INSTANCE-DEPENDENT CONTINUOUS-TIME REINFORCEMENT LEARNING VIA MAXIMUM LIKELIHOOD ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous-time reinforcement learning (CTRL) provides a natural framework for sequential decision-making in dynamic environments where interactions evolve continuously over time. While CTRL has shown growing empirical success, its ability to adapt to varying levels of problem difficulty remains poorly understood. In this work, we investigate the instance-dependent behavior of CTRL and introduce a simple, model-based algorithm built on maximum likelihood estimation (MLE) with a general function approximator. Unlike existing approaches that estimate system dynamics directly, our method estimates the state marginal density to guide learning. We establish instance-dependent performance guarantees by deriving a regret bound that scales with the total reward variance and measurement resolution. Notably, the regret becomes independent of the specific measurement strategy when the observation frequency adapts appropriately to the problem's complexity. To further improve performance, our algorithm incorporates a randomized measurement schedule that enhances sample efficiency without increasing measurement cost. These results highlight a new direction for designing CTRL algorithms that automatically adjust their learning behavior based on the underlying difficulty of the environment.

1 INTRODUCTION

Many real-world systems—such as autonomous robots, financial markets, and medical interventions—evolve in continuous time, where actions and feedback unfold without discrete intervals. This motivates the study of continuous-time reinforcement learning (CTRL), a framework where the agent learns to interact with a dynamic environment in real time to maximize cumulative reward. Unlike its discrete-time counterpart, CTRL is grounded in the natural temporal structure of many applications, making it particularly well-suited for control in physical and continuous systems. Recent work has highlighted its empirical potential, drawing on tools from continuous control theory (Greydanus et al., 2019; Yildiz et al., 2021; Lutter et al., 2021; Treven et al., 2024a) and the emerging use of diffusion-based models (Yoon et al., 2024; Xie et al., 2023). These developments underscore CTRL's growing relevance and its advantage in capturing fine-grained interactions that discrete-time methods often approximate only coarsely.

In this paper, we focus on the adaptivity of CTRL—that is, the ability of a learning algorithm to adjust its behavior and complexity in response to the difficulty of the problem instance. Intuitively, simpler environments should require less exploration and faster convergence, while more complex dynamics or reward structures may demand prolonged learning and finer control. For example, in robotic manipulation, navigating an open space may require significantly less precision and feedback sensitivity compared to threading a needle or interacting with deformable objects. Despite its importance, adaptivity remains largely underexplored in the CTRL literature: existing methods often lack theoretical guarantees or empirical mechanisms to modulate learning effort according to task complexity. This motivates our first core question:

Can we design a CTRL algorithm that is provably adaptive to problem difficulty, offering instance-dependent performance guarantees?

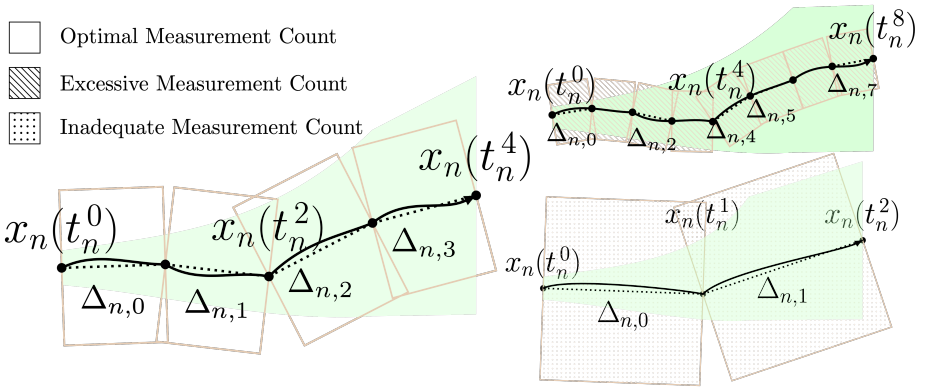

Figure 1: We depict the state trajectory $x_n(t)$ over $t \in [0, T]$ in episode n, with $x_n(0) = x_{\text{ini}}$ and $x_n(T)$ at the endpoints. Observation times t_n^k are marked by black dots. Each measurement interval $\Delta_{n,k} = t_n^{k+1} - t_n^k$ is overlaid by a brown rectangle of width $\Delta_{n,k}$ and height proportional to $\Delta_{n,k}$, so that its area encodes $\Delta_{n,k}^2$ in our regret bound. The green shading illustrates the total variance Var^{u_n}. Proper measurement gap should be selected in accordance with policy variance Var^{u_n} to achieve an optimal instance-dependent performance.

A natural starting point to investigate adaptivity in CTRL is to approximate the continuous-time process using discrete-time reinforcement learning with equidistant measurement spaced observations. This allows us to draw from a rich body of work on adaptivity in discrete-time RL, where regret bounds and learning dynamics have been thoroughly analyzed (Zhao et al., 2023; Zhou et al., 2023; Wang et al., 2024b;a). However, equidistant measurement imposes a fixed temporal structure that may be ill-suited for many continuous-time environments. For tasks characterized by unevenly evolving system dynamics—such as sudden shifts in patient vital signs or abrupt state transitions in mechanical systems—uniform sampling might either overlook important events or waste resources on redundant observations. This rigidity limits the ability of CTRL to adapt its learning to the underlying complexity of the environment. Consequently, a key question arises:

How does the choice of measurement strategy in CTRL influence its ability to adapt across problem instances?

In this work, we aim to address the two core questions outlined above. Our main contributions are summarized as follows.

- We introduce a conceptually simple model-based algorithm for CTRL, termed CT-MLE (Continuous-Time Reinforcement Learning with Maximum Likelihood Estimation). Unlike previous methods that estimate the underlying system dynamics directly (Treven et al., 2024a; Zhao et al., 2025), CT-MLE instead estimates the marginal state density using maximum likelihood estimation (MLE) with a general function approximator (e.g., neural networks or kernel models). This shift—from modeling dynamics to modeling marginal distributions—offers greater modeling flexibility and improved sample efficiency in practice. Additionally, CT-MLE is modular and compatible with a broad range of policy classes and sampling strategies, making it applicable to a wide variety of CTRL settings.
- From a theoretical perspective, we establish a regret bound for CT-MLE over the first N episodes of interaction. Specifically, we show that the regret satisfies

$$d^2 + d\sqrt{\sum_{n=1}^{N}\sum_{k=0}^{m_n-1}\Delta_{n,k}^2 + \sum_{n=1}^{N}\text{Var}^{u_n}},$$

  where d denotes the complexity of the function class used for marginal density estimation, m_n represents the number of measurements in episode n, $\Delta_{n,k}$ represents the k-th measurement gap in episode n, and Var^{u_n} quantifies the total variance of the integrated reward under policy u_n. A central insight of our analysis is that when the measurement schedule is adapted to the problem

instance—i.e., when $\sum_{k=0}^{m_n-1} \Delta_{n,k}^2$ is chosen in accordance with Var^{u_n}—the regret becomes primarily dependent on the reward variance and is nearly independent of the measurement schedule itself. This instance-dependent property highlights a key distinction from traditional discrete-time reinforcement learning, where measurements are typically uniform and agnostic to problem complexity. Figure 1 provides a demonstration of this phenomenon. Our results underscore the importance of adaptive measurement strategies for achieving instance-optimal performance in the continuous-time setting.

- A core technical innovation in CT-MLE is its Monte Carlo-type randomized measurement strategy, which augments the default measurement grid with additional observation points sampled within each interval. This randomization enables unbiased estimation of the reward integral across each measurement gap, while maintaining the total number of measurements (i.e., measurement complexity) at the same order. This design not only enhances the practical effectiveness of CT-MLE but also introduces a general technique that may be of independent interest for continuous-time decision-making problems.

Notation. We use lower case letters to denote scalars, and use lower and upper case bold face letters to denote vectors and matrices respectively. We denote by $[n]$ the set $\{1, \ldots, n\}$. For two positive functions $a(x)$ and $b(x)$ defined on a common domain, we write $a(x) \lesssim b(x)$ if there exists an absolute constant $C > 0$ such that $a(x) \leq Cb(x)$ for all x in the domain. Given a distribution $p(x)$, we use $\mathbb{E}_{x\sim p}[\cdot]$ to denote expectation and $\mathbb{V}_{x\sim p}[\cdot]$ to denote variance. For two distributions p and q, we define their squared Hellinger distance as $\mathbb{H}^2(p \,\|\, q) := 1 - \int \sqrt{p(x)q(x)}\, dx$.

