataset $\mathcal{D}$, collected according to the procedure described in Section 3.2. The update rule is hence

$$\theta_{t+1} = \theta_t + \alpha \bar{g}(\theta_t), \tag{3}$$

where the mean of the negative semi-gradient is

$$\begin{aligned}\bar{g}(\theta_t) &= \overline{\phi r} + \overline{\phi(\gamma^h \phi' - \phi)}\theta_t \\ &= \frac{1}{M(N-1)} \sum_{i=1}^{M} \sum_{k=0}^{N-2} \phi\left(x_i(t_k)\right) \Big(r_i(t_k) + \left(\gamma^h \phi\left(x_i(t_{k+1})\right) - \phi\left(x_i(t_k)\right)\right)\theta_t\Big),\end{aligned} \tag{4}$$

where $\overline{\phi r}$ and $\overline{\phi(\gamma^h \phi' - \phi)}\theta_t$ are shorthands denoting taking the mean over the triplet (ϕ, r, ϕ') in the dataset.

3.4 Objective: mean-squared error of value estimation

We characterize the Mean-Squared Error of the value estimate from the offline mean-path semi-gradient TD(0) algorithm described above. It is a function of the parameter estimate θ_t after t updates: $\text{MSE}_t = \mathbb{E}\left[(V_{\theta_t} - V)^2\right]$ where V_{θ_t} and V are the infinite-horizon value estimate after t-step updates and the true value, respectively. V_{θ_t} is determined by the parameters $h, B, T, \sigma, \alpha, \theta_0, t$.

4 Theoretical results on Offline Mean-Path TD

The main goal of this section is to gather insights on the behaviour of the MSE with respect to the temporal resolution parameter h, through the analysis of the evolution of the parameter θ_t. Recall that the ground truth value is $V = -\frac{\sigma^2}{(\ln\gamma)(\ln\gamma + 2a)}$. With t step update with the semi-gradient, we have the value estimate $V_{\theta_t} = -\frac{1}{\ln\gamma}\sigma^2\theta_t$. The corresponding MSE can be expressed as follows:

$$\text{MSE}_t = \mathbb{E}\left[(V_{\theta_t} - V)^2\right] = \frac{\sigma^4}{(\ln\gamma)^2}\left(\mathbb{E}[\theta_t^2] - \frac{2\mathbb{E}[\theta_t]}{\ln\gamma + 2a} + \left(\frac{1}{\ln\gamma + 2a}\right)^2\right), \tag{5}$$

where the expectation is taken w.r.t. the distribution of the data generated by the process $x(\cdot)$.

4.1 MSE for offline mean-path semi-gradient TD(0)

The following theorem provides the characterization of the MSE for Offline Mean-Path Semi-gradient TD(0) after t updates, provided the discretization step-size is small: $h \in (0, 1)$.

Theorem 4.1 (Mean Squared Error)**.** *After t updates, the mean squared error is*

$$\begin{aligned}\text{MSE}_t = \frac{\sigma^4}{(\ln\gamma)^2}\Bigg\{ & \left[t^2\alpha^2\mathcal{I}_3 + 2t\alpha\theta_0\left(\mathcal{I}_1 + (2t-1)\alpha\mathcal{I}_5\right) + \theta_0^2\left(1 + 2t\alpha\mathcal{I}_2 + t(3t-2)\alpha^2\mathcal{I}_4\right)\right] \\ & - \frac{2}{\ln\gamma + 2a}\left[\theta_0 + t\alpha(\mathcal{I}_1 + \mathcal{I}_2\theta_0) + \frac{t(t-1)}{2}\alpha^2\left(\mathcal{I}_5 + \mathcal{I}_4\theta_0\right)\right] + \left(\frac{1}{\ln\gamma + 2a}\right)^2\Bigg\} + \mathcal{O}(h^3)\end{aligned} \tag{6}$$

where $\mathcal{I}_1, \cdots, \mathcal{I}_5$ are auxiliary terms dependent on h but not t, α, introduced in Appendix A.2. Importantly, the MSE can be expressed as:

$$\text{MSE}_t = C_0 + C_1 h + C_2 h^2 + \mathcal{O}(h^3) \tag{7}$$

where $C_0 \geq 0$, $C_1 \leq 0$, $C_2 \geq 0$ are constants with respect to h, given by:

$$
\begin{aligned}
C_0 &= \frac{\sigma^4}{(\ln\gamma)^2}\left(\theta_0 - \frac{1}{\ln\gamma + 2a}\right)^2, \\
C_1 &= \frac{t\alpha\sigma^4}{(\ln\gamma)^2}\left(\theta_0 - \frac{1}{\ln\gamma + 2a}\right)^2\left[-2\left(2a+\ln\gamma\right)C_{11} + \frac{\alpha(2t-1)\left(2a+\ln\gamma\right)^2 C_{31}}{B}\right], \\
C_2 &= \frac{t\alpha\sigma^4}{(\ln\gamma)^2}\left(\theta_0 - \frac{1}{\ln\gamma + 2a}\right)^2\Big[2C_{23} - 2(2a+\ln\gamma)C_{12} + \\
&\qquad\qquad (C_{11}^2 + \frac{C_{320}}{B})(2a+\ln\gamma)^2(2t-1)\alpha\Big].
\end{aligned}
$$

The constants $C_{11} < 0, C_{12} > 0, C_{23} > 0, C_{31} < 0, C_{320} > 0$ depend only on $a, T, \ln\gamma, \sigma^4$, and their precise forms are given in Appendix A.2.

The theorem presents the expression for the t-step MSE in Equation 6. In order to clearly exhibit the order of h in the MSE, we derive another approximate form of t-step MSE in Equation 7, offering more interpretable insights. For small h, the MSE approximately follows a quadratic relation in h, and the minimum is attained when h is strictly positive, i.e., $h^* > 0$. It confirms the existence of a trade-off in the temporal resolution parameter for the offline mean-path semi-gradient TD(0).

4.2 Optimal temporal resolution h^*

The optimal discretization step-size h^* represents the time interval at which we would ideally sample our dynamical system in order to have the best estimation of the value in term of the MSE. A precise form for this optimal parameter can be found by exploiting the approximate expression of the MSE in Equation 7, as shown in the next corollary.

Corollary 4.2 (Optimal Discretization)**.** *The optimal h^* based on the approximation Equation 7 after t updates is*

$$
h^* \approx -\frac{C_1}{2C_2} = -\frac{-2\left(2a+\ln\gamma\right)C_{11} + \frac{\alpha(2t-1)(2a+\ln\gamma)^2 C_{31}}{B}}{2\left[2C_{23} - 2(2a+\ln\gamma)C_{12} + (C_{11}^2 + \frac{C_{320}}{B})(2a+\ln\gamma)^2(2t-1)\alpha\right]}, \tag{8}
$$

and the minimum MSE is

$$
\mathrm{MSE}_t^* \approx \frac{\sigma^4\left(\theta_0 - \frac{1}{\ln\gamma+2a}\right)^2}{(\ln\gamma)^2}
\left[1 - \frac{4t a\alpha\left(-2\left(2a+\ln\gamma\right)C_{11} + \frac{\alpha(2t-1)(2a+\ln\gamma)^2 C_{31}}{B}\right)^2}{2C_{23} - 2(2a+\ln\gamma)C_{12} + (C_{11}^2 + \frac{C_{320}}{B})(2a+\ln\gamma)^2(2t-1)\alpha}\right]. \tag{9}
$$

The expression in Equation 8 is clearly dependent on the specific dynamical system or environment at hand. Therefore setting the time discretization to the optimal value would be impossible without full knowledge of the dynamics. Although it is possible to empirically find the optimal temporal resolution by sweeping over different discretization intervals, it would be impractical to sample the dataset at different frequencies just to maintain the one that has proved the most effective in terms of the MSE for the value estimation. On the other hand, if the $1/B$ terms are relatively small, the resulting optimal h would be insensitive to the change in B. We will show empirically in Section 5 that it is indeed the case.

For large enough data budgets B, we can show that the optimal time discretization h^* is independent from the data budget, and further simplify the expressions, shown in the next corollary.

Corollary 4.3 (Asymptotic Optimal Discretization)**.** *(i) If the budget B is large while the horizon T is fixed and finite, one can obtain*

$$\begin{aligned}\mathrm{MSE}_t = &\left\{1 + t\alpha\left[-2\left(2a + \ln\gamma\right)\left(C_{11}h + C_{12}h^2\right) + 2C_{23}h^2 + C_{11}^2 h^2 (2a + \ln\gamma)^2 (2t-1)\alpha\right]\right\} \\ &* \frac{\sigma^4}{(\ln\gamma)^2}\left(\theta_0 - \frac{1}{\ln\gamma + 2a}\right)^2 + \mathcal{O}(\frac{1}{B}) + \mathcal{O}(h^3). \\ h^* \approx &-\frac{-2\left(2a + \ln\gamma\right)C_{11}}{2\left[2C_{23} - 2(2a+\ln\gamma)C_{12} + C_{11}^2(2a+\ln\gamma)^2(2t-1)\alpha\right]}.\end{aligned}$$

(ii) If the horizon T is large (and thus B is also large, since $B = \frac{TM}{h}$), we have

$$\begin{aligned}\mathrm{MSE}_t = &\frac{\sigma^4}{(\ln\gamma)^2}\left(\theta_0 - \frac{1}{\ln\gamma + 2a}\right)^2\left\{1 + t\alpha\left[\frac{\sigma^4(2a+\ln\gamma)(2a+3\ln\gamma)}{2a^2\ln\gamma}h\right.\right. \\ &\left.\left. + (2a+\ln\gamma)^2\left(\frac{3\sigma^4}{4a^2} + \frac{\sigma^8(2a+3\ln\gamma)^2(2t-1)\alpha}{16a^4(\ln\gamma)^2}\right)h^2\right]\right\} + \mathcal{O}(\frac{1}{T}) + \mathcal{O}(h^3). \\ h^* \approx &-\frac{4a^2\ln\gamma(2a+3\ln\gamma)}{(2a+\ln\gamma)\left(12a^2(\ln\gamma)^2 + \sigma^4(2a+3\ln\gamma)^2(2t-1)\alpha\right)}\end{aligned}$$

Remark 4.4. The two cases in Corollary 4.3 are consistent: letting T be large in (i) recovers the expression in (ii).

How to choose temporal resolution for TD The fact that h^* is insensitive to the data budget B has important practical implications. An optimal h^* can be efficiently determined by performing a grid search on h using a baseline data budget B_0. Concretely, we can consider an initial "burn-in" phase: collect a dataset of size B_0, estimate the value V via Monte Carlo as in Zhang et al. (2023), and perform a grid search over h based on the empirical MSE, by sub-sampling this dataset. Then increasing B can verify if h^* remains stable. If so, the same h can be reused for larger data budgets, thereby reducing hyperparameter search costs while maintaining accurate value estimation.

4.3 Comparison with Monte Carlo

Recent work by Zhang et al. (2023) established that Monte Carlo (MC) estimation exhibits a trade-off in MSE w.r.t. h, under the same problem setting as ours. They derived the exact MSE expression (Theorem 3.6 in Zhang et al. (2023)) and showed that $\mathrm{MSE}_{\mathrm{MC}} = \mathcal{O}(\frac{1}{hB} + h)$. They further demonstrated that the optimal h scales polynomially with B, namely: $h^*_{\mathrm{MC}} \approx B^{-1/2}$. In contrast, our analysis indicates that for TD learning, the optimal step-size h^* behaves differently – it remains largely constant w.r.t. B.

To build intuition, consider how variance reacts to the changes in the data budget B. TD implicitly performs a maximum-likelihood fit of the value-function parameters within its chosen model (Sutton & Barto, 2018). Once sufficient data are available to obtain a stable parameter estimate, additional samples yield little further variance reduction. This explains why the trade-off and hence h^* is largely insensitive to B. In contrast, the Monte-Carlo estimator in Zhang et al. (2023) directly averages returns. Increasing B continues to reduce trajectory variance, hence affecting the trade-off.