2 Related Work

2.1 Continuous-time reinforcement learning

Our work resides within the paradigm of CTRL, a foundational research thread in the control community. Early studies emphasized planning in analytically tractable settings such as the linear–quadratic regulator (LQR) (Doya, 2000; Vrabie & Lewis, 2009; Faradonbeh & Faradonbeh, 2023; Caines & Levanony, 2019; Huang et al., 2024; Basei et al., 2022; Szpruch et al., 2024). A pivotal advance occurred when Chen et al. (2018) introduced neural function approximation for learning nonlinear dynamics and value functions, thereby catalysing data-driven CTRL. Building on this foundation, Yildiz et al. (2021) proposed an episodic model-based framework that alternates between fitting ODE models to collected trajectories and solving the resulting optimal-control problem with a continuous-time actor–critic. Subsequently, Holt et al. (2024) showed that under costly observations, uniform time sampling is suboptimal and that state-dependent schedules can yield higher returns. Parallel efforts (Karimi, 2023; Ni & Jang, 2022; Holt et al., 2023) have bridged continuous-time theory with practical implementations by considering deterministic systems with discrete measurements or control updates. More recent analyses have extended these ideas to deterministic and stochastic dynamics with nonlinear approximation (Treven et al., 2024a;b), and Zhao et al. (2025) further broadened the approximation class while relaxing assumptions on epistemic-uncertainty estimators. Yet the existing theory largely provides worst-case guarantees. We close this gap by establishing the first variance-aware, nearly horizon-free *second-order* regret bound for stochastic CTRL under general function approximation—measured via the eluder dimension—and show that a simple, standard MLE-based model-based algorithm attains this bound.

2.2 Variance-aware reinforcement learning

There has been a series of work studying variance-aware or horizon-free sample complexity for discrete-time reinforcement learning (Simchowitz & Jamieson, 2019; Jin et al., 2020; Dann et al., 2021; Xu et al., 2021; Wagenmaker et al., 2022; He et al., 2021a;b; Zhou et al., 2021b;a; Zhao et al., 2022; Zhou & Gu, 2022). To mention a few, early online-learning work provided second-order bounds: Cesa-Bianchi et al. (2007) derived refined regret bounds based on squared losses in expert advice, and Ito et al. (2020) established tight first- and second-order regret for adversarial linear bandits using Bernstein-type concentration. Extending to MDPs, Zanette & Brunskill (2019)'s EULER algorithm achieves regret scaling with the maximum per-step return variance rather than H, and Foster & Krishnamurthy (2021) used triangular-discrimination bonuses to obtain small-loss

bounds in contextual bandits. For structured function approximation, Kim et al. (2022) obtained horizon-free, variance-adaptive regret for linear mixture MDPs via weighted least-squares, Zhao et al. (2023) provided computationally efficient variance-dependent bounds for linear bandits and mixtures, and Zhang et al. (2021) devised variance-aware confidence sets giving logarithmic horizon dependence. Distributional RL has delivered second-order guarantees under general classes by modeling full return distributions (Zhang et al., 2022), and Huang et al. (2024) achieved sublinear regret for continuous-time stochastic LQR by estimating transition. Despite these advances, all require specialized variance or distributional machinery; our work shows that a standard MLE-based model-based RL approach attains nearly horizon-free, second-order variance-dependent bounds under general function approximation without bespoke variance estimation or distributional techniques, similar to Wang et al. (2024b) but under the continuous-time setup.

3 Problem Setup

Stochastic Differential Equation Formulation. We consider a general nonlinear continuous-time dynamical system governed by a stochastic differential equation (SDE). Let $x(\cdot)$ denote the state trajectory over a fixed planning horizon $[0, T]$, where $x(t) \in \mathcal{X} \subseteq \mathbb{R}^l$ for all $t \in [0, T]$. The system dynamics under a deterministic policy $u \in \Pi : \mathcal{X} \to \mathcal{U} \subseteq \mathbb{R}^r$ are described by

$$dx(t) = f(x(t), u(x(t)))\, dt + g(x(t), u(x(t)))\, dw(t),$$

where $w(t) \in \mathbb{R}^l$ is a standard Wiener process and the SDE is interpreted in the Itô sense. Here, $f \in \mathcal{F}$ and $g \in \mathcal{G}$, where each $f : \mathcal{X} \times \mathcal{U} \to \mathbb{R}^l$ and $g : \mathcal{X} \times \mathcal{U} \to \mathbb{R}^{l \times l}$ denote the drift and diffusion functions, respectively. Given an initial state $x(0) = x$, we denote by $p_{f,g}(u, x)$ the law of the trajectory $x(\cdot)$. We write $p_{f,g}(u, x, s)$ for the marginal distribution of $x(s)$ and use $p_{f,g}(\cdot \mid u, x, s)$ to denote its corresponding density function.

Learning Protocol. The learning process unfolds in episodes. In each episode $n = 1, \ldots, N$, the agent executes a policy u_n and observes the trajectory $x(\cdot) \sim p_{f^*,g^*}(u_n, x_{\text{ini}})$, where (f^*, g^*) denotes the unknown environment and x_{ini} is the fixed initial state. During execution, the agent selects a set of measurement times $\{t_n^k\}_{k=1}^{m_n} \subset [0, T]$ at which observations are collected. These observations are used to update the policy for the next episode. The agent's objective is to find a policy that maximizes the expected cumulative reward under the reward function $b : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$:

$$u^* = \arg\max_{u \in \Pi} R_{f^*,g^*}(u), \quad \text{where} \quad R_{f,g}(u) := V_{f,g}(u, x_{\text{ini}}, 0),$$

and the value function is given by

$$V_{f,g}(u, x, s) := \mathbb{E}_{x(\cdot) \sim p_{f,g}(u, x_{\text{ini}})} \left[\int_{t=s}^{T} b(x(t), u(x(t)))\, dt \,\middle|\, x(s) = x \right].$$

Performance Metrics. We evaluate algorithmic performance using several metrics. *The regret* is defined as

$$\text{Regret}(N) := \sum_{n=1}^{N} \left(R_{f^*,g^*}(u^*) - R_{f^*,g^*}(u_n) \right),$$

We say a policy u is ϵ-optimal if $R_{f^*,g^*}(u^*) - R_{f^*,g^*}(u) \leq \epsilon$. For any CTRL algorithm that returns an ϵ-optimal policy after N episodes, we define the *episode complexity* as N, and the *measurement complexity* as $\sum_{n=1}^{N} m_n$, where m_n denotes the number of measurements in episode n. We also consider the λ-*total complexity* for any $\lambda \in [0, 1]$, defined as the weighted sum: $(1 - \lambda)N + \lambda \sum_{n=1}^{N} m_n$. This interpolates between pure episode complexity ($\lambda = 0$) and pure measurement complexity ($\lambda = 1$).

4 CTRL with Maximum Likelihood Estimation

In this section, we introduce our algorithm, CT-MLE, as described in Algorithm 1. At a high level, each episode n follows the standard optimistic model-based approach in CTRL (Treven et al., 2024b).

Algorithm 1 Continuous-Time Reinforcement Learning with Maximum Likelihood Estimation

Require: Episode number N, policy class Π, initial state x_{ini}, drift class $\mathcal{F}$, diffusion class $\mathcal{G}$, reward function b, confidence radius β, planning horizon T.

1: For each $n \in [N]$, determine a fixed measurement time sequence $0 = t_n^0 < \cdots < t_n^{m_n} = T$. For any $0 \le k < m_n$, denote measurement gaps $\Delta_{n,k} := t_n^{k+1} - t_n^k$.

2: **for** episode $n = 1, \ldots, N$ **do**

3: Set confidence sets of (f, g) as $\mathcal{P}_n$, where

$$\mathcal{P}_n := \Bigg\{ (f,g) \in \mathcal{F} \times \mathcal{G} : \sum_{i=1}^{n-1} \sum_{k=0}^{m_i - 1} \log p_{f,g}(x_i(t_i^{k+1}) | u_i, x_i(t_i^k), \Delta_{i,k}) \geq \max_{(f',g') \in \mathcal{F} \times \mathcal{G}} \sum_{i=1}^{n-1} \sum_{k=0}^{m_i - 1} \log p_{f',g'}(x_i(t_i^{k+1}) | u_i, x_i(t_i^k), \Delta_{i,k}) - \beta \Bigg\}.$$

4: (Randomized strategy) set $\widehat{\mathcal{P}}_n$ following Algorithm 2.

5: Set policy u_n, f_n, g_n as $u_n, f_n, g_n = \arg\max_{u \in \Pi, (f,g) \in \mathcal{P}_n \cap \widehat{\mathcal{P}}_n} R_{f,g}(u)$.

6: Execute the n-th episode and observe $x_n(t_n^0), \ldots, x_n(t_n^{m_n})$.

7: (Randomized strategy) obtain additional observations to build $\widehat{\mathcal{P}}_n$ following Algorithm 2.

8: **end for**

9: **return** Randomly pick an $n \in [N]$ uniformly and output $\widehat{u}$ as u_n.

Specifically, the agent constructs a confidence set for the unknown drift f^* and diffusion g^*, and then applies the principle of optimism to select a near-optimal policy $u_n \in \Pi$. The selected policy is executed in the environment, yielding a continuous-time trajectory $x_n(\cdot)$. The agent then collects informative observations from this trajectory to refine its confidence set for the next episode. This framework parallels optimistic approaches in discrete-time RL (Abbasi-Yadkori et al., 2011; Jin et al., 2019; Russo & Van Roy, 2013; Jin et al., 2021), though applied to the continuous-time setting.

A key distinction in CTRL is that the agent must decide *when* to observe the trajectory, since data is generated in continuous time. To address this, CT-MLE introduces a sequence of measurement times $(t_n^k)_{k=1}^{m_n}$ for each episode n. The agent collects observations only at these time points, i.e., $\{x_n(t_n^k)\}_{k=1}^{m_n}$. Importantly, we allow the measurement times to be non-uniformly spaced, meaning the measurement gap $\Delta_{n,k} := t_n^{k+1} - t_n^k$ can vary across time.

Maximum Likelihood Estimation. To construct the confidence set, we begin by examining the learning objective in CTRL. Due to the Markov property of the Itô process, for any drift-diffusion pair (f, g), policy u, state x, time s, and measurement gap Δ, the following identity holds:

$$V_{f,g}(u, x, s) = \mathbb{E}_{x' \sim p_{f,g}(u,x,\Delta)} \left[V_{f,g}(u, x', s + \Delta) \right] + \mathbb{E}_{x(\cdot) \sim p_{f,g}(u,x)} \left[\int_{t=0}^{\Delta} b(x(t), u(x(t)))\, dt \right]. \quad (4.1)$$

This can be viewed as a continuous-time analogue of the Bellman equation. It implies that to evaluate the value function $V_{f^*,g^*}(u, x_{\text{ini}}, 0)$, it suffices to estimate the marginal distribution $p_{f^*,g^*}(u, x, \Delta)$ and the trajectory distribution $p_{f^*,g^*}(u, x)$ over the interval $[0, \Delta]$. To estimate the first term in equation 4.1, we construct a confidence set $\mathcal{P}_n$ based on MLE over historical observations, as defined in line 3 of Algorithm 1, inspired by existing works about MLE for discrete-time RL (Agarwal et al., 2020; Liu et al., 2022; Wang et al., 2024a;b). Specifically, $\mathcal{P}_n$ contains all drift-diffusion pairs (f, g) whose likelihood on the conditional distribution $p_{f,g}(x_i(t_i^{k+1}) \mid u_i, x_i(t_i^k), \Delta_{i,k})$ is sufficiently close to that of the MLE solution. The proximity is controlled via a confidence radius parameter β.

We note that existing approaches (Treven et al., 2024a; Zhao et al., 2025) typically aim to learn the underlying dynamics (f^*, g^*) by directly estimating the drift term $f^*(x(t))$. In the corresponding deterministic setting where the diffusion term is zero, this drift is equivalent to the time derivative $\dot{x}(t)$. However, estimating this term from discrete and noisy trajectory data often requires non-trivial procedures like finite-difference approximations, which introduces additional algorithmic complexity and sensitivity to noise. In contrast, our approach relies solely on the observed states at discrete measurement times, making the estimation process both simpler and more robust.

Algorithm 2 Monte Carlo-Type Estimation

Require: Current episode n, history observations $\{x_i(t_i^k), x_i(t_i^k + \widehat{\Delta}_{i,k})\}_{i=1,\dots,n-1,k=0,\dots,m_i-1}$, measurement gaps $\{\Delta_{n,k}\}_{k=0,\dots,m_n-1}$.

1: Build confidence set $\widehat{\mathcal{P}}_n$ as

$$\widehat{\mathcal{P}}_n := \Bigg\{(f,g) \in \mathcal{F}\times\mathcal{G} : \sum_{i=1}^{n-1}\sum_{k=0}^{m_i-1} \log p_{f,g}(x_i(t_i^k + \widehat{\Delta}_{i,k})|u_i, x_i(t_i^k), \widehat{\Delta}_{i,k}) \geq \max_{(f',g')\in\mathcal{F}\times\mathcal{G}} \sum_{i=1}^{n-1}\sum_{k=0}^{m_i-1} \log p_{f',g'}(x_i(t_i^k + \widehat{\Delta}_{i,k})|u_i, x_i(t_i^k), \widehat{\Delta}_{i,k}) - \beta\Bigg\}.$$

2: Set $\widehat{\Delta}_{n,k} \sim \text{Unif}(0, \Delta_{n,k})$ for all $0 \leq k < m_n$.

3: **return** Confidence set $\widehat{\mathcal{P}}_n$ and observations $x_n(\widehat{t}_n^0 + \widehat{\Delta}_{n,0}), \dots, x_n(t_n^{m_n-1} + \widehat{\Delta}_{n,m_n-1})$.

Randomized Additional Measurement. The second term in equation 4.1 involves an integral over the trajectory segment $x(\cdot)$ governed by the law $p_{f,g}(u,x)$. While this integral could in principle require full knowledge of the process, it can instead be estimated using a single sample point via a Monte Carlo-style approach. To implement this, we augment CT-MLE with an additional randomized measurement step, as described in Algorithm 2. Specifically, for each interval $[t_i^k, t_i^{k+1})$, we sample a random time $\widehat{t}_{i,k} = t_i^k + \widehat{\Delta}_{i,k}$ uniformly from the interval and record the state $x_i(\widehat{t}_{i,k})$. It is worth noting that this modification requires only one additional measurement per interval, effectively doubling the number of measurements compared to CT-MLE without Algorithm 2. Using these additional samples, we construct a second confidence set $\widehat{\mathcal{P}}_n$, based on the conditional distribution $p_{f,g}(x_i(\widehat{t}_{i,k}) \mid u_i, x_i(t_i^k), \widehat{\Delta}_{i,k})$. Notably, our algorithm does not explicitly compute the integral in equation 4.1; instead, the randomized measurements serve to implicitly capture the integral's behavior by refining the confidence set around the true dynamics (f^*, g^*). This enables us to eliminate the continuity assumption without compromising performance guarantees.

5 Analysis of CT-MLE

We present the theoretical analysis of Algorithm 1. We begin by introducing the following regularity assumption, which summarizes all the conditions we impose on the system dynamics.

Assumption 5.1. The continuous-time system dynamics satisfy the following conditions:

- The policy class Π, drift class $\mathcal{F}$, and diffusion class $\mathcal{G}$ are all finite.
- The reward function $b(x,u)$ and the initial state x_{ini} are known to the agent.
- The reward function is bounded: $0 \leq b(x,u) \leq 1$ for all $(x,u) \in \mathcal{X}\times\mathcal{U}$. Furthermore, for any trajectory $x(\cdot) \sim p_{f^*,g^*}(u, x_{\text{ini}})$, the cumulative reward is bounded as $\int_0^T b(x(t), u(x(t)))\,dt \leq 1$.

Remark 5.2. The finiteness assumption on Π, $\mathcal{F}$, and $\mathcal{G}$ is made to simplify the theoretical analysis, following prior works on sample complexity in discrete-time RL (Wang et al., 2023; 2024b). Our results can be extended to infinite model classes by replacing Π, $\mathcal{F}$, and $\mathcal{G}$ with their appropriate covering sets, without changing the core analysis (Wang et al., 2023; 2024b).

Remark 5.3. The boundedness assumption on b is made for simplicity. For any general reward function b satisfying $0 \leq b(x,u) \leq B_1$ and $\int_0^T b(x(t), u(x(t)))\,dt \leq B_2$, one can normalize the reward by defining $b' := b/\max(B_1, B_2)$ and apply the algorithm and analysis to b'.

Next, we introduce the notion of *total variance* for a policy u, a concept originating from discrete-time reinforcement learning (Wang et al., 2024b; Zhou et al., 2023), which serves as an instance-dependent measure of problem hardness.

Definition 5.4. For any policy $u \in \Pi$, we define its total variance Var^u and the maximal total variance Var^Π as

$$\mathrm{Var}^u := \mathbb{V}_{x(\cdot)\sim p_{f^*,g^*}(u,x_{\mathrm{ini}})}\left[\int_0^T b\big(x(t),u(x(t))\big)\,dt\right], \quad \mathrm{Var}^\Pi := \max_{u\in\Pi}\mathrm{Var}^u.$$

By Assumption 5.1, it immediately follows that $\mathrm{Var}^u \le 1$ for any $u \in \Pi$. The total variance Var^u quantifies the uncertainty in the cumulative reward under the stochastic dynamics, and is tightly connected to the diffusion term g. The following proposition formally characterizes this dependence.