In the next section, we present numerical experiments that illustrate and confirm these theoretical differences between TD and MC estimation.

5 Numerical experiments

To empirically validate our theoretical analysis in the previous section, we conduct simulations on continuous-time stochastic linear quadratic systems. While our theoretical framework characterizes the trade-off in Langevin dynamics, we investigate whether these insights hold for TD in practice, especially for multi-step updates. By systematically varying temporal resolution, data budget, and system parameters, we quantify how the discretization choices impact the MSE of the value estimation of TD. We also perform a comparison between TD and Monte Carlo methods.

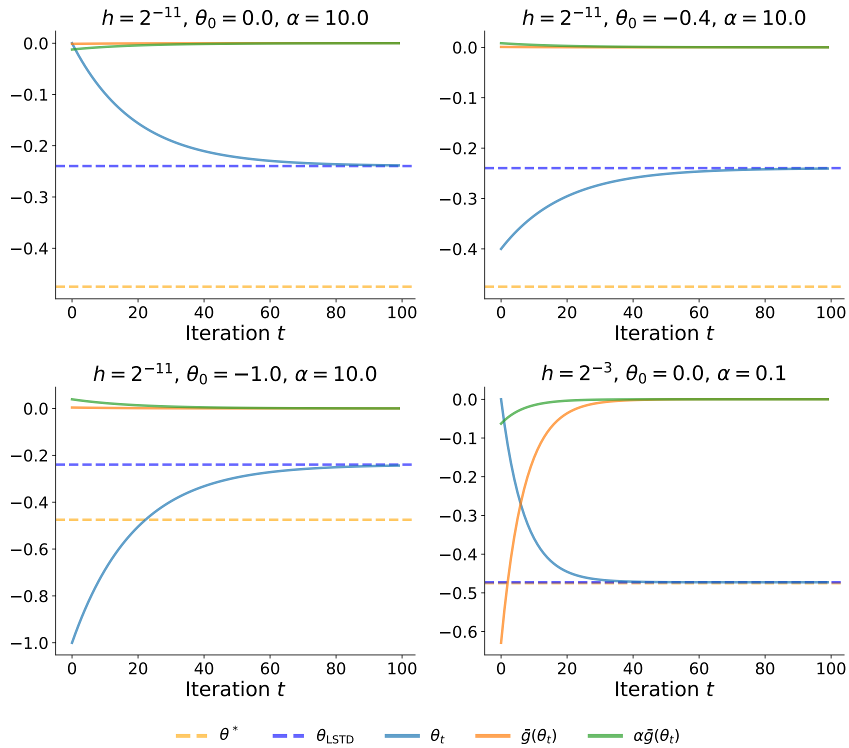

Figure 2: Trajectory of the parameter θ_t as it converges to the fixed point θ_{LSTD}

5.1 Offline TD on linear quadratic systems

In our experiments, we perform 50 independent runs to approximate the expectation in the MSE computation. In each run, we generate a new dataset by simulating the Langevin process of Section 3.1 with a unique random seed, following the procedure outlined in Section 3.2. We then apply the offline mean-path semi-gradient TD(0) algorithm, as described in Section 3.3, to obtain an estimate and compute the squared error relative to the true value. The lines in the plots represent the mean squared error averaged over the 50 runs, while the shaded regions indicate the standard error. We fix the parameter $\sigma = 1$ throughout the experiments. The values of h is chosen from this grid: $h \in \left(\{2^{-15}, 2^{-14}, \cdots, 2^{-2}\}\right) T$.

Trajectory and convergence of the iterates: In order to understand the evolution over updates of the parameter θ_t, when following the gradient dynamics in equation 3, we can start by looking at the fixed points of the latter. If $\bar{\theta}$ is a fixed point of the gradient dynamics, then from equation 3 we have that $\bar{\theta}$ must satisfy $\bar{g}(\bar{\theta}) = 0$. From equation 4 we then derive:

$$\begin{aligned}\bar{\theta} &= -\left(\overline{\phi(\gamma^h \phi' - \phi)}\right)^{-1} \overline{\phi r} \\ &= -\left(\sum_{i=1}^{M}\sum_{k=0}^{N-2} \phi\left(x_i(t_k)\right)\left[\gamma^h \phi\left(x_i(t_{k+1})\right) - \phi\left(x_i(t_k)\right)\right]\right)^{-1} \sum_{i=1}^{M}\sum_{k=0}^{N-2} \phi\left(x_i(t_k)\right) r_i(t_k), \qquad (10)\end{aligned}$$

which represents the unique fixed point, and it coincide with the LSTD estimate θ_{LSTD}. Convergence to the LSTD estimate is empirically shown in Figure 2, where the evolution of the parameter θ_t converges to the unique fixed point, and indeed the average gradient converges to 0. From Figure 2 one can note that θ_t converges to the LSTD estimate even if it starts closer to the true parameter θ^*, as is the case in top right plot, while convergence to the optimal parameter is achieved only if the latter coincide with θ_{LSTD}, as shown in bottom right plot.

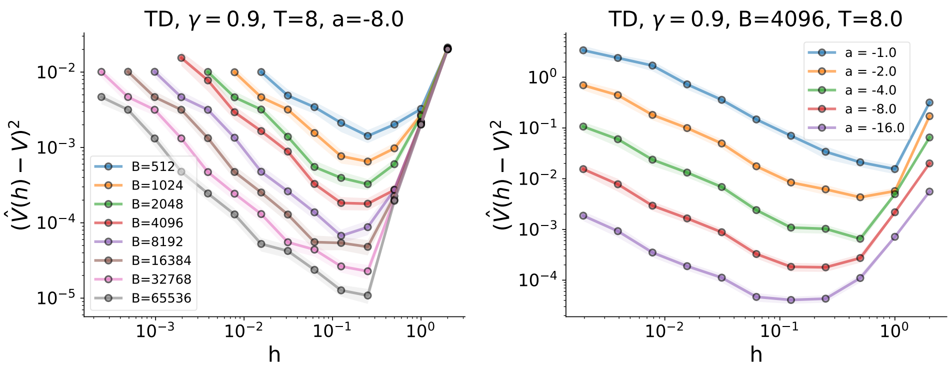

Figure 4: MSE under varying B and a, respectively

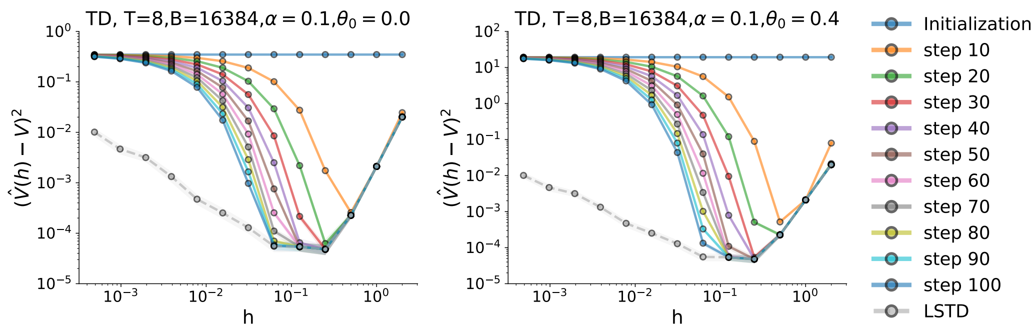

Figure 5: Empirical MSE as a function of h as number of steps increase

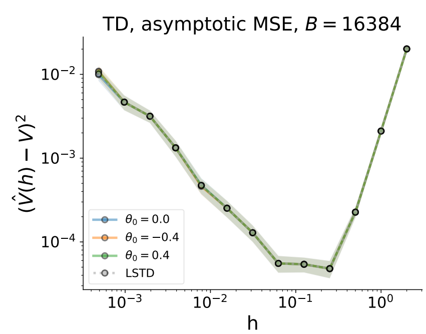

Figure 3: Asymptotic MSE

Asymptotic MSE vs h: In Figure 3, we illustrate how the asymptotic MSE varies with h, under the parameters $a = -8, T = 8, \gamma = 0.9$. For each h, the learning rate is optimized from $\{0.1, 1.0, 10.0\}$ and TD is run until convergence. The plot shows the MSE for three different initializations of θ_0. In all cases, the iterates converge to the LSTD estimate, consistent with the earlier discussion on convergence.

Dependence of MSE and h^* on the data budget B: We plot the asymptotic MSE of TD as a function of B while keeping other parameters fixed to $a = -8, T = 8, \gamma = 0.9, \theta_0 = 0$. As shown in Figure 4 (left), increasing B generally reduces the MSE, since more data yields more accurate estimates. However, varying B has negligible effect on the optimal step size h^*. It aligns with the trend in Figure 10 for one-step TD (Appendix), where h^* remain stable across different B.

MSE under varying dynamics parameter a: Figure 4 (right) illustrates the asymptotic MSE when we vary system dynamics parameter a over $\{-1, -2, -4, -6, -8, -16\}$. The other parameters are fixed to $T = 8, B = 4096, \gamma = 0.9, \theta_0 = 0$. As $|a|$ increases, the MSE across all step sizes h decreases as the system decays faster.

MSE at various number of updates t: Figure 5 illustrates how the MSE evolves w.r.t h over update steps, under two different algorithm parameter settings while keeping the system parameters fixed at $a = -8, T = 8, B = 16384, \gamma = 0.9$. In both plots, the algorithm is run for 100 update steps for each fixed h, with learning rate $\alpha = 0.1$. The left plot starts from $\theta_0 = 0$, while the

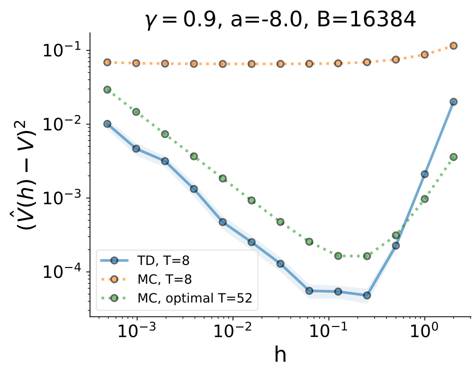

Figure 6: MSE of TD compared with MC

right plot starts from $\theta_0 = 0.4$. In both cases, the MSE decreases with the number of updates and converges quickly. However, the trade-off in MSE w.r.t h persists as the updates progress. Notably, the optimal step size h^* appears to remain stable once the number of updates t is sufficiently large.

5.2 Comparing the value estimation accuracy of TD and Monte-Carlo

To gain more insights into the value estimation accuracy of TD and MC, we evaluate the MSE of TD with multi-step updates, and compare it against both MC with the same T and the theoretically optimal MSE* that MC could achieve, in Figure 6. The optimal MC performance is obtained by optimizing the expression of its MSE w.r.t both T and h, which occurs at $T \approx 52$. The results show TD outperforms the optimal MC performance. This demonstrates that, when appropriately tuned, TD is a highly effective method for value estimation.

6 Limitations and future work

While our work provides a framework for understanding the impact of temporal resolution in TD, it has a limited scope. Our analysis is confined to a specific class of systems and algorithms. In particular, we focus on one-dimensional Langevin systems and study the offline mean-path semi-gradient TD(0) algorithm. As a result, the extent to which our findings generalize to more complex dynamical systems and alternative TD algorithms remains an open question. Exploring how value estimation responds to temporal resolution in broader settings, including higher-dimensional, non-linear environments and different learning paradigms, is an important direction for future work.