Proposition 5.5. Suppose the following conditions hold:

- The reward function b is L_b-Lipschitz continuous: for all $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{U}$,
$$|b(x,y) - b(x',y')| \le L_b\left(\|x-x'\|_2 + \|y-y'\|_2\right).$$
- The drift $f \in \mathcal{F}$ is L_f-Lipschitz continuous, and the policy $u \in \Pi$ is L_u-Lipschitz continuous:
$$\|f(x,y) - f(x',y')\|_2 \le L_f\left(\|x-x'\|_2 + \|y-y'\|_2\right), \quad \|u(x)-u(y)\|_2 \le L_u\|x-y\|_2.$$
- The diffusion term g has bounded Frobenius norm: $\|g(x,y)\|_F \le G$ for all $x \in \mathcal{X}$ and $y \in \mathcal{U}$.

Then, for any $u \in \Pi$, the total variance is bounded as

$$\mathrm{Var}^u \le \min\left\{1,\, G^2 \cdot \frac{TL_b^2(1+L_u)}{2L_f}\left(e^{2L_f(1+L_u)T} - 1\right)\right\}.$$

Proposition 5.5 shows that the total variance Var^u is controlled by the magnitude of the diffusion term G. In particular, in a deterministic environment ($G = 0$), we have $\mathrm{Var}^u = 0$ for all $u \in \Pi$. Furthermore, if the policy u is less sensitive to its input (i.e., has small L_u), the total variance is also reduced. These observations support the use of Var^u as a meaningful measure of instance difficulty in continuous-time reinforcement learning.

Next, we introduce the notion of the *eluder dimension* (Russo & Van Roy, 2013; Wang et al., 2023; 2024b; Zhao et al., 2025), which we use to characterize the complexity of the system dynamics class $\mathcal{F} \times \mathcal{G}$.

Definition 5.6. Let Ψ be a class of real-valued functions defined on a domain $\mathcal{Y}$. The *ϵ-eluder dimension* $\mathrm{DE}_p(\Psi, \mathcal{Y}, \epsilon)$ is the length of the longest sequence $y^1, \ldots, y^L \subseteq \mathcal{Y}$ such that for all $t \in [L]$, there exists $\psi \in \Psi$ satisfying $\sum_{\ell=1}^{t-1} |\psi(y^\ell)|^p \le \epsilon^p$ and $|\psi(y^t)| > \epsilon$.

In this work, we specify $\mathcal{Y} = \Pi \times \mathcal{X} \times [0,T]$ and define the function class $\Psi = \{\psi_{f,g}\}_{(f,g)\in\mathcal{F}\times\mathcal{G}}$, where

$$\psi_{f,g}(u,x,t) := \mathbb{H}^2\left(p_{f,g}(u,x,t) \,\|\, p_{f^*,g^*}(u,x,t)\right).$$

We use $d_{1/\epsilon}$ to denote $\mathrm{DE}_1(\Psi, \mathcal{Y}, \epsilon)$.

Remark 5.7. The function class Ψ is chosen for analytical clarity. First, by assuming a known reward function (Assumption 5.1), we isolate the core challenge to learning the unknown dynamics (f^*, g^*). This allows for a focused analysis of how the measurement strategy and stochasticity affect regret. While a unified analysis incorporating the reward function is common in other settings Jin et al. (2021); He et al. (2021b), its extension to continuous time is a nontrivial challenge deferred to future work. Second, using the squared Hellinger distance provides a direct analytical bridge between the statistical error of our estimator and the regret decomposition, which is central to the proof for the final regret bound.

Remark 5.8. Treven et al. (2024a) introduced a model complexity notion $\mathcal{I}_N$ based on an external estimator for the epistemic uncertainty of f^*, g^*. In contrast, our eluder dimension requires no such estimator, offering a broader, self-contained characterization. Zhao et al. (2025) also considered eluder dimension in CTRL, but theirs targets only the nonlinearity in estimating f^*, while ours captures the nonlinearity of the full induced distribution p_{f^*,g^*}, yielding a more general measure.

The following proposition shows that under a *quadratic* density model, the eluder dimension of Ψ is low.

Proposition 5.9. Suppose the marginal density function $p_{f,g}(\cdot \mid u, x, t)$ admits the form

$$p_{f,g}(\cdot \mid u, x, t) = \big(\phi(u, x, t)^\top \mu_{f,g}(\cdot)\big)^2, \quad \phi, \mu_{f,g} \in \mathbb{R}^d, \quad \|\phi\|_2 \le 1, \int_y \|\mu_{f,g}(y)\|_2^2 dy \le B,$$

then the corresponding $\psi_{f,g}$ satisfies $\mathrm{DE}_1(\Psi, \mathcal{Y}, \epsilon) \lesssim d^2 \log\left(1 + \frac{B^2}{\epsilon^2}\right)$.

While the quadratic density form in Proposition 5.9 may seem restrictive, such structures naturally emerge in continuous-time systems (see Appendix A.3 for details).

We now present our main theory.

Theorem 5.10. For any fixed grid (t_n^k), define $\boldsymbol{\Delta}_n := \sqrt{\sum_{k=0}^{m_n-1} \Delta_{n,k}^2}$ and $\boldsymbol{m}_N := \sum_{n=1}^N m_n$. Given $0 < \delta < 1$, set $\beta = \log(|\mathcal{F}||\mathcal{G}|N/\delta)$ and $\iota := \log(N/\delta)\log(\boldsymbol{m}_N)$. Denote $d_{\boldsymbol{m}_N} := \mathrm{DE}_1(\Psi, \mathcal{Y}, 1/\boldsymbol{m}_N)$ and $d_{8\beta\boldsymbol{m}_N} := \mathrm{DE}_1(\Psi, \mathcal{Y}, 1/(8\beta\boldsymbol{m}_N))$ following Definition 5.6. Then, under Assumption 5.1, with probability at least $1 - 8\delta$, we have

$$\mathrm{Regret}(N) \lesssim \iota\left(d_{8\beta\boldsymbol{m}_N}\beta + \sqrt{d_{\boldsymbol{m}_N}\beta\left(\sum_{n=1}^N \boldsymbol{\Delta}_n^2 + \sum_{n=1}^N \mathrm{Var}^{u_n}\right)}\right). \tag{5.1}$$

Proof. See Appendix B. □

To the best of our knowledge, the resulting regret bound of Algorithm 1 is the first *instance-dependent second-order regret bound* established in CTRL. Notably, the dependence on Var^{u_n} is independent of the measurement strategy, highlighting it as a fundamental quantity characterizing the intrinsic difficulty of the continuous-time system dynamics. We summarize several key remarks below.

Remark 5.11. The regret bound equation 5.1 remains unchanged as long as the total measurement budget $\boldsymbol{\Delta}_n$ is fixed. This implies that CTRL is *robust* to different choices of measurement schedules, provided the total measurement effort remains the same. This aligns with recent observations (Treven et al., 2024b) suggesting that CTRL is relatively insensitive to the minimum measurement gap $\min_k \Delta_{n,k}$. In particular, while equidistant measurements may seem natural—as they mirror discrete-time RL—they are not the only strategy capable of achieving near-optimal regret guarantees.

Remark 5.12. Many prior works on CTRL derive regret or sample complexity bounds that scale exponentially with the planning horizon T, i.e., contain terms of the form $\exp(T)$ (Treven et al., 2024a; Zhao et al., 2025), making the bounds vacuous for large T. In contrast, our regret bound in equation 5.1 depends on T only *logarithmically*, due to the use of the total variance Var^{u_n}, which is bounded by 1 under Assumption 5.1. We emphasize that avoiding the exponential dependence on T is made possible by analyzing the problem through the lens of total variance. Without this perspective, one would recover an exponential dependence on T, as shown in Proposition 5.5.

Next we discuss a more refined version of regret bound and λ-total complexity of CT-MLE.

Corollary 5.13. Suppose there exists a constant $d > 0$ such that $d \ge \max\{d_{8\beta\boldsymbol{m}_N}, d_{\boldsymbol{m}_N}, \beta\}$, where we denote $d_{\boldsymbol{m}_N} := \mathrm{DE}_1(\Psi, \mathcal{Y}, 1/\boldsymbol{m}_N)$ and $d_{8\beta\boldsymbol{m}_N} := \mathrm{DE}_1(\Psi, \mathcal{Y}, 1/(8\beta\boldsymbol{m}_N))$ following Definition 5.6. Using the notations defined in Theorem 5.10, and selecting equidistant measurements $\Delta_{n,k} = \Delta$, the regret is bounded as

$$\mathrm{Regret}(N) \lesssim \log(N/\delta)\log(TN/\Delta)\big(d^2 + d\sqrt{NT\Delta + N\mathrm{Var}^{\Pi}}\big).$$

Furthermore, to find an ϵ-optimal policy $\widehat{u}$, the λ-total complexity is bounded, up to logarithmic factors, by

$$(1-\lambda)\left(\frac{d^2}{\epsilon} + \frac{d^2\mathrm{Var}^{\Pi}}{\epsilon^2}\right) + \lambda\frac{d^2T^2}{\epsilon^2} + \frac{(1-\lambda)d^2T\Delta}{\epsilon^2} + \left(\frac{d^2}{\epsilon} + \frac{d^2\mathrm{Var}^{\Pi}}{\epsilon^2}\right)\frac{\lambda T}{\Delta}. \tag{5.2}$$

We have the following remarks about the total complexity equation 5.2.