7 Conclusion

In this work, we provided a theoretical and empirical investigation into the impact of temporal resolution on offline Temporal Difference value estimation. By analyzing the Mean-Squared Error of the mean-path semi-gradient TD(0) algorithm in continuous-time stochastic linear quadratic systems, we demonstrated the existence of a non-trivial trade-off in step size h where an optimal discretization improves estimation accuracy. Our analysis further revealed that unlike Monte Carlo estimation, where the optimal h scales polynomially with the data budget B (Zhang et al., 2023), the optimal h for TD remains largely invariant to B. This provides practical guidance: one can select an appropriate temporal resolution under small data budgets without re-tuning for larger data.

Through extensive numerical experiments, we verified our theoretical predictions and explored the behavior of TD estimation across different system parameters. Additionally, we compared TD with MC and showed that TD can outperform MC under the same data budget.

This work establishes a framework for analyzing the role of temporal resolution in TD methods, contributing to a deeper understanding of how step size influences learning dynamics. Future directions include extending this analysis to more complex environments, higher-dimensional systems, and alternative TD formulations.

Reproducibility statement

The assumptions underlying our theoretical results are stated in the main text, and complete proofs are provided in the Appendix. The supplementary materials contain the Mathematica scripts and data used for symbolic computations supporting our analysis of one-step and multi-step MSE. To illustrate the complexity of the expressions, we also provide the exact formula for the one-step MSE. In addition, we include the Python code used to conduct the offline TD numerical experiments.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Online least squares estimation with self-normalized processes: An application to bandit problems. *arXiv preprint arXiv:1102.2670*, 2011.

Kavosh Asadi, Shoham Sabach, Yao Liu, Omer Gottesman, and Rasool Fakoor. Td convergence: An optimization perspective. *Advances in Neural Information Processing Systems*, 36, 2024.

Leemon C Baird. Reinforcement learning in continuous time: Advantage updating. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 4, pp. 2448–2453. IEEE, 1994.

Matteo Basei, Xin Guo, Anran Hu, and Yufei Zhang. Logarithmic regret for episodic continuous-time linear-quadratic reinforcement learning over a finite-time horizon. *Journal of Machine Learning Research*, 23(178):1–34, 2022.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pp. 1691–1692. PMLR, 2018.

Steven J. Bradtke and Michael O. Duff. Reinforcement learning methods for continuous-time Markov decision problems. In *Advances in Neural Information Processing Systems*, 1994.

Kaylee Burns, Tianhe Yu, Chelsea Finn, and Karol Hausman. Offline reinforcement learning at multiple frequencies. In *Conference on Robot Learning*, pp. 2041–2051. PMLR, 2023.

Will Dabney, Georg Ostrovski, and Andre Barreto. Temporally extended ϵ-greedy exploration. In *Proceedings of the International Conference on Learning Representations*, 2021.

Sarah Dean, Horia Mania, Nikolai Matni, Benjamin Recht, and Stephen Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2020.

Kenji Doya. Reinforcement learning in continuous time and space. *Neural computation*, 12(1): 219–245, 2000.

Homayoon Farrahi and A Rupam Mahmood. Reducing the cost of cycle-time tuning for real-world policy optimization. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2023.

Yunhan Huang and Quanyan Zhu. Infinite-horizon linear-quadratic-Gaussian control with costly measurements. *arXiv preprint arXiv:2012.14925*, 2020.

Yunhan Huang, Veeraruna Kavitha, and Quanyan Zhu. Continuous-time Markov decision processes with controlled observations. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 32–39. IEEE, 2019.

Tommi Jaakkola, Michael Jordan, and Satinder Singh. Convergence of stochastic iterative dynamic programming algorithms. *Advances in neural information processing systems*, 6, 1993.

Yanwei Jia and Xun Yu Zhou. Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *Journal of Machine Learning Research*, 23(154):1–55, 2022a.

Yanwei Jia and Xun Yu Zhou. Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *Journal of Machine Learning Research*, 23(154):1–55, 2022b.

Aravind S. Lakshminarayanan, Sahil Sharma, and Balaraman Ravindran. Dynamic action repetition for deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

Chandrashekar Lakshminarayanan and Csaba Szepesvari. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1347–1355. PMLR, 09–11 Apr 2018.

Anders Lindquist. Linear stochastic systems. *SIAM Review*, 32(2):325–328, 1990. doi: 10.1137/1032067.

Michael Lutter, Boris Belousov, Shie Mannor, Dieter Fox, Animesh Garg, and Jan Peters. Continuous-time fitted value iteration for robust policies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018.

Alberto Maria Metelli, Flavio Mazzolini, Lorenzo Bisi, Luca Sabbioni, and Marcello Restelli. Control frequency adaptation via action persistence in batch reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2020.

Seohong Park, Jaekyeom Kim, and Gunhee Kim. Time discretization-invariant safe action repetition for policy gradient methods. *Advances in Neural Information Processing Systems*, 34, 2021.

Gandharv Patil, Prashanth L. A., Dheeraj Nagaraj, and Doina Precup. Finite time analysis of temporal difference learning with linear function approximation: Tail averaging and regularisation, 2024.

Sahil Sharma, Aravind S. Lakshminarayanan, and Balaraman Ravindran. Learning to repeat: Fine grained action repetition for deep reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2017.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3:9–44, 1988.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

Richard S. Sutton, Doina Precup, and Santinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112:181–211, 1999.

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

Stephen Tu and Benjamin Recht. Least-squares temporal difference learning for the linear quadratic regulator. In *International Conference on Machine Learning*, pp. 5005–5014. PMLR, 2018.

Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21(198): 1–34, 2020.

Chenjun Xiao, Bo Dai, Jincheng Mei, Oscar A Ramirez, Ramki Gummadi, Chris Harris, and Dale Schuurmans. Understanding and leveraging overparameterization in recursive value estimation. In *International Conference on Learning Representations*, 2021.

Zichen Zhang, Johannes Kirschner, Junxi Zhang, Francesco Zanini, Alex Ayoub, Masood Dehghan, and Dale Schuurmans. Managing temporal resolution in continuous value estimation: A fundamental trade-off. In *Advances in Neural Information Processing Systems*, volume 36, pp. 62519–62548, 2023.

A Appendix

A.1 Parameters used for the example in Fig. 1

The parameters that we use in the example are: $a = -8, T = 8, B = 16384, \sigma = 1, \gamma = 0.9, \alpha = 0.1$. The two plots differ in their initialization of θ: $\theta_0 = 0$ for the left and $\theta_0 = 0.4$ for the right. We run the offline mean-path semi-gradient TD(0) algorithm described in Section 3.3 under each h for 10 updates. The result is averaged across 50 runs and shaded area is the standard error.

A.2 Proof of Theorem 4.1

In this section, we provide the technical derivations that support the theoretical results in Section 4. We begin by stating key lemmas related to the moments of the stochastic process, which form the basis of the MSE derivation. We then derive the exact MSE after one-step update, followed by an approximate form for multi-step updates, to highlight the dependence on the step-size h.

The analysis of the MSE needs a clear understanding of $\mathbb{E}[\theta_t^2]$ and $\mathbb{E}[\theta_t]$. We shall first state the following lemma that is used in the computation.

Lemma A.1. *(Even moments) Let $x(t) = \sigma \int_0^t e^{a(t-s)} \mathrm{d}w(s)$ be the solution of the Langevin equation, the moments of $x(t)$ is given by*

$$\mathbb{E}\left[x^{2n}(t)\right] = \frac{(2n-1)!!\sigma^{2n}}{(2a)^n}(e^{2at}-1)^n, \tag{11}$$

The joint moments, e.g. $\mathbb{E}\left[x^{2n}(t)x^{2m}(s)\right]$ for $s \le t$ and all nonegative integers m, n can be derived from Equation 11.

Proof. Equation 11 can be derived using induction and Itô's formula as follows.

Let $f(x) = x^{2n}$, by Itô's formula, we have

$$\mathrm{d}f(x(t)) = f'(x(t))\mathrm{d}x(t) + \frac{1}{2}f''(x(t))d[x,x]_t.$$

Substituting $f'(x) = 2nx^{2n-1}$, $f''(x) = 2n(2n-1)x^{2n-2}$ and $\mathrm{d}x(t) = ax(t)\mathrm{d}t + \sigma\mathrm{d}w(t)$, one has

$$\mathrm{d}f(x(t)) = 2nx^{2n-1}(t)\left(ax(t)\mathrm{d}t + \sigma\mathrm{d}w(t)\right) + \frac{1}{2}2n(2n-1)x^{2n-2}(t)\sigma^2\mathrm{d}t.$$

Noticing the stochastic integral has zero expectation ($\mathbb{E}[x^{2n-1}(t)dw(t)] = 0$), the expectation of the above equation gives

$$\frac{d}{dt}\mathbb{E}[x^{2n}(t)] = 2na\mathbb{E}[x^{2n}(t)] + n(2n-1)\sigma^2\mathbb{E}[x^{2n-2}(t)]\,. \tag{12}$$

Taking $M_n(t) = \mathbb{E}[x^{2n}(t)]$, the above equation gives

$$\frac{dM_n(t)}{dt} = 2naM_n(t) + n(2n-1)\sigma^2 M_{n-1}(t)\,. \tag{13}$$

Now, we show $M_n(t) = \frac{(2n-1)!!\sigma^{2n}}{(2a)^n}(e^{2at}-1)^n$ by induction using the above recurrence form.

It is trivial when $n = 1$. Assume for some $n \ge 1$, $M_n(t) = \frac{(2n-1)!!\sigma^{2n}}{(2a)^n}(e^{2at}-1)^n$ holds. By Equation 13

$$\begin{aligned}\frac{dM_{n+1}(t)}{dt} &= 2(n+1)aM_{n+1}(t) + (n+1)(2n+1)\sigma^2 M_n(t)\\ &= 2(n+1)aM_{n+1}(t) + (n+1)(2n+1)\sigma^2\frac{(2n-1)!!\sigma^{2n}}{(2a)^n}(e^{2at}-1)^n\,.\end{aligned} \tag{14}$$

We solve this linear ODE with integrating factor $e^{-2(n+1)at}$ and obtain

$$\frac{d}{dt}\left(e^{-2(n+1)at}M_{n+1}(t)\right) = (n+1)(2n+1)\sigma^{2(n+1)}\frac{(2n-1)!!}{(2a)^n}e^{-2(n+1)at}(e^{2at}-1)^n\,.$$

Integrating both sides from 0 to t,

$$\left(e^{-2(n+1)at}M_{n+1}(t)\right) = (n+1)(2n+1)\sigma^{2(n+1)}\frac{(2n-1)!!}{(2a)^n}\int_0^t e^{-2(n+1)as}(e^{2as}-1)^n ds\,,$$

which can be solved directly using change of variables (e.g., $z = e^{2as}-1$), and therefore we obtain

$$M_{n+1}(t) = \mathbb{E}\left[x^{2(n+1)}(t)\right] = \frac{(2n+1)!!\sigma^{2(n+1)}}{(2a)^{n+1}}(e^{2at}-1)^{n+1}\,,$$

which complete the proof of Equation 11.

To derive the joint moments $\mathbb{E}\left[x^{2n}(t)x^{2m}(s)\right]$ for $s \le t$ and all nonegative integers m, n, we simply decompose $x(t) = x(s) + \sigma\int_s^t e^{a(t-u)}\mathrm{d}w(u)$. Then

$$\begin{aligned}\mathbb{E}\left[x^{2n}(t)x^{2m}(s)\right] &= \mathbb{E}\left[\left(x(s)+\sigma\int_s^t e^{a(t-u)}\mathrm{d}w(u)\right)^{2n}x^{2m}(s)\right]\\ &= \sum_{k=0}^{n}\binom{2n}{2k}\mathbb{E}\left[x^{2(m+k)}(s)\right]\mathbb{E}\left[\left(\sigma\int_s^t e^{a(t-u)}\mathrm{d}w(u)\right)^{2(n-k)}\right].\end{aligned}$$

Both the terms $\mathbb{E}\left[x^{2(m+k)}(s)\right]$ and $\mathbb{E}\left[\left(\sigma\int_s^t e^{a(t-u)}\mathrm{d}w(u)\right)^{2(n-k)}\right]$ can be computed in exactly the same way as the computation of Equation 11.

Similarly, the general form $\mathbb{E}\left[x_1^{2n_1}(t_1)x_2^{2n_2}(t_2)\cdots x_k^{2n_k}(t_k)\right]$ can be computed in the same manner by first sorting $t_1, \cdots, t_k$. □

To characterize the MSE in Equation 5, let's start by analyzing one update, i.e., when $t = 1$. The result is in the following lemma:

Lemma A.2. *(One-step MSE) The MSE after one-step update ($t = 1$) is:*

$$\begin{aligned}\mathrm{MSE}_1 = \frac{\sigma^4}{(\ln\gamma)^2}\Bigg(&(1+2\alpha\mathcal{I}_2+\alpha^2\mathcal{I}_4)\theta_0^2 + \left(\frac{1}{\ln\gamma+2a}\right)^2\\ &+2\alpha(\mathcal{I}_1+\alpha\mathcal{I}_5)\theta_0+\alpha^2\mathcal{I}_3 - 2\frac{\theta_0+\alpha(\mathcal{I}_1+\mathcal{I}_2\theta_0)}{\ln\gamma+2a}\Bigg).\end{aligned}$$

where the quantities $\mathcal{I}_1, \cdots, \mathcal{I}_5$ are auxiliary expectation terms introduced in the proof.

Proof. Since $\theta_1 = \theta_0 + \alpha\bar{g}(\theta_0)$, we have $\mathbb{E}[\hat{\theta}] = \mathbb{E}[\theta_1] = \theta_0 + \alpha\mathbb{E}[\bar{g}(\theta_0)]$.

We can rewrite Equation 4 as

$$\mathbb{E}[\bar{g}(\theta_0)] = \mathbb{E}\left[\overline{\phi r} + \overline{\phi(\gamma^h\phi'-\phi)}\theta_0\right] \tag{15}$$

$$= \underbrace{\mathbb{E}\left[\overline{\phi r}\right]}_{\mathcal{I}_1} + \underbrace{\mathbb{E}\left[\overline{\phi(\gamma^h\phi'-\phi)}\right]}_{\mathcal{I}_2}\theta_0 \tag{16}$$

We can write out the two terms in Equation (16). For the first term,

$$
\begin{aligned}
\mathcal{I}_1 = \mathbb{E}\left[\,\overline{\phi r}\,\right] &= \mathbb{E}\left[\frac{1}{M}\sum_{i=1}^{M}\frac{1}{N-1}\sum_{k=0}^{N-2}\phi(x_i(kh))r_i(kh)\right] \\
&= \frac{1}{M}\sum_{i=1}^{M}\frac{1}{N-1}\sum_{k=0}^{N-2}\mathbb{E}\left[\phi(x_i(kh))r_i(kh)\right] \\
&= \frac{1}{N-1}\sum_{k=0}^{N-2}\mathbb{E}\left[\phi(x(kh))r(kh)\right] \qquad (17)\\
&= -\frac{h}{N-1}\sum_{k=0}^{N-2}\left\{\mathbb{E}\left[x^4(kh)\right] - \frac{\sigma^2}{\ln\gamma}\mathbb{E}\left[x^2(kh)\right]\right\} \\
&= -\frac{h\sigma^4}{2a(N-1)}\sum_{k=0}^{N-2}\left\{\left[\frac{3}{2a}(e^{2akh}-1)^2\right] - \left[\frac{1}{\ln\gamma}(e^{2akh}-1)\right]\right\}
\end{aligned}
$$

By taking the summation, $\mathcal{I}_1$ can be written as

$$
\mathcal{I}_1 = -\frac{h\sigma^4}{2a(N-1)}\left[(N-1)\left(\frac{3}{2a}+\frac{1}{\ln\gamma}\right) - \left(\frac{3}{a}+\frac{1}{\ln\gamma}\right)\frac{1-e^{2a(T-h)}}{1-e^{2ah}} + \frac{3}{2a}\frac{1-e^{4a(T-h)}}{1-e^{4ah}}\right] \tag{18}
$$

For the second term $\mathcal{I}_2$,

$$
\begin{aligned}
\mathcal{I}_2 &= \mathbb{E}\left[\,\overline{\phi(\gamma^h\phi'-\phi)}\,\right] \\
&= \mathbb{E}\left[\frac{1}{M}\sum_{i=1}^{M}\frac{1}{N-1}\sum_{k=0}^{N-2}\phi(x_i(kh))(\gamma^h\phi(x_i(kh+h))-\phi(x_i(kh)))\right] \qquad (19)\\
&= \frac{1}{M}\sum_{i=1}^{M}\frac{1}{N-1}\sum_{k=0}^{N-2}\mathbb{E}\left[\phi(x_i(kh))(\gamma^h\phi(x_i(kh+h))-\phi(x_i(kh)))\right] \qquad (20)\\
&= \frac{1}{N-1}\sum_{k=0}^{N-2}\mathbb{E}\left[\phi(x(kh))(\gamma^h\phi(x(kh+h))-\phi(x(kh)))\right] \qquad (21)\\
&= \frac{1}{N-1}\sum_{k=0}^{N-2}\left\{\gamma^h\mathbb{E}\left[x^2(kh)x^2(kh+h)\right] + \frac{\sigma^4}{(\ln\gamma)^2}(\gamma^h-1) - \mathbb{E}\left[x^4(kh)\right]\right.\\
&\qquad\left. + \frac{\sigma^2}{\ln\gamma}\left((2-\gamma^h)\mathbb{E}\left[x^2(kh)\right] - \gamma^h\mathbb{E}\left[x^2(kh+h)\right]\right)\right\} \qquad (22)
\end{aligned}
$$

Substituting the moments from Lemma A.1, we have

$$
\begin{aligned}
\mathcal{I}_2 = \frac{1}{N-1}\sum_{k=0}^{N-2}&\left\{\gamma^h\frac{\sigma^4}{4a^2}(e^{2akh}-1)e^{2ah}\left[(1-e^{-2ah})+3(e^{2akh}-1)\right]\right.\\
&+\frac{\sigma^4}{(\ln\gamma)^2}(\gamma^h-1) - \frac{3\sigma^4}{4a^2}(e^{2akh}-1)^2\\
&\left.+\frac{\sigma^4}{2a\ln\gamma}\left((2-\gamma^h)(e^{2akh}-1)-\gamma^h(e^{2a(k+1)h}-1)\right)\right\},
\end{aligned} \tag{23}
$$

which can be computed and simplified symbolically using *Mathematica*.

On the other hand, $\mathbb{E}\left[\theta_1^2\right] = \theta_0^2 + 2\theta_0\alpha\mathbb{E}\left[\bar{g}(\theta_0)\right] + \alpha^2\mathbb{E}\left[\bar{g}^2(\theta_0)\right]$, we need to compute $\mathbb{E}\left[\bar{g}^2(\theta_0)\right]$ to find the MSE. By definition,

$$
\mathbb{E}\left[\bar{g}^2(\theta_0)\right] = \underbrace{\mathbb{E}\left[(\overline{\phi r})^2\right]}_{\mathcal{I}_3} + \theta_0^2\underbrace{\mathbb{E}\left[(\overline{\phi(\gamma^h\phi'-\phi)})^2\right]}_{\mathcal{I}_4} + 2\theta_0\underbrace{\mathbb{E}\left[(\overline{\phi r})(\overline{\phi(\gamma^h\phi'-\phi)})\right]}_{\mathcal{I}_5}, \tag{24}
$$

where each term can be computed as follows.

$$
\begin{aligned}
\mathcal{I}_3 &= \mathbb{E}\left[(\overline{\phi r})^2 \right] \\
&= \mathbb{E}\left[\frac{1}{(B-M)^2} \left(\sum_{i=1}^{M} \sum_{k=0}^{N-2} [\phi(x_i(kh)) r_i(kh)] \right)^2 \right] \\
&= \mathbb{E}\left[\frac{h^2}{(B-M)^2} \left(\sum_{i=1}^{M} \sum_{k=0}^{N-2} \left[\left(x_i^2(kh) - \frac{\sigma^2}{\ln\gamma} \right) x_i^2(kh) \right] \right)^2 \right] \\
&= \frac{h^2 N}{B(N-1)^2} \Bigg[\sum_{k=0}^{N-2} \sum_{\ell=0}^{N-2} \mathbb{E}\left[\left(x^2(kh) - \frac{\sigma^2}{\ln\gamma} \right) x^2(kh) \left(x^2(\ell h) - \frac{\sigma^2}{\ln\gamma} \right) x^2(\ell h) \right] \\
&\qquad + (M-1) \left(\sum_{k=0}^{N-2} \mathbb{E}\left[\left(x^2(kh) - \frac{\sigma^2}{\ln\gamma} \right) x^2(kh) \right] \right)^2 \Bigg]
\end{aligned} \tag{25}
$$

where each expectation can be computed from Lemma A.1 and summation can be computed symbolically from *Mathematica*.

Similarly,

$$
\begin{aligned}
\mathcal{I}_4 &= \mathbb{E}\left[\left(\frac{1}{M} \sum_{i=1}^{M} \frac{1}{N-1} \sum_{k=0}^{N-2} \phi(x_i(kh))(\gamma^h \phi(x_i(kh+h)) - \phi(x_i(kh))) \right)^2 \right] \\
&= \frac{N}{B(N-1)^2} \Bigg[\sum_{k=0}^{N-2} \sum_{\ell=0}^{N-2} \mathbb{E}[\phi(x(kh))(\gamma^h \phi(x(kh+h)) - \phi(x(kh))) \\
&\qquad \phi(x(\ell h))(\gamma^h \phi(x(\ell h+h)) - \phi(x(\ell h)))] \Bigg] + \\
&\quad \frac{(M-1)}{(N-1)^2} \Bigg[\sum_{k=0}^{N-2} \sum_{\ell=0}^{N-2} \mathbb{E}\left[\phi(x(kh))(\gamma^h \phi(x(kh+h)) - \phi(x(kh))) \right] \\
&\qquad \mathbb{E}\left[\phi(x(\ell h))(\gamma^h \phi(x(\ell h+h)) - \phi(x(\ell h))) \right] \Bigg],
\end{aligned} \tag{26}
$$

and

$$
\begin{aligned}
\mathcal{I}_5 = \mathbb{E}\Bigg[& \left(\frac{1}{M} \sum_{i=1}^{M} \frac{1}{N-1} \sum_{k=0}^{N-2} [\phi(x_i(kh)) r_i(kh)] \right) \\
& \left(\frac{1}{M} \sum_{i=1}^{M} \frac{1}{N-1} \sum_{k=0}^{N-2} \phi(x_i(kh))(\gamma^h \phi(x_i(kh+h)) - \phi(x_i(kh))) \right) \Bigg] \\
= & \frac{N}{B(N-1)^2} \left[\sum_{k=0}^{N-2} \sum_{\ell=0}^{N-2} \mathbb{E}[\phi(x(kh)) r(kh) \phi(x(\ell h))(\gamma^h \phi(x(\ell h+h)) - \phi(x(\ell h)))] \right] + \\
& \frac{(M-1)}{(N-1)^2} \left[\sum_{k=0}^{N-2} \sum_{\ell=0}^{N-2} \mathbb{E}\left[\phi(x(kh)) r(kh)\right] \mathbb{E}\left[\phi(x(\ell h))(\gamma^h \phi(x(\ell h+h)) - \phi(x(\ell h))) \right] \right].
\end{aligned} \tag{27}
$$

By employing Lemma A.1 and the computation from *Mathematica*, one can derive the MSE after one-step update as

$$\begin{aligned}\mathrm{MSE}_1 = \frac{\sigma^4}{(\ln\gamma)^2}\Bigg(&(2\alpha+1)\theta_0^2 + 2\theta_0\alpha^2(\mathcal{I}_1+\mathcal{I}_2\theta_0) + \alpha^2(\mathcal{I}_3+\mathcal{I}_4\theta_0^2+2\theta_0\mathcal{I}_5)\\ &- 2\frac{\theta_0+\alpha(\mathcal{I}_1+\mathcal{I}_2\theta_0)}{\ln\gamma+2a} + \left(\frac{1}{\ln\gamma+2a}\right)^2\Bigg).\\ = \frac{\sigma^4}{(\ln\gamma)^2}\Bigg(&(1+2\alpha\mathcal{I}_2+\alpha^2\mathcal{I}_4)\theta_0^2 + 2\alpha(\mathcal{I}_1+\alpha\mathcal{I}_5)\theta_0 + \alpha^2\mathcal{I}_3\\ &- 2\frac{\theta_0+\alpha(\mathcal{I}_1+\mathcal{I}_2\theta_0)}{\ln\gamma+2a} + \left(\frac{1}{\ln\gamma+2a}\right)^2\Bigg).\end{aligned}$$

□

We now generalize the one-step update to the case of t-step updates. Recall that the update rule (Equation 3) is $\theta_{t+1} = \theta_t + \alpha\bar{g}(\theta_t)$. To derive the MSE after t-steps update, we first express θ_t in terms of the initial parameter θ_0 by repeated substitution. This recursive expansion is provided by the following lemma.