Remark 5.14. When $\lambda = 0$, i.e., we only care about the episode complexity and ignore the measurement complexity, selecting the measurement gap as $\Delta = \mathrm{Var}^{\Pi}/T$ yields an episode complexity of $d^2\mathrm{Var}^{\Pi}/\epsilon^2$. This result suggests that to fully exploit the instance-dependent property of Algorithm 1, it suffices to choose an instance-dependent measurement gap. In particular, achieving instance-adaptive performance requires measuring more frequently in less stochastic environments. Meanwhile, the measurement complexity becomes d^2T^2/ϵ^2, which is independent of the specific problem instance.

Remark 5.15. When $\lambda = 1$, i.e., we focus solely on the measurement complexity and ignore the episode complexity, the optimal choice is $\Delta = T$. The total measurement complexity is proportional to $\frac{d^2\mathrm{Var}^{\Pi}}{\epsilon^2} \cdot \frac{T}{\Delta}$. To minimize this expression, Δ must be maximized. This implies a sparse sampling strategy where for each episode, we collect samples at the start and end points, $x(0)$ and $x(T)$, along with one additional sample at a random time $\widehat{t} \in [0, T]$. This result highlights a theoretical trade-off, favoring many "measurement-cheap" episodes over a few "measurement-expensive" ones. Interestingly, the measurement complexity asymptotically matches the complexity when episode complexity is the sole focus $\lambda = 0$. This observation leads to an interesting conjecture: the problem instance influences only the episode complexity, but not the measurement complexity. Verifying the tightness of these bounds remains an open direction for future work.

6 Conclusion and Limitations

Conclusion. In this work, we presented CT-MLE, a simple and general model-based algorithm for CTRL that learns through marginal density estimation rather than explicit dynamic modeling. Our approach leverages MLE with flexible function approximators, enabling compatibility with a wide range of policy classes and continuous-time settings. We introduced a randomized measurement strategy, including a Monte Carlo-style scheme that provides unbiased integral estimation while preserving measurement efficiency. Theoretically, we established regret bounds that reveal the benefit of instance-dependent measurement schedules, and we demonstrated that the regret can be made primarily dependent on total reward variance, effectively decoupling it from fixed measurement grids.

Limitations. While our work provides a theoretical foundation, several gaps remain. First, we assume access to general function approximators, but do not provide a computationally efficient, provably correct algorithm. A key next step is to develop an adaptive method that estimates variance online and sets measurement gaps accordingly. Second, our analysis relies on a simplified continuous-time structure for tractability, which may not hold in practice. Future work could identify realistic dynamics that still support Eluder-dimension-based analysis. Third, our framework assumes a known deterministic reward and stationary policy. Extending to stochastic rewards and time-varying policies $u(t, x)$ would require generalizing existing tools to the joint state-time domain. Lastly, empirical validation on standard CTRL benchmarks is needed to assess the practical utility of our approach.

Ethics Statement

Our study develops and analyzes algorithms for continuous-time reinforcement learning (CTRL) using a theoretical SDE-based formulation and episodic learning protocol; it does not involve human subjects, personally identifiable information, or sensitive data, and all experiments are performed in simulator settings (standard RL environments) rather than on physical systems. The work focuses on algorithmic methods (Algorithm 1, 2) and formal analysis, not deployment, thereby avoiding direct safety risks in real-world control; nevertheless, we caution that applying any learned policy to safety-critical domains (e.g., robotics, healthcare, finance) should include appropriate risk assessment, domain-specific safeguards, and compliance checks.

Reproducibility statement

We facilitate reproducibility by referencing precise locations of all necessary components: the formal problem setup (Section 3) and learning protocol, the complete algorithmic specification (Algorithm 1 and randomized measurement Algorithm 2), and full theoretical details, assumptions, and proofs in the appendix (Appendix B with supporting lemmas). Experimental settings, implementation specifics, and environment configurations are documented in the "Numerical Experiments" appendix (Appendix C), including "Implementation Details," main results, and ablations, with further clarifications in "Additional Details". Together, these materials specify objectives, schedules, and measurement strategies sufficient to reproduce the reported results or re-create them under equivalent simulator conditions.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 32:96, 2019.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

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.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Peter E Caines and David Levanony. Stochastic ε-optimal linear quadratic adaptation: An alternating controls policy. *SIAM Journal on Control and Optimization*, 57(2):1094–1126, 2019.

Nicolo Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66:321–352, 2007.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Christoph Dann, Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34:1–12, 2021.

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

Kacha Dzhaparidze and JH Van Zanten. On bernstein-type inequalities for martingales. *Stochastic processes and their applications*, 93(1):109–117, 2001.

Xiequan Fan, Ion Grama, and Quansheng Liu. Martingale inequalities of type dzhaparidze and van zanten. *Statistics*, 51(6):1200–1213, 2017.

Mohamad Kazem Shirani Faradonbeh and Mohamad Sadegh Shirani Faradonbeh. Online reinforcement learning in stochastic continuous-time systems. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 612–656. PMLR, 2023.

Dylan J Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems*, 34: 18907–18919, 2021.

Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. *Advances in neural information processing systems*, 32, 2019.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pp. 4171–4180. PMLR, 2021a.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Uniform-pac bounds for reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems*, 34: 14188–14199, 2021b.

Samuel Holt, Alihan Hüyük, and Mihaela van der Schaar. Active observing in continuous-time control. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Samuel Holt, Alihan Hüyük, and Mihaela van der Schaar. Active observing in continuous-time control. *Advances in Neural Information Processing Systems*, 36, 2024.

Yilie Huang, Yanwei Jia, and Xun Yu Zhou. Sublinear regret for a class of continuous-time linear–quadratic reinforcement learning problems. *arXiv preprint arXiv:2407.17226*, 2024.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Shinji Ito, Shuichi Hirahara, Tasuku Soma, and Yuichi Yoshida. Tight first-and second-order regret bounds for adversarial linear bandits. *Advances in Neural Information Processing Systems*, 33: 2028–2038, 2020.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. *arXiv preprint arXiv:1907.05388*, 2019.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, pp. 4870–4879. PMLR, 2020.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021.

Amirmohammad Karimi. Decision frequency adaptation in reinforcement learning using continuous options with open-loop policies, 2023.

Yeoneung Kim, Insoon Yang, and Kwang-Sung Jun. Improved regret analysis for variance-adaptive linear bandits and horizon-free linear mixture mdps. *Advances in Neural Information Processing Systems*, 35:1060–1072, 2022.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Qinghua Liu, Alan Chung, Csaba Szepesvári, and Chi Jin. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pp. 5175–5220. PMLR, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Michael Lutter, Shie Mannor, Jan Peters, Dieter Fox, and Animesh Garg. Value iteration in continuous actions, states and time. In *International Conference on Machine Learning*, pp. 7224–7234. PMLR, 2021.

Tianwei Ni and Eric Jang. Continuous control on time. In *ICLR 2022 Workshop on Generalizable Policy Learning in Physical World*, 2022. URL https://openreview.net/forum?id=BtbG3NT4y-c.

Chirag Pabbaraju, Dhruv Rohatgi, Anish Prasad Sevekari, Holden Lee, Ankur Moitra, and Andrej Risteski. Provable benefits of score matching. *Advances in Neural Information Processing Systems*, 36:61306–61326, 2023.

Hannes Risken and Till Frank. *The Fokker-Planck Equation: Methods of Solution and Applications*, volume 18. Springer Science & Business Media, 1996.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pp. 387–395. Pmlr, 2014.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. *Advances in Neural Information Processing Systems*, 32, 2019.

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in artificial intelligence*, pp. 574–584. PMLR, 2020.

Lukasz Szpruch, Tanut Treetanthiploet, and Yufei Zhang. Optimal scheduling of entropy regularizer for continuous-time linear-quadratic reinforcement learning. *SIAM Journal on Control and Optimization*, 62(1):135–166, 2024.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulao, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

Lenart Treven, Jonas Hübotter, Florian Dorfler, and Andreas Krause. Efficient exploration in continuous-time model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024a.

Lenart Treven, Bhavya Sukhija, Yarden As, Florian Dörfler, and Andreas Krause. When to sense and control? a time-adaptive approach for continuous-time rl. *arXiv preprint arXiv:2406.01163*, 2024b.