Lemma A.3 (Expansion of θ_t)**.** *The parameter θ_t after t updates satisfies:*

$$\theta_t = \theta_0 + \sum_{\ell=1}^{t}\binom{t}{\ell}\alpha^\ell\left[\overline{\phi(\gamma^h\phi'-\phi)}\right]^{\ell-1}\left[\overline{\phi r} + \overline{\phi(\gamma^h\phi'-\phi)}\theta_0\right]. \tag{28}$$

Proof. By induction, if Equation 28 is true, then

$$\begin{aligned}\theta_{t+1} &= \theta_t + \alpha\left(\overline{\phi r} + \overline{\phi(\gamma^h\phi'-\phi)}\theta_t\right)\\ &= \theta_0 + \sum_{\ell=1}^{t}\binom{t}{\ell}\alpha^\ell\left[\overline{\phi(\gamma^h\phi'-\phi)}\right]^{\ell-1}\left[\overline{\phi r} + \overline{\phi(\gamma^h\phi'-\phi)}\theta_0\right]\\ &\quad + \alpha\left[\overline{\phi r} + \overline{\phi(\gamma^h\phi'-\phi)}\theta_0\right]\\ &\quad + \sum_{\ell=1}^{t}\binom{t}{\ell}\alpha^{\ell+1}\left[\overline{\phi(\gamma^h\phi'-\phi)}\right]^{\ell}\left[\overline{\phi r} + \overline{\phi(\gamma^h\phi'-\phi)}\theta_0\right].\end{aligned}$$

Rearrange the indices for the two summations,

$$\begin{aligned}\theta_{t+1} &= \theta_0 + (t+1)\alpha\left[\overline{\phi r} + \overline{\phi(\gamma^h\phi'-\phi)}\theta_0\right]\\ &\quad + \sum_{\ell=2}^{t}\binom{t}{\ell}\alpha^\ell\left[\overline{\phi(\gamma^h\phi'-\phi)}\right]^{\ell-1}\left[\overline{\phi r} + \overline{\phi(\gamma^h\phi'-\phi)}\theta_0\right]\\ &\quad + \sum_{\ell=2}^{t}\binom{t}{\ell-1}\alpha^\ell\left[\overline{\phi(\gamma^h\phi'-\phi)}\right]^{\ell-1}\left[\overline{\phi r} + \overline{\phi(\gamma^h\phi'-\phi)}\theta_0\right]\\ &= \theta_0 + \sum_{\ell=1}^{t+1}\binom{t+1}{\ell}\alpha^\ell\left[\overline{\phi(\gamma^h\phi'-\phi)}\right]^{\ell-1}\left[\overline{\phi r} + \overline{\phi(\gamma^h\phi'-\phi)}\theta_0\right].\end{aligned}$$

where the last equation is a direct application of Pascal's formula. This completes the proof. □

By straightforward manipulation, the formula can be rewritten as

$$\theta_t = [1+\alpha(\overline{\phi(\gamma^h\phi'-\phi)})]^t\theta_0 + \alpha\overline{\phi r}\sum_{j=0}^{t-1}[1+\alpha(\overline{\phi(\gamma^h\phi'-\phi)})].$$

Recall from Equation 5 that the MSE after t-step update is

$$\text{MSE}_t = \frac{\sigma^4}{(\ln\gamma)^2}\left(\mathbb{E}[\theta_t^2] - 2\frac{1}{\ln\gamma + 2a}\mathbb{E}[\theta_t] + \left(\frac{1}{\ln\gamma + 2a}\right)^2\right), \tag{29}$$

where

$$\theta_t = \theta_0 + \sum_{\ell=1}^{t}\binom{t}{\ell}\alpha^\ell\left[\overline{\phi(\gamma^h\phi' - \phi)}\right]^{\ell-1}\left[\overline{\phi r} + \overline{\phi(\gamma^h\phi' - \phi)}\theta_0\right]. \tag{30}$$

by Lemma A.3. The exact computation in this case cannot be obtained as we did in the one-step case, since $\mathbb{E}\left[\theta_t^2\right]$ and $\mathbb{E}\left[\theta_t\right]$ cannot be easily derived. We shall consider these quantities using approximation in order to give a clear insight of MSE_t.

Lemma A.4 (Order of Higher Moments in Gradient Terms)**.** *For small $h \in (0,1)$, the following hold for all integers $n \geq 3$:*

$$\mathbb{E}\left[\left(\overline{\phi(\gamma^h\phi' - \phi)}\right)^n\right] = \mathcal{O}(h^n),$$
$$\mathbb{E}\left[\left(\overline{\phi(\gamma^h\phi' - \phi)}\right)^{n-1}\overline{\phi r}\right] = \mathcal{O}(h^n),$$
$$\mathbb{E}\left[\left(\overline{\phi(\gamma^h\phi' - \phi)}\right)^{n-2}\left(\overline{\phi r}\right)^2\right] = \mathcal{O}(h^n),$$

Proof. We begin with the following expansion from applying Taylor expansion on γ^h:

$$\begin{aligned}&\gamma^h\phi(x(kh+h)) - \phi(x(kh))\\ =\ &[\phi(x(kh+h)) - \phi(x(kh))] + h\ln\gamma\phi(x(kh+h)) + h^2(\ln\gamma)^2\phi(x(kh+h)) + \mathcal{O}(h^3)\end{aligned}$$

A direct expansion of $\left(\gamma^h\phi(x(kh+h)) - \phi(x(kh))\right)^n$ will be the summation of

$$[\phi(x(kh+h)) - \phi(x(kh))]^{n_1}\left(h\ln\gamma\phi(x(kh+h))\right)^{n_2}\left(h^2(\ln\gamma)^2\phi(x(kh+h))\right)^{n_3} + \mathcal{O}(h^3),$$

with $n_1 + n_2 + n_3 = n$. Noticing that $[\phi(x(kh+h)) - \phi(x(kh))]$ will contribute one factor h, therefore the leading order of $\mathbb{E}\left[\left(\overline{\phi(\gamma^h\phi' - \phi)}\right)^n\right] = \mathcal{O}(h^n)$ will have the power of h at least $n_1 + n_2 + n_3 = n$. The second and the third equations can be shown similarly by noticing $\overline{\phi r}$ has a factor h. ☐

This lemma shows that the high-order moments of the TD update components vanish rapidly with small h. As a result, when analyzing the MSE, truncating at $\mathcal{O}(h^3)$ suffices for a valid approximation.

Based on Lemma A.4, we immediately have

$$\mathbb{E}\left[\theta_t\right] = \theta_0 + t\alpha(\mathcal{I}_1 + \mathcal{I}_2\theta_0) + \frac{t(t-1)}{2}\alpha^2\left(\mathcal{I}_5 + \mathcal{I}_4\theta_0\right) + \mathcal{O}(h^3), \tag{31}$$

$$\mathbb{E}\left[\theta_t^2\right] = t^2\alpha^2\mathcal{I}_3 + t\alpha\theta_0\left(2\mathcal{I}_1 + (3t-1)\alpha\mathcal{I}_5\right) + \theta_0^2\left(1 + 2t\alpha\mathcal{I}_2 + t(2t-1)\alpha^2\mathcal{I}_4\right) + \mathcal{O}(h^3). \tag{32}$$

Therefore, we can write the MSE after t-step update as

$$\begin{aligned}\text{MSE}_t = \frac{\sigma^4}{(\ln\gamma)^2}\Bigg[&\left(t^2\alpha^2\mathcal{I}_3 + t\alpha\theta_0\left(2\mathcal{I}_1 + (3t-1)\alpha\mathcal{I}_5\right) + \theta_0^2\left(1 + 2t\alpha\mathcal{I}_2 + t(2t-1)\alpha^2\mathcal{I}_4\right)\right)\\ &- 2\frac{1}{\ln\gamma + 2a}\left(\theta_0 + t\alpha(\mathcal{I}_1 + \mathcal{I}_2\theta_0) + \frac{t(t-1)}{2}\alpha^2\left(\mathcal{I}_5 + \mathcal{I}_4\theta_0\right)\right)\\ &+ \left(\frac{1}{\ln\gamma + 2a}\right)^2\Bigg] + \mathcal{O}(h^3)\end{aligned} \tag{33}$$

which can be solved by the computation of $\mathcal{I}_1, \cdots, \mathcal{I}_5$.

Although the exact forms of $\mathcal{I}_1, \cdots, \mathcal{I}_5$ can be obtained by symbolic computation in *Mathematica*, it is too complex to parse. To interpret the relationship between the MSE and h clearly, we shall approximate $\mathcal{I}_1, \cdots, \mathcal{I}_5$ in terms of h.

Lemma A.5 (Approximation of I_i terms)**.** *For small $h \in (0,1)$, the quantities I_i can be expanded as follows:*

$$\mathcal{I}_1 = C_{11}h + C_{12}h^2 + \mathcal{O}(h^3), \tag{34}$$

$$\mathcal{I}_2 = C_{21}h + C_{22}h^2 + \mathcal{O}(h^3), \tag{35}$$

$$\mathcal{I}_3 = \frac{C_{31}}{B}h + \left(\frac{C_{320}}{B} + C_{321}\right)h^2 + \mathcal{O}(h^3), \tag{36}$$

$$\mathcal{I}_4 = \frac{C_{41}}{B}h + \left(\frac{C_{420}}{B} + C_{421}\right)h^2 + \mathcal{O}(h^3), \tag{37}$$

$$\mathcal{I}_5 = \frac{C_{51}}{B}h + \left(\frac{C_{520}}{B} + C_{521}\right)h^2 + \mathcal{O}(h^3), \tag{38}$$

where coefficients C_{ij} and C_{ijk} depends on a, γ, σ and T, but are independent of h and the data budget B. The expressions for each coefficient are listed below.

(i)

$$C_{11} = -\frac{\sigma^4\left(4a - 4ae^{2aT} + 8a^2T + 3\ln\gamma\left(3 - 4e^{2aT} + e^{4aT} + 4aT\right)\right)}{16a^3T\ln\gamma}, \tag{39}$$

$$C_{21} = -(2a + \ln\gamma)C_{11}, \tag{40}$$

$$C_{321} = C_{11}^2, \tag{41}$$

$$C_{421} = (2a + \ln\gamma)^2C_{11}^2, \tag{42}$$

$$C_{521} = -(2a + \ln\gamma)C_{11}^2 \tag{43}$$

(ii)

$$\begin{aligned} C_{31} = {} & \frac{\sigma^8\left(5 - e^{4aT} + 4aT - 4e^{2aT} + 8aTe^{2aT}\right)}{8a^4T(\ln\gamma)^2} + \\ & \frac{3\sigma^8\left(11 + e^{6aT} + 8aT - 9e^{2aT} + 20aTe^{2aT} - 3e^{4aT} - 4aTe^{4aT}\right)}{8a^5T\ln\gamma} + \\ & \frac{\sigma^8\left(61 + 29e^{6aT} - 3e^{8aT} + 42aT - 18e^{4aT} - 90aTe^{4aT} - 69e^{2aT} + 108aTe^{2aT}\right)}{8a^6T}, \end{aligned} \tag{44}$$

$$C_{41} = (\ln\gamma + 2a)^2C_{31}, \tag{45}$$

$$C_{51} = -(\ln\gamma + 2a)C_{31}. \tag{46}$$

(iii)

$$
\begin{aligned}
C_{320} =& \frac{\sigma^8\left(4e^{2aT}(1-2aT+3a^2T^2)+e^{4aT}(3aT-1)-aT-5\right)}{4a^4T^2(\ln\gamma)^2} \\
&+ \frac{3\sigma^8\left(e^{6aT}(9aT-2)+2e^{4aT}(3-8aT-12a^2T^2)\right)}{8a^5T^2\ln\gamma} \\
&+ \frac{3\sigma^8\left(e^{2aT}(18-37aT+60a^2T^2)-4aT-22\right)}{8a^5T^2\ln\gamma} \\
&+ \frac{\sigma^8\left(6e^{8aT}(1-6aT)+29e^{6aT}(-2+9aT)+9e^{4aT}(4-7aT-60a^2T^2)\right)}{8a^6T^2} \\
&+ \frac{\sigma^8\left(3e^{2aT}(46-87aT+108a^2T^2)-21aT-122\right)}{8a^6T^2}, \qquad (47) \\
C_{420} =& (2a+\ln\gamma)^2 C_{320}^2, \qquad (48) \\
C_{520} =& -(2a+\ln\gamma) C_{320}^2. \qquad (49)
\end{aligned}
$$

(iv)

$$
\begin{aligned}
C_{12} =& -\frac{\sigma^4\left(1-e^{2aT}-aT+3aTe^{2aT}\right)}{4a^2T^2\ln\gamma} \\
&- \frac{3\sigma^4\left(3-4e^{2aT}+e^{4aT}-2aT+12aTe^{2aT}-6aTe^{4aT}\right)}{16a^3T^2}, \qquad (50) \\
C_{22} =& -(2a+\ln\gamma)C_{12}+C_{23}, \qquad (51) \\
C_{23} =& \frac{3\sigma^4(2a+\ln\gamma)^2(3-4e^{2aT}+e^{4aT}+4aT)}{32a^3T}. \qquad (52)
\end{aligned}
$$

Proof. By relying on the previous computations of $\mathcal{I}_1, \cdots, \mathcal{I}_5$, we can derive the approximations using Taylor expansion at h.

For $\mathcal{I}_1$, we take the expansion of Equation 18 with the relationship that $T = Nh$ to obtain the following directly.

$$
\begin{aligned}
C_{11} =& -\frac{\sigma^4}{2a}\left\{\left(\frac{3}{2a}+\frac{1}{\ln\gamma}\right)+\left(\frac{3}{a}+\frac{1}{\ln\gamma}\right)\frac{(1-e^{2aT})}{2aT}+\frac{3(e^{4aT}-1)}{8a^2T}\right\} \qquad (53) \\
=& -\frac{\sigma^4\left(4a-4ae^{2aT}+8a^2T+3\ln\gamma\left(3-4e^{2aT}+e^{4aT}+4aT\right)\right)}{16a^3T\ln\gamma}, \qquad (54) \\
C_{12} =& -\frac{\sigma^4}{2a}\left\{\left(\frac{3}{a}+\frac{1}{\ln\gamma}\right)\frac{\left(1-e^{2aT}-aT+3ae^{2aT}T\right)}{2aT^2}-\frac{3\left(1-e^{4aT}-2aT+6ae^{4aT}\right)}{8a^2T^2}\right\} \qquad (55) \\
=& -\frac{\sigma^4\left(1-e^{2aT}-aT+3aTe^{2aT}\right)}{4a^2T^2\ln\gamma} \\
&- \frac{3\sigma^4\left(3-4e^{2aT}+e^{4aT}-2aT+12aTe^{2aT}-6aTe^{4aT}\right)}{16a^3T^2}. \qquad (56)
\end{aligned}
$$

For $\mathcal{I}_2$, we first use the expansion that $\gamma^h = 1 + h\ln\gamma + h^2(\ln\gamma)^2 + \mathcal{O}(h^3)$ to rewrite $\mathcal{I}_2$ as

$$
\begin{aligned}
\mathcal{I}_2 =& \mathbb{E}\left[\overline{\phi(\gamma^h\phi'-\phi)}\right] \\
=& \mathbb{E}\left[\overline{\phi(\phi'-\phi)}\right] + \mathbb{E}\left[\overline{\phi\phi'}\right]\left(h\ln\gamma + h^2(\ln\gamma)^2 + \mathcal{O}(h^3)\right), \qquad (57)
\end{aligned}
$$

the terms of each can be derived as follows.

$$
\begin{aligned}
&\mathbb{E}\left[\overline{\phi(\phi'-\phi)}\right] \\
&= \frac{1}{N-1}\sum_{k=0}^{N-1}\left(\mathbb{E}\left[x^2(kh)x^2(kh+h)-x^4(kh)\right]-\frac{\sigma^2}{\ln\gamma}\mathbb{E}\left[x^2(kh+h)-x^2(kh)\right]\right) \\
&= \frac{1}{N-1}\sum_{k=0}^{N-1}\frac{\sigma^4(e^{2ah}-1)}{4a^2}\left(3(e^{2akh}-1)^2+(e^{2akh}-1)\right) \\
&\quad -\frac{1}{N-1}\sum_{k=0}^{N-1}\frac{\sigma^4}{2a\ln\gamma}e^{2akh}(e^{2ah}-1) \\
&= \frac{\sigma^4}{4a^2}\Bigg\{\frac{7-10e^{2aT}+3e^{4aT}+8aT}{2T}h+ \\
&\qquad \frac{7-10e^{2aT}+3e^{4aT}+3aT+20aTe^{2aT}-15ae^{4aT}T+8a^2T^2}{2T^2}h^2\Bigg\} \\
&\quad -\frac{\sigma^4}{2a\ln\gamma}\left(\frac{e^{2aT}-1}{T}h+\frac{e^{2aT}-1-2aTe^{2aT}}{T^2}h^2\right)+\mathcal{O}(h^3).
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\left[\overline{\phi\phi'}\right] = \sigma^4\left\{\frac{1}{(\ln\gamma)^2}+\frac{1-e^{2aT}+2aT}{2a^2T\ln\gamma}+\frac{3\left(3-4e^{2aT}+e^{4aT}+4aT\right)}{16a^3T}+\frac{1-e^{2aT}+2aT}{2a^2T^2\ln\gamma}h\right\} \\
&+\sigma^4\frac{9-12e^{2aT}+3e^{4aT}+8aT+16aTe^{2aT}-12aTe^{4aT}+16a^2T^2}{16a^3T^2}h+\mathcal{O}(h^2).
\end{aligned}
$$

We therefore have the result for $\mathcal{I}_2$ relying on previous computation through coefficients

$$
\begin{aligned}
C_{21} &= \frac{\sigma^4(2a+\ln\gamma)\left(4a-4ae^{2aT}+8a^2T+3\ln\gamma\left(3-4e^{2aT}+e^{4aT}+4aT\right)\right)}{16a^3T\ln\gamma}, \\
C_{22} &= \frac{\sigma^4}{4a^2}\left(\frac{7-10e^{2aT}+3e^{4aT}+3aT+20aTe^{2aT}-15ae^{4aT}T+8a^2T^2}{2T^2}\right) \\
&\quad -\frac{\sigma^4}{2a\ln\gamma}\left(\frac{e^{2aT}-1-2aTe^{2aT}}{T^2}\right)+\sigma^4\frac{1-e^{2aT}+2aT}{2a^2T^2} \\
&\quad +\sigma^4\ln\gamma\frac{9-12e^{2aT}+3e^{4aT}+8aT+16aTe^{2aT}-12aTe^{4aT}+16a^2T^2}{16a^3T^2} \\
&\quad +\sigma^4\left\{1+\ln\gamma\frac{1-e^{2aT}+2aT}{2a^2T}+\frac{3(\ln\gamma)^2\left(3-4e^{2aT}+e^{4aT}+4aT\right)}{16a^3T}\right\} \\
&\quad = -(2a+\ln\gamma)C_{12}+C_{23}.
\end{aligned}
$$

For $\mathcal{I}_3$, the expectations in Equation 25 can be computed from Lemma A.1. Based on earlier similar computation, we can derive $\mathcal{I}_3$ as follows:

$$
\begin{aligned}
\mathcal{I}_3 &= \frac{h^2N}{B(N-1)^2}\Bigg[\sum_{k=0}^{N-2}\sum_{\ell=0}^{N-2}\mathbb{E}\left[\left(x^2(kh)-\frac{\sigma^2}{\ln\gamma}\right)x^2(kh)\left(x^2(\ell h)-\frac{\sigma^2}{\ln\gamma}\right)x^2(\ell h)\right] \\
&\qquad +(M-1)\left(\sum_{k=0}^{N-2}\mathbb{E}\left[\left(x^2(kh)-\frac{\sigma^2}{\ln\gamma}\right)x^2(kh)\right]\right)^2\Bigg] \\
&= \frac{C_{31}}{B}h+\left(\frac{C_{320}}{B}+C_{321}\right)h^2+\mathcal{O}(h^3),
\end{aligned}
$$

where

$$C_{31} = \frac{\sigma^8\left(5 - e^{4aT} + 4aT - 4e^{2aT} + 8aTe^{2aT}\right)}{8a^4T(\ln\gamma)^2} + \frac{3\sigma^8\left(11 + e^{6aT} + 8aT - 9e^{2aT} + 20aTe^{2aT} - 3e^{4aT} - 4aTe^{4aT}\right)}{8a^5T\ln\gamma} + \frac{\sigma^8\left(61 + 29e^{6aT} - 3e^{8aT} + 42aT - 18e^{4aT} - 90aTe^{4aT} - 69e^{2aT} + 108aTe^{2aT}\right)}{8a^6T}.$$

and

$$\begin{aligned} C_{320} = &\frac{\sigma^8\left(4e^{2aT}(1 - 2aT + 3a^2T^2) + e^{4aT}(3aT - 1) - aT - 5\right)}{4a^4T^2(\ln\gamma)^2} \\ &+ \frac{3\sigma^8\left(e^{6aT}(9aT - 2) + 2e^{4aT}(3 - 8aT - 12a^2T^2) + e^{2aT}(18 - 37aT + 60a^2T^2) - 4aT - 22\right)}{8a^5T^2\ln\gamma} \\ &+ \frac{\sigma^8\left(6e^{8aT}(1 - 6aT) + 29e^{6aT}(-2 + 9aT) + 9e^{4aT}(4 - 7aT - 60a^2T^2)\right)}{8a^6T^2} \\ &+ \frac{\sigma^8\left(3e^{2aT}(46 - 87aT + 108a^2T^2) - 21aT - 122\right)}{8a^6T^2}. \\ C_{321} = &\frac{\sigma^8\left(1 + 2aT - e^{2aT}\right)^2}{16a^4T^2(\ln\gamma)^2} + \frac{9\sigma^8\left(3 - 4e^{2aT} + e^{4aT} + 4aT\right)^2}{256a^6T^2} \\ &+ \frac{3\sigma^8\left(3 - e^{6aT} + 10aT + 8a^2T^2 + e^{4aT}(5 + 2aT) - e^{2aT}(7 + 12aT)\right)}{32a^5T^2\ln\gamma} \\ = &\frac{\sigma^8\left(4a\left(1 + 2aT - e^{2aT}\right) + 3\ln\gamma\left(3 - 4e^{2aT} + e^{4aT} + 4aT\right)\right)^2}{256a^6T^2(\ln\gamma)^2}. \end{aligned}$$