Draguna Vrabie and Frank Lewis. Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems. *Neural Networks*, 22(3):237–246, 2009.

Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In *International Conference on Machine Learning*, pp. 22384–22429. PMLR, 2022.

Kaiwen Wang, Kevin Zhou, Runzhe Wu, Nathan Kallus, and Wen Sun. The benefits of being distributional: Small-loss bounds for reinforcement learning. *Advances in neural information processing systems*, 36:2275–2312, 2023.

Kaiwen Wang, Owen Oertell, Alekh Agarwal, Nathan Kallus, and Wen Sun. More benefits of being distributional: Second-order bounds for reinforcement learning. *arXiv preprint arXiv:2402.07198*, 2024a.

Zhiyong Wang, Dongruo Zhou, John Lui, and Wen Sun. Model-based rl as a minimalist approach to horizon-free and second-order bounds. *arXiv preprint arXiv:2408.08994*, 2024b.

Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4230–4239, 2023.

Haike Xu, Tengyu Ma, and Simon Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. In *Conference on Learning Theory*, pp. 4438–4472. PMLR, 2021.

Cagatay Yildiz, Markus Heinonen, and Harri Lähdesmäki. Continuous-time model-based reinforcement learning. In *International Conference on Machine Learning*, pp. 12009–12018. PMLR, 2021.

TaeHo Yoon, Kibeom Myoung, Keon Lee, Jaewoong Cho, Albert No, and Ernest Ryu. Censored sampling of diffusion models using 3 minutes of human feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312. PMLR, 2019.

Zihan Zhang, Jiaqi Yang, Xiangyang Ji, and Simon S Du. Improved variance-aware confidence sets for linear bandits and linear mixture mdp. *Advances in Neural Information Processing Systems*, 34:4342–4355, 2021.

Zihan Zhang, Xiangyang Ji, and Simon Du. Horizon-free reinforcement learning in polynomial time: the power of stationary policies. In *Conference on Learning Theory*, pp. 3858–3904. PMLR, 2022.

Heyang Zhao, Dongruo Zhou, Jiafan He, and Quanquan Gu. Bandit learning with general function classes: Heteroscedastic noise and variance-dependent regret bounds. *CoRR*, 2022.

Heyang Zhao, Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Variance-dependent regret bounds for linear bandits and reinforcement learning: Adaptivity and computational efficiency. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 4977–5020. PMLR, 2023.

Runze Zhao, Yue Yu, Adams Yiyue Zhu, Chen Yang, and Dongruo Zhou. Sample and computationally efficient continuous-time reinforcement learning with general function approximation. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025. URL https://openreview.net/forum?id=JFSZewaXaG.

Yaofeng Desmond Zhong and Naomi Leonard. Unsupervised learning of lagrangian dynamics from images for prediction and control. *Advances in Neural Information Processing Systems*, 33: 10741–10752, 2020.

Dongruo Zhou and Quanquan Gu. Computationally efficient horizon-free reinforcement learning for linear mixture mdps. *Advances in neural information processing systems*, 35:36337–36349, 2022.

Dongruo Zhou, Quanquan Gu, and Csaba Szepesvari. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pp. 4532–4576. PMLR, 2021a.

Dongruo Zhou, Jiafan He, and Quanquan Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pp. 12793–12802. PMLR, 2021b.

Runlong Zhou, Zhang Zihan, and Simon Shaolei Du. Sharp variance-dependent bounds in reinforcement learning: Best of both worlds in stochastic and deterministic environments. In *International Conference on Machine Learning*, pp. 42878–42914. PMLR, 2023.

CONTENTS OF THE APPENDIX

THE USE OF LARGE LANGUAGE MODELS (LLMS)

LLMs were used solely for language polishing; all ideas, analyses, and conclusions are the authors' own, and the authors take full responsibility for the final text.

A ADDITIONAL RESULTS FROM MAIN PAPER

A.1 PROOF OF PROPOSITION 5.5

Proof. For any deterministic policy u, we have

$$\mathrm{Var}^u = \mathbb{E}_{x(\cdot)\sim p_{f^*,g^*}(u,x_{\text{ini}})}\left[\int_0^T b(x(t),u(x(t))) - \mathbb{E}_{x(\cdot)\sim p_{f^*,g^*}(u,x_{\text{ini}})}\int_0^T b(x(t),u(x(t)))\right]^2.$$

Applying the Cauchy–Schwarz inequality yields

$$\begin{aligned}\mathrm{Var}^u &\le \mathbb{E}_{x(\cdot)\sim p_{f^*,g^*}(u,x_{\text{ini}})}\left[\int_0^T \left(b(x(t),u(x(t))) - \mathbb{E}_{x\sim p_t} b(x,u(x))\right)^2 dt\right] \\ &= \int_0^T \mathbb{E}_{x\sim p_t}\left(b(x,u(x)) - \mathbb{E}_{x'\sim p_t} b(x',u(x'))\right)^2 dt, \end{aligned} \tag{A.1}$$

where we denote $p_t = p_{f^*,g^*}(u, x_{\text{ini}}, t)$ for simplicity.

For the integrand in equation A.1, by the Lipschitz continuity of b and u, we have

$$\begin{aligned}&\mathbb{E}_{x\sim p_t}\left(b(x,u(x)) - \mathbb{E}_{x'\sim p_t} b(x',u(x'))\right)^2 \\ &= \mathbb{E}_{x,x'\sim p_t}\left(b(x,u(x)) - b(x',u(x'))\right)^2 \\ &\le \mathbb{E}_{x,x'\sim p_t}\left(L_b(\|x-x'\|_2 + L_u\|x-x'\|_2)\right)^2 \\ &= L_b^2(1+L_u)^2\mathbb{E}_{x,x'\sim p_t}\|x-x'\|_2^2 \\ &= 2L_b^2(1+L_u)^2\,\mathbb{E}_{x\sim p_t}\left\|x - \mathbb{E}_{x'\sim p_t}x'\right\|_2^2.\end{aligned}$$

Define

$$V(t) := \mathbb{E}_{x\sim p_t}\left\|x - \mathbb{E}_{x'\sim p_t}x'\right\|_2^2, \quad \mu(t) := \mathbb{E}_{x\sim p_t}[x].$$

Next we calculate the derivate of $V(t)$. First, by applying Itô's formula to $\|x(t)\|^2$, we have

$$\begin{aligned}d\|x(t)\|_2^2 = {}& 2\langle x(t), f(x(t),u(x(t)))\rangle\,dt + \|g(x(t),u(x(t)))\|_F^2\,dt \\ &+ 2\langle x(t), g(x(t),u(x(t)))\,dw(t)\rangle.\end{aligned}$$

Then taking expectation for both side and using the fact $\mathbb{E}dw(t) = 0$, we have

$$\frac{d}{dt}\mathbb{E}\|x(t)\|_2^2 = 2\,\mathbb{E}\langle x(t), f(x(t),u(x(t)))\rangle + \mathbb{E}\|g(x(t),u(x(t)))\|_F^2.$$

Next, we have

$$\frac{d}{dt}\|\mu(t)\|_2^2 = 2\big\langle \mu(t), \mathbb{E}f(x(t),u(x(t)))\big\rangle.$$

Then by the fact that $V(t) = \mathbb{E}\|x(t)\|_2^2 - \|\mu(t)\|_2^2$ we obtain

$$\begin{aligned}\frac{d}{dt}V(t) &= 2\,\mathbb{E}\left[\langle x(t)-\mu(t), f(x(t),u(x(t))) - f(\mu(t),u(\mu(t)))\rangle\right] + \mathbb{E}\left[\|g(x(t),u(x(t)))\|_F^2\right] \\ &\le 2\,\mathbb{E}\|x(t)-\mu(t)\|_2 \cdot \|f(x(t),u(x(t))) - f(\mu(t),u(\mu(t)))\|_2 + \mathbb{E}\|g(x(t),u(x(t)))\|_F^2 \\ &\le 2L_f(1+L_u)V(t) + G^2,\end{aligned}$$

where the last inequality follows from the Lipschitz continuity of f and u.