For $\mathcal{I}_4$ (Equation 26) and $\mathcal{I}_5$ (Equation 27), we use the same expansion we exploited for $\mathcal{I}_2$ to rewrite the form $\gamma^h\phi(x(kh+h)) - \phi(x(kh))$ as $[\phi(x(kh+h)) - \phi(x(kh))] + h\ln\gamma\phi(x(kh+h)) + h^2(\ln\gamma)^2\phi(x(kh+h)) + \mathcal{O}(h^3)$. Then, both the two terms can be computed similarly to what we did for $\mathcal{I}_3$. The leading term of h in $\mathcal{I}_5$ will be h since $r(kh)$ has a factor h, e.g., $C_{420} = (2a + \ln\gamma)^2 C_{320}$, $C_{520} = -(2a + \ln\gamma)C_{320}$ and

$$C_{421} = \frac{\sigma^8(2a + \ln\gamma)^2\left(4a\left(1 + 2aT - e^{2aT}\right) + 3\ln\gamma\left(3 - 4e^{2aT} + e^{4aT} + 4aT\right)\right)^2}{256a^6T^2(\ln\gamma)^2},$$

$$C_{521} = -\frac{\sigma^8(2a + \ln\gamma)\left(4a\left(1 + 2aT - e^{2aT}\right) + 3\ln\gamma\left(3 - 4e^{2aT} + e^{4aT} + 4aT\right)\right)^2}{256a^6T^2(\ln\gamma)^2}.$$

We then directly obtain $C_{421} = (2a + \ln\gamma)^2 C_{11}^2$ and $C_{521} = -(2a + \ln\gamma)C_{11}^2$. $\square$

Corollary A.6 (Sign Structure of Expansion Coefficients)**.** *Let $a < 0, 0 < \gamma < 1, \sigma > 0, T > 0$, the coefficients defined in Lemma A.5 satisfy the following sign conditions,*

$$\begin{aligned} &C_{11} < 0, C_{12} > 0, \\ &C_{21} < 0, C_{22} > 0, \\ &C_{31} < 0, C_{320} > 0, C_{321} > 0, \\ &C_{41} < 0, C_{420} > 0, C_{421} > 0, \\ &C_{51} < 0, C_{520} > 0, C_{521} > 0. \end{aligned}$$

Proof. We begin with C_{11}. Recall from Lemma A.5 that

$$C_{11} = -\frac{\sigma^4\left(4a - 4ae^{2aT} + 8a^2T + 3\ln\gamma\left(3 - 4e^{2aT} + e^{4aT} + 4aT\right)\right)}{16a^3T\ln\gamma}$$

Since $a < 0, 0 < \gamma < 1$, the denominator is positive. In the numerator, the term $4a - 4ae^{2aT} + 8a^2T = 4a(1 + 2aT - e^{2aT}) > 0$. Then let's define

$$f(y) \;=\; 3 \;-\; 4e^{2y} \;+\; e^{4y} \;+\; 4y, \quad y \le 0.$$

The derivative with respect y is:

$$f'(y) = -8e^{2y} + 4e^{4y} + 4 = 4\Big(e^{4y} - 2e^{2y} + 1\Big) = 4\big(e^{2y} - 1\big)^2 \geq 0,$$

and the equality holds only when $e^{2y} = 1$ (i.e. $y = 0$). It follows that $f'(y) > 0$ for all $y < 0$. Thus, f is strictly increasing on $(-\infty, 0)$ and $f(y) < f(0) = 0$. Substituting $y = aT$, we have $3 - 4e^{2aT} + e^{4aT} + 4aT < 0$.

Combining these facts shows that the numerator and the denominator of C_{11} are both positive, hence $C_{11} < 0$. Following a similar argument, we can show that $C_{12} > 0$.

$C_{31} < 0$ follows from that each term in the summation of C_{31} is negative, e.g., the nominator of the first term in the summation of C_{31} is $5 - e^{4aT} + 4aT - 4e^{2aT} + 8aTe^{2aT}$, which is strictly negative when $aT < 0$. Similarly, one can check the sign of each term in the summations of C_{320} to show that $C_{320} > 0$. $C_{41} < 0$ and $C_{51} < 0$ follows immediately from the sign of C_{31}.

A straightforward term-by-term check yields the sign of the remaining coefficients: $C_{321} > 0, C_{420} > 0, C_{421} > 0, C_{520} > 0, C_{521} > 0$. This completes the proof.

□

Now we turn to the proof of our main result, Theorem 4.1.

Proof. We shall show

$$\mathrm{MSE}_t = C_0 + C_1 h + C_2 h^2 + \mathcal{O}(h^3) \tag{58}$$

where $C_0 \geq 0$, $C_1 \leq 0$, $C_2 \geq 0$ are constants as follows. First, we take C_0 to be the constants with respect to h, that is

$$C_0 = \frac{\sigma^4}{(\ln\gamma)^2}\left(\theta_0 - \frac{1}{\ln\gamma + 2a}\right)^2.$$

It is trivial that $C_0 \geq 0$ and $C_0 = 0$ if and only if when $\theta_0 = \theta^* = \frac{1}{\ln\gamma + 2a}$. Furthermore, C_0 is quadratic in terms of θ_0.

Then, we collect all the coefficients of h and obtain

$$\begin{aligned} C_1 = {} & \frac{2t\alpha\sigma^4 C_{11}}{(\ln\gamma)^2}\left(\theta_0 - \frac{1}{\ln\gamma + 2a}\right) + \frac{2t\alpha\theta_0\sigma^4 C_{21}}{(\ln\gamma)^2}\left(\theta_0 - \frac{1}{\ln\gamma + 2a}\right) \\ & + \frac{t\alpha\sigma^4}{B(\ln\gamma)^2}\left(t\alpha(C_{31} + 2C_{51}\theta_0 + \theta_0^2 C_{41}) + (t-1)\alpha(C_{51} + \theta_0 C_{41})\left(\theta_0 - \frac{1}{\ln\gamma + 2a}\right)\right). \end{aligned}$$

By applying the analysis in Lemma A.5, we rewrite C_1 as follows.

$$\begin{aligned} C_1 = {} & \frac{2t\alpha\sigma^8\,(2a + \ln\gamma)}{(\ln\gamma)^2}\left(\theta_0 - \frac{1}{\ln\gamma + 2a}\right)^2 \\ & \frac{\left(4a(1 - e^{2aT} + 2aT) + 3\ln\gamma\left(3 - 4e^{2aT} + e^{4aT} + 4aT\right)\right)}{16a^3 T\ln\gamma} \\ & + \frac{t\alpha\sigma^4}{B(\ln\gamma)^2}\alpha(2t-1)\,(2a + \ln\gamma)^2\left(\theta_0 - \frac{1}{\ln\gamma + 2a}\right)^2 C_{31} \\ = {} & -\frac{2t\alpha\sigma^4\,(2a + \ln\gamma)}{(\ln\gamma)^2}\left(\theta_0 - \frac{1}{\ln\gamma + 2a}\right)^2 C_{11} \\ & + \frac{t\alpha\sigma^4}{B(\ln\gamma)^2}\alpha(2t-1)\,(2a + \ln\gamma)^2\left(\theta_0 - \frac{1}{\ln\gamma + 2a}\right)^2 C_{31} \\ = {} & \frac{t\alpha\sigma^4}{(\ln\gamma)^2}\left(\theta_0 - \frac{1}{\ln\gamma + 2a}\right)^2\left[-2\,(2a + \ln\gamma)\,C_{11} + \frac{\alpha(2t-1)\,(2a + \ln\gamma)^2\,C_{31}}{B}\right] \end{aligned}$$

Since $C_{11} < 0$ and $C_{31} < 0$, we directly have $C_1 \leq 0$. Particularly, $C_1 = 0$ if and only if when $\theta_0 = \theta^* = \frac{1}{\ln\gamma + 2a}$.

For C_2, we compute it similarly to what we did for C_1 as follows.

$$C_2 = \frac{t\alpha\sigma^4}{(\ln\gamma)^2}\Bigg[2\left(\theta_0 - \frac{1}{\ln\gamma+2a}\right)C_{12} + 2\theta_0\left(\theta_0 - \frac{1}{\ln\gamma+2a}\right)C_{22} + t\alpha C_{321}$$
$$+ \alpha\theta_0\left(\theta_0(2t-1) - \frac{t-1}{\ln\gamma+2a}\right)C_{421} + \alpha\left(\theta_0(3t-1) - \frac{t-1}{\ln\gamma+2a}\right)C_{521}\Bigg]$$
$$+ \frac{t\alpha\sigma^4}{B(\ln\gamma)^2}\left\{t\alpha C_{320} + \alpha\theta_0\left(\theta_0(2t-1) - \frac{t-1}{\ln\gamma+2a}\right)C_{420} + \alpha\left(\theta_0(3t-1) - \frac{t-1}{\ln\gamma+2a}\right)C_{520}\right\}.$$

Hence

$$C_2 = \frac{t\alpha\sigma^4}{(\ln\gamma)^2}\left(\theta_0 - \frac{1}{\ln\gamma+2a}\right)^2$$
$$\left[2C_{23} - 2(2a+\ln\gamma)C_{12} + (C_{11}^2 + \frac{C_{320}}{B})(2a+\ln\gamma)^2(t\alpha + (t-1)\alpha)\right],$$

where $C_{23}, C_{12}, C_{11}, C_{320}$ are given in Lemma A.5. It is straighforward to verify that $C_2 \geq 0$ since $C_{23} > 0, C_{12} > 0, C_{11}^2 > 0, C_{320} > 0$. Furthermore, $C_2 = 0$ if and only if $\theta_0 = \theta^* = \frac{1}{\ln\gamma+2a}$. $\square$

For ease of reference, we restate the following quantities.

$$C_{11} = -\frac{\sigma^4\left(4a - 4ae^{2aT} + 8a^2T + 3\ln\gamma\left(3 - 4e^{2aT} + e^{4aT} + 4aT\right)\right)}{16a^3T\ln\gamma},$$

$$C_{12} = -\frac{\sigma^4\left(1 - e^{2aT} - aT + 3aTe^{2aT}\right)}{4a^2T^2\ln\gamma}$$
$$- \frac{3\sigma^4\left(3 - 4e^{2aT} + e^{4aT} - 2aT + 12aTe^{2aT} - 6aTe^{4aT}\right)}{16a^3T^2},$$

$$C_{23} = \frac{3\sigma^4(2a+\ln\gamma)^2(3 - 4e^{2aT} + e^{4aT} + 4aT)}{32a^3T},$$

$$C_{31} = \frac{\sigma^8\left(5 - e^{4aT} + 4aT - 4e^{2aT} + 8aTe^{2aT}\right)}{8a^4T(\ln\gamma)^2}$$
$$+ \frac{3\sigma^8\left(11 + e^{6aT} + 8aT - 9e^{2aT} + 20aTe^{2aT} - 3e^{4aT} - 4aTe^{4aT}\right)}{8a^5T\ln\gamma}$$
$$+ \frac{\sigma^8\left(61 + 29e^{6aT} - 3e^{8aT} + 42aT - 18e^{4aT} - 90aTe^{4aT} - 69e^{2aT} + 108aTe^{2aT}\right)}{8a^6T},$$

$$C_{320} = \frac{\sigma^8\left(4e^{2aT}(1 - 2aT + 3a^2T^2) + e^{4aT}(3aT - 1) - aT - 5\right)}{4a^4T^2(\ln\gamma)^2}$$
$$+ \frac{3\sigma^8\left(e^{6aT}(9aT - 2) + 2e^{4aT}(3 - 8aT - 12a^2T^2) + e^{2aT}(18 - 37aT + 60a^2T^2) - 4aT - 22\right)}{8a^5T^2\ln\gamma}$$
$$+ \frac{\sigma^8\left(6e^{8aT}(1 - 6aT) + 29e^{6aT}(-2 + 9aT) + 9e^{4aT}(4 - 7aT - 60a^2T^2)\right)}{8a^6T^2}$$
$$+ \frac{\sigma^8\left(3e^{2aT}(46 - 87aT + 108a^2T^2) - 21aT - 122\right)}{8a^6T^2}$$