Applying Grönwall's lemma, we get

$$V(t) \le \frac{G^2}{2L_f(1+L_u)}\left(e^{2L_f(1+L_u)t} - 1\right) \le \frac{G^2}{2L_f(1+L_u)}\left(e^{2L_f(1+L_u)T} - 1\right). \tag{A.2}$$

Substituting equation A.2 into equation A.1 completes the proof. □

A.2 Proof of Proposition 5.9

Proof. By the definition of Hellinger distance, we have H

$$\begin{aligned}\mathbb{H}^2(p_{f,g}(u,x,t)\|p_{f^*,g^*}(u,x,t)) &= 1-\int_y \sqrt{p_{f,g}(y|u,x,t)\cdot p_{f^*,g^*}(y|u,x,t)}\,dy \\ &= 1-\int_y \phi(u,x,t)^\top \mu_{f,g}(y)\mu_{f^*,g^*}(y)^\top \phi(u,x,t)\,dy \\ &= 1-\phi(u,x,t)^\top \left[\int_y \mu_{f,g}(y)\mu_{f^*,g^*}(y)^\top dy\right]\phi(u,x,t). \quad \text{(A.3)}\end{aligned}$$

Therefore, the squared Hellinger distance is a linear function of the feature matrix $\phi(u,x,t)\phi(u,x,t)^\top \in \mathbb{R}^{d\times d}$. Since $\|\phi(u,x,t)\|_2 \le 1$, it follows that

$$\|\phi(u,x,t)\phi(u,x,t)^\top\|_F \le 1. \quad \text{(A.4)}$$

Now we bound the Frobenius norm of the matrix inside the integral:

$$\begin{aligned}\left\|\int_y \mu_{f,g}(y)\mu_{f^*,g^*}(y)^\top dy\right\|_F &\le \int_y \|\mu_{f,g}(y)\|_2\cdot\|\mu_{f^*,g^*}(y)\|_2\,dy \\ &\le \left(\int_y \|\mu_{f,g}(y)\|_2^2 dy\right)^{1/2}\left(\int_y \|\mu_{f^*,g^*}(y)\|_2^2 dy\right)^{1/2} \\ &\le B, \quad \text{(A.5)}\end{aligned}$$

where the last inequality uses the Cauchy–Schwarz inequality and the assumed boundedness of the μ functions.

Putting together the bounds in equation A.3, equation A.4, and equation A.5, and invoking Proposition 19 in Liu et al. (2022) and Proposition 6 in Russo & Van Roy (2013), we conclude that

$$\mathrm{DE}_1(\Psi,\mathcal{Y},\epsilon) \le \mathrm{DE}_2(\Psi,\mathcal{Y},\epsilon) \lesssim d^2\log\left(1+\frac{B^2}{\epsilon^2}\right).$$

□

A.3 Construction example for Proposition 5.9

The quadratic form presented in Proposition 5.9 is well-motivated and can be constructed explicitly. For simplicity, let us consider a quadratic density function $p(y \mid t)$ that is independent of policy u and state x. Let us assume $p(y \mid t) = (\phi(t)^\top \mu(y))^2$ with $\phi(t), \mu(y) \in \mathbb{R}^2$. Then we can take $\phi(t) = (\cos(t), \sin(t))^\top$ and $\mu(y) = (c_1 e^{-y^2}, c_2 y e^{-y^2})^\top$, where $c_1 = (2/\pi)^{1/4}$ and $c_2 = 2(2/\pi)^{1/4}$. The resulting density is

$$p(y \mid t) = \left[(2/\pi)^{1/4}\cos(t)e^{-y^2} + 2(2/\pi)^{1/4}\sin(t)ye^{-y^2}\right]^2.$$

This defines a valid, time-evolving probability density because the basis functions in $\mu(y)$ are orthonormal, satisfying $\int \mu_i(y)\mu_j(y)\,dy = \delta_{ij}$, and the coefficients in $\phi(t)$ satisfy $\cos^2(t)+\sin^2(t) = 1$, ensuring $\int p(y \mid t)\,dy = 1$ for all t.

The drift $f(y,t)$ and diffusion $g(y,t)$ of an SDE generating this density can be obtained from the Fokker–Planck equation $\partial_t p = -\partial_y(fp) + \frac{1}{2}\partial_y^2(g^2 p)$. Setting $g = 1$, we have

$$f(y,t) = \frac{1}{p(y \mid t)}\int_{-\infty}^{y}\left[\frac{1}{2}\frac{\partial^2 p}{\partial z^2} - \frac{\partial p}{\partial t}\right]dz.$$

Although the resulting drift does not have a simple closed form, it can be computed explicitly given $p(y \mid t)$. In this sense, the SDE with $(f, 1)$ provides a valid example satisfying Proposition 5.9.

Additionally, though classical SDEs do not directly yield the quadratic form of Proposition 5.9, we can identify related structures in well-known processes. The classical Ornstein–Uhlenbeck (OU) process provides a case that satisfies a linear form $p(y \mid t) = \phi(t)^\top \mu(y)$. Its spectral representation (see Chapter 5.4 of Risken & Frank (1996)) is given by

$$p(y \mid t, y_0) = \sum_{n=0}^{\infty} e^{\lambda_n t} \psi_n(y) \psi_n(y_0),$$

where $\lambda_n = -n\gamma$ with $\gamma > 0$ denoting the mean-reversion rate, and $\{\psi_n(y)\}$ are the Hermite eigenfunctions of the corresponding OU generator. This representation constitutes a linear inner product in an infinite-dimensional space, illustrating that such structures arise naturally even when the SDE itself has simple drift and diffusion coefficients.

B Proof of Main Theorem

We first define several notations for convenience. Let

$$p_n^*(x,t) := p_{f^*,g^*}(u_n, x, t), \qquad p_n(x,t) := p_{f_n,g_n}(u_n, x, t),$$
$$V_n^*(x,t) := V_{f^*,g^*}(u_n, x, t), \qquad V_n(x,t) := V_{f_n,g_n}(u_n, x, t).$$

B.1 Auxiliary lemmas

The following lemma shows that the difference in expectations between two distributions can be bounded by the variance of one distribution and their Hellinger distance, which plays a key role in deriving our variance-dependent regret bound.

Lemma B.1 (Wang et al. 2024b;a)**.** Let $p, q \in \Delta([0,1])$ be two probability distributions over $[0,1]$. Define the variance of p as

$$\mathrm{VaR}_p := \mathbb{E}_{x\sim p}\left[(x - \mathbb{E}_{x\sim p}[x])^2\right].$$

Then the following inequality holds:

$$|\mathbb{E}_{x\sim p}[x] - \mathbb{E}_{x\sim q}[x]| \lesssim \sqrt{\mathrm{VaR}_p \cdot \mathbb{H}^2(p \,\|\, q)} + \mathbb{H}^2(p \,\|\, q).$$

The following lemma provides a concentration inequality for martingale difference sequences without boundedness assumptions, which is essential for handling heavy-tailed or unbounded noise in our analysis.

Lemma B.2 (Unbounded Freedman's inequality, Dzhaparidze & Van Zanten (2001); Fan et al. (2017))**.** Let $\{x_i\}_{i=1}^n$ be a stochastic process adapted to a filtration $\{\mathcal{G}_i\}_{i=1}^n$, where $\mathcal{G}_i = \sigma(x_1, \ldots, x_i)$. Suppose $\mathbb{E}[x_i \mid \mathcal{G}_{i-1}] = 0$ and $\mathbb{E}[x_i^2 \mid \mathcal{G}_{i-1}] < \infty$ almost surely. Then, for any $a, v, y > 0$, we have

$$\mathbb{P}\left(\sum_{i=1}^n x_i > a,\ \sum_{i=1}^n \left(\mathbb{E}[x_i^2 \mid \mathcal{G}_{i-1}] + x_i^2 \cdot \mathbb{1}\{|x_i| > y\}\right) < v^2\right) \le \exp\left(\frac{-a^2}{2(v^2 + ay/3)}\right).$$

Equivalently, with probability at least $1-\delta$, the following high-probability bound holds:

$$\sum_{i=1}^n x_i \le \sqrt{2\sum_{i=1}^n \left(\mathbb{E}[x_i^2 \mid \mathcal{G}_{i-1}] + x_i^2 \cdot \mathbb{1}\{|x_i| > y\}\right)\log(1/\delta)} + \frac{y}{3}\log(1/\delta).$$

The next lemma bounds the sum of truncated random variables in terms of their conditional expectations, which is useful for controlling tail contributions in martingale-adapted processes.

Lemma B.3 (Lemma 8, Zhang et al. 2022)**.** Let $\{x_i\}_{i=1}^n$ be a nonnegative stochastic process adapted to a filtration $\{\mathcal{G}_i\}_{i\ge 1}$, i.e., $x_i \ge 0$ almost surely. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\sum_{i=1}^n \min\{x_i, y\} \le 4y\log(4/\delta) + 4\log(4/\delta)\sum_{i=1}^n \mathbb{E}[x_i \mid \mathcal{G}_{i-1}].$$

We also include the following two auxiliary lemmas that will be used in our analysis.

Lemma B.4 (Lemma 11, Wang et al. 2024b)**.** Let $G > 0$ and $a < G/2$ be positive constants. Let $\{C_i\}_{i=0}^{M}$ be a sequence of positive real numbers, where $M = \lceil \log_2(H/G) \rceil$, satisfying:

- $C_i \leq 2^i G + \sqrt{aC_{i+1}} + a$ for all $i \geq 0$;
- $C_i \leq H$ for all $i \geq 0$, where $H > 0$ is a positive constant.

Then it holds that $C_0 \leq 4G$.

Lemma B.5. For any random variable $X \in [0, 1]$, we have $\mathrm{Var}(X^2) \leq 4\mathrm{Var}(X)$.

Proof. Let Y be an independent copy of X. Then,

$$\mathrm{Var}(X^2) = \tfrac{1}{2}\,\mathbb{E}[(X^2 - Y^2)^2] = \tfrac{1}{2}\,\mathbb{E}[(X-Y)^2(X+Y)^2] \leq \tfrac{1}{2} \cdot 4\,\mathbb{E}[(X-Y)^2] = 2\,\mathbb{E}[(X-Y)^2],$$

where we used $(X+Y)^2 \leq 4$ since $X, Y \in [0, 1]$.

Next, we observe:

$$\mathbb{E}[(X-Y)^2] = \mathbb{E}[X^2] + \mathbb{E}[Y^2] - 2\,\mathbb{E}[X]\,\mathbb{E}[Y] = 2\,\mathbb{E}[X^2] - 2\,\mathbb{E}[X]^2 = 2\,\mathrm{Var}(X),$$

where the last equality follows from $\mathbb{E}[X] = \mathbb{E}[Y]$ and $\mathbb{E}[X^2] = \mathbb{E}[Y^2]$. Combining both steps, we get

$$\mathrm{Var}(X^2) \leq 2 \cdot 2\,\mathrm{Var}(X) = 4\,\mathrm{Var}(X),$$

which concludes the proof. □

B.2 Lemmas on confidence sets

We now introduce a sequence of lemmas that are instrumental for proving Theorems 5.10.

Lemma B.6. With probability at least $1 - \delta$, the following holds for all $n \in [N]$: $(f^*, g^*) \in \mathcal{P}_n$, and

$$\sum_{i=1}^{n-1} \sum_{k=0}^{m_i - 1} \mathbb{H}^2 \left(p_{f_n, g_n}(u_i, x_i(t_i^k), \Delta_{i,k}) \,\|\, p_{f^*, g^*}(u_i, x_i(t_i^k), \Delta_{i,k}) \right) \leq \beta, \tag{B.1}$$

where $\beta = \log\left(|\mathcal{F}||\mathcal{G}|N/\delta\right)$.

Proof. We apply Theorem E.4 in Wang et al. (2023) to the function class $\mathcal{F} \times \mathcal{G}$, using delta distributions $D_{i,k}$ centered at $(u_i, x_i(t_i^k), \Delta_{i,k})$. This guarantees that $(f^*, g^*) \in \mathcal{P}_n$ and that inequality equation B.1 holds with probability at least $1 - \delta$, for the specified choice of β. □

Next, we present a key lemma that uses the eluder dimension to bound the accumulated Hellinger distances.

Lemma B.7. Let $\mathcal{E}_{B.6}$ denote the event described in Lemma B.6. Then, under event $\mathcal{E}_{B.6}$, there exists a subset $\mathcal{N} \subseteq [N]$ such that:

- $|\mathcal{N}| \leq 13 \log^2(4\beta \boldsymbol{m}_N) \cdot d_{8\beta \boldsymbol{m}_N}$;
- For each $n \in [N]$, the indicator $n \in \mathcal{N}$ corresponds to a stopping time;
- The cumulative Hellinger distance outside $\mathcal{N}$ is bounded:

$$\sum_{i \in [N] \setminus \mathcal{N}} \sum_{k=0}^{m_i - 1} \mathbb{H}^2 \left(p_{f_i, g_i}(u_i, x_i(t_i^k), \Delta_{i,k}) \,\|\, p_{f^*, g^*}(u_i, x_i(t_i^k), \Delta_{i,k}) \right) \leq 3 d_{\boldsymbol{m}_N} + 7 d_{\boldsymbol{m}_N} \beta \log(\boldsymbol{m}_N).$$

Proof. We apply Lemma 6 from Wang et al. (2024b), using the distribution class $p_{f,g}$, the input space $\Pi \times \mathcal{X} \times [T]$, and the function class Ψ. □

B.3 LEMMAS ABOUT REGRET DECOMPOSITION

The following lemma provides a decomposition of the regret into four interpretable components based on differences between the learned and ground-truth dynamics.

Lemma B.8 (Simulation Lemma, Agarwal et al. 2019)**.** At episode n, the following decomposition holds:

$$V_n(x_{\text{ini}}, 0) - V_n^*(x_{\text{ini}}, 0) = I_{0,n} + \sum_{k=0}^{m_n-1} \left(I_{1,n}^k + I_{2,n}^k + I_{3,n}^k + I_{4,n}^k \right),$$

where the individual terms are defined as follows:

$$\begin{aligned}
I_{0,n} &:= \int_0^T b(x_n(t), u_n(t))\, dt - V_n^*(x_{\text{ini}}, 0), \\
I_{1,n}^k &:= \mathbb{E}_{x \sim p_n^*(x_n(t_n^k), \Delta_{n,k})} V_n(x, t_n^{k+1}) - V_n(x_n(t_n^{k+1}), t_n^{k+1}), \\
I_{2,n}^k &:= \mathbb{E}_{x(\cdot) \sim p_n^*(x_n(t_n^k))} \left[\int_0^{\Delta_{n,k}} b(x(t), u_n(t))\, dt \right] - \int_{t_n^k}^{t_n^{k+1}} b(x_n(t), u_n(t))\, dt, \\
I_{3,n}^k &:= \mathbb{E}_{x \sim p_n(x_n(t_n^k), \Delta_{n,k})} V_n(x, t_n^{k+1}) - \mathbb{E}_{x \sim p_n^*(x_n(t_n^k), \Delta_{n,k})} V_n(x, t_n^{k+1}), \\
I_{4,n}^k &:= \mathbb{E}_{x(\cdot) \sim p_n(x_n(t_n^k))} \left[\int_0^{\Delta_{n,k}} b(x(t), u_n(t))\, dt \right] - \mathbb{E}_{x(\cdot) \sim p_n^*(x_n(t_n^k))} \left[\int_0^{\Delta_{n,k}} b(x(t), u_n(t))\, dt \right].
\end{aligned}$$

Proof. We apply a telescoping argument over the discretization grid $\{t_n^k\}_{k=0}^{m_n}$. From the definition of the value function, we have

$$\begin{aligned}
V_n(x_{\text{ini}}, 0) &= \mathbb{E}_{x(\cdot) \sim p_n(x_{\text{ini}})} \left[\int_0^T b(x(t), u_n(t))\, dt \right] \\
&= \mathbb{E}_{x(\cdot) \sim p_n(x_{\text{ini}})} \left[\int_0^{t_n^1} b(x(t), u_n(t))\, dt \right] + \mathbb{E}_{x \sim p_n(x_{\text{ini}}, \Delta_{n,0})} V_n(x, t_n^1).
\end{aligned}$$

Subtracting the realized cumulative reward yields

$$\begin{aligned}
&V_n(x_{\text{ini}}, 0) - \int_0^T b(x_n(t), u_n(t))\, dt \\
&= \underbrace{\mathbb{E}_{x(\cdot) \sim p_n(x_{\text{ini}})} \left[\int_0^{t_n^1} b(x(t), u_n(t))\, dt \right] - \int_0^{t_n^1} b(x_n(t), u_n(t))\, dt}_{I_{2,n}^0 + I_{4,n}^0} \\
&\quad + \underbrace{\mathbb{E}_{x \sim p_n(x_{\text{ini}}, \Delta_{n,0})} V_n(x, t_n^1) - \mathbb{E}_{x \sim p_n^*(x_{\text{ini}}, \Delta_{n,0})} V_n(x, t_n^1)}_{I_{3,n}^0} \\
&\quad + \underbrace{\mathbb{E}_{x \sim p_n^*(x_{\text{ini}}, \Delta_{n,0})} V_n(x, t_n^1) - V_n(x_n(t_n^1), t_n^1)}_{I_{1,n}^0} \\
&\quad + V_n(x_n(t_n^1), t_n^1) - \int_{t_n^1}^T b(x_n(t), u_n(t))\, dt. \qquad \text{(B.2)}
\end{aligned}$$

By the Markov property of the Itô SDE, we have

$$\mathbb{E}_{x(\cdot) \sim p_{f,g}(u,x)} \left[\int_{t_n^k}^{t_n^{k+1}} b(x(t), u(t))\, dt \right] = \mathbb{E}_{x(\cdot) \sim p_{f,g}(u,x)} \left[\int_0^{\Delta_{n,k}} b(x(t), u(t))\, dt \right].$$

Using this identity recursively up to some $0 \le m^\dagger