A.3 Proof of Corollary 4.2 and Corollary 4.3

Proof of Corollary 4.2. Based on Theorem 4.1, we can further find h^* as follows.

$$h^* = \arg\min_h \mathrm{MSE}_t$$
$$= -\frac{-2\,(2a+\ln\gamma)\,C_{11} + \frac{\alpha(2t-1)(2a+\ln\gamma)^2 C_{31}}{B}}{2\left[2C_{23} - 2(2a+\ln\gamma)C_{12} + (C_{11}^2 + \frac{C_{320}}{B})(2a+\ln\gamma)^2(2t-1)\alpha\right]},$$

which is not affected by the initial value θ_0. The corresponding minimal t-step MSE follows by substituting $h = h^*$,

$$\mathrm{MSE}_t^* \approx \frac{\sigma^4\left(\theta_0 - \frac{1}{\ln\gamma+2a}\right)^2}{(\ln\gamma)^2} \left[1 - \frac{4ta\alpha\left(-2\,(2a+\ln\gamma)\,C_{11} + \frac{\alpha(2t-1)(2a+\ln\gamma)^2 C_{31}}{B}\right)^2}{2C_{23} - 2(2a+\ln\gamma)C_{12} + (C_{11}^2 + \frac{C_{320}}{B})(2a+\ln\gamma)^2(2t-1)\alpha}\right]. \quad (59)$$

□

Proof of Corollary 4.3. Let's rewrite the t-step MSE MSE_t as follows:

$$\mathrm{MSE}_t = \Bigg\{1 + t\alpha\left[-2\,(2a+\ln\gamma)\,(C_{11}h + C_{12}h^2) + 2C_{23}h^2 + C_{11}^2h^2(2a+\ln\gamma)^2(2t-1)\alpha\right]$$
$$+ \frac{t(2t-1)\alpha^2\,(2a+\ln\gamma)^2\,(C_{31}h + C_{320}h^2)}{B}\Bigg\} \frac{\sigma^4}{(\ln\gamma)^2}\left(\theta_0 - \frac{1}{\ln\gamma+2a}\right)^2 + \mathcal{O}(h^3).$$

If the budge B is large, the approximation becomes

$$\mathrm{MSE}_t = \left\{1 + t\alpha\left[-2\,(2a+\ln\gamma)\,(C_{11}h + C_{12}h^2) + 2C_{23}h^2 + C_{11}^2h^2(2a+\ln\gamma)^2(2t-1)\alpha\right]\right\}$$
$$\frac{\sigma^4}{(\ln\gamma)^2}\left(\theta_0 - \frac{1}{\ln\gamma+2a}\right)^2 + \mathcal{O}(\frac{1}{B}) + \mathcal{O}(h^3).$$

Therefore, we obtain the optimal h from the above approximation.

$$h^* \approx -\frac{-2\,(2a+\ln\gamma)\,C_{11}}{2\left[2C_{23} - 2(2a+\ln\gamma)C_{12} + C_{11}^2(2a+\ln\gamma)^2(2t-1)\alpha\right]}.$$

If the horizon T is large (therefore, B is large), we have

$$C_{11} \to -\frac{\sigma^4(2a+3\ln\gamma)}{4a^2\ln\gamma},$$
$$C_{23} \to \frac{3\sigma^4(2a+\ln\gamma)^2}{8a^2},$$
$$C_{12} \to 0$$
$$C_{31}/B \to 0,$$
$$C_{320} \to 0.$$

The MSE becomes

$$\mathrm{MSE}_t = \frac{\sigma^4}{(\ln\gamma)^2}\left(\theta_0 - \frac{1}{\ln\gamma+2a}\right)^2 \Bigg\{1 + t\alpha\Bigg[\frac{\sigma^4(2a+\ln\gamma)(2a+3\ln\gamma)}{2a^2\ln\gamma}h$$
$$+ (2a+\ln\gamma)^2\left(\frac{3\sigma^4}{4a^2} + \frac{\sigma^8(2a+3\ln\gamma)^2(2t-1)\alpha}{16a^4(\ln\gamma)^2}\right)h^2\Bigg]\Bigg\} + \mathcal{O}(\frac{1}{T}) + \mathcal{O}(h^3).$$

In this case, we can obtain the optimal h as

$$h^* = -\frac{8a^2\ln\gamma(2a+3\ln\gamma)}{(2a+\ln\gamma)\left(12a^2(\ln\gamma)^2 + \sigma^4(2a+3\ln\gamma)^2(2t-1)\alpha\right)}.$$

□

A.4 Numerical evaluation of the analytical one-Step MSE

In Lemma A.2, we derived an exact, closed-form expression for the MSE after one-step update. The full expression, along with the *Mathematica* code for symbolic computation, are included in the supplementary material. Although this expression does not capture the behavior of the asymptotic MSE, analysis of one-step MSE offers valuable insights, as it is an exact form free from any approximation error. However, the expression is too complex to parse and difficult to interpret. Instead, we gain intuition by numerically evaluating the MSE over a range of key parameter values and examining its characteristics in this section. We fix the parameters $\sigma = 1, \alpha = 1$ throughout this section unless otherwise specified.

Behaviour of $\mathrm{MSE}_1(h)$. To begin, we visualize the MSE as a function of the discretization h in Figure 7, for two different parameter settings, both starting from $\theta_0 = -1$. The figure show that for small h, MSE behaves quadratically in h. Moreover, each curve exhibits a clear trade-off in h, highlighting the importance of choosing an appropriate discretization step for accurate value estimation in TD learning.

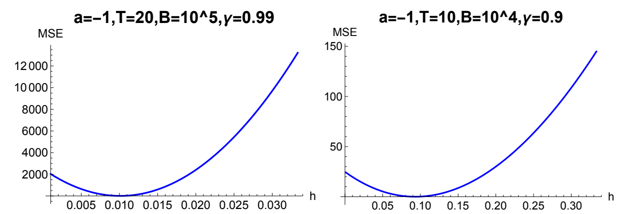

Figure 7: MSE_1 as a function of h

Dependence of h^* and MSE^* on B and T. Next, we examine how the optimal step-size h^* and the corresponding minimal MSE, $\mathrm{MSE}^* := \mathrm{MSE}_1(h^*)$, vary with the horizon T, data budget B. Fixing $a = -1, \gamma = 0.99, \alpha = 1.0, \theta_0 = 0$, we sweep T and B over a range of values. The minimizer h^* is computed by optimizing the exact MSE in *Mathematica* over the interval of h in $[T/B, T/2]$. Figure 8 shows h^* and MSE^* over $T \in [0.5, 50]$ and $B \in [1e3, 1e6]$, with $\theta_0 = 0$, where the color represents the value of h^* and MSE^*, respectively. Both h^* and MSE^* exhibit some variation across different regions of the parameter space, with MSE^* spanning a much wider range than h^*. They remain relatively consistent for the majority of values in T and B. However, for small B, the changes are more pronounced: h^* increases rapidly with T, and a similar trend is observed in MSE^*. This behavior arises because the lower bound of h is determined by T/B where we assign the entire data budget to a single trajectory. When B is small, this lower bound would be larger and dominates h^*, and MSE^* is driven up accordingly. In contrast, as B increases, h^* seems to stabilize and shows little sensitivity to T, as it is no longer constrained by the lower bound T/B.

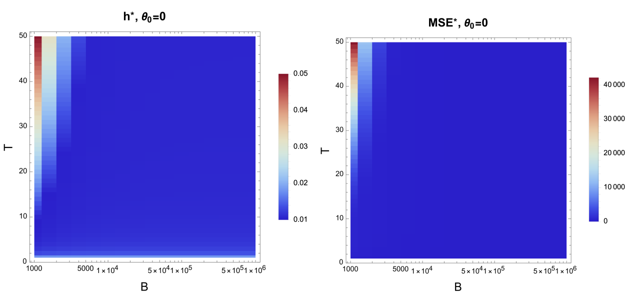

Figure 8: h^* and MSE^* of one-step TD over B and T

The phenomenon is further illustrated in Figure 9 which plots h^* as a function of T at $B = 1e3$ (left) and $B = 1e6$ (right). For $B = 1e3$, there is a transition point at $T = 9.5, h^* = 0.01$, exhibiting a noticeable shift in the behavior of the curve. At this point, h^* coincides with the lower bound of T/B and thereafter increases linearly with T. In contrast, for $B = 1e6$, h^* is mostly constant except when $T < 5$. This suggests that for short estimation horizon T, a slightly larger h (fewer samples per trajectory) can be optimal.

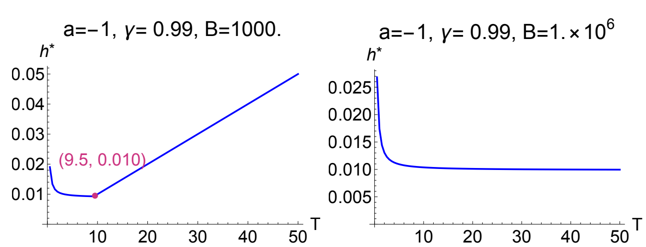

Figure 9: h^* as a function of T for different values of B

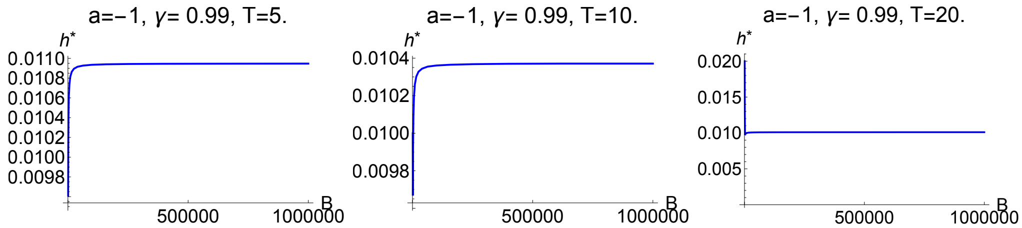

Figure 10: h^* as a function of B for different values of T

Similarly, we plot h^* as a function of B while fixing T to different values, shown in Figure 10. The results illustrates that the optimal step size h remains largely constant for large B. In contrast, in the small data regime ($B \approx 1000$), h^* is constrained by the lower bound T/B at some value of T. Once B increases enough to free h^* from that constraint, h^* decreases rapidly to the unconstrained optimum and then stabilizes around a certain value.

A.5 Computational resources

The *Mathematica* scripts were executed on a laptop with Intel i5 CPU and 8 GB of RAM. All numerical experiments in Section 5 can be run on a standard desktop equipped with 64 GB of RAM (sufficient to load the full dataset into memory). For computing MSE for various parameters, we leveraged a compute cluster to parallelize and accelerate the workload. The desktop-version code are provided in the supplementary materials; cluster-specific code is omitted to preserve anonymity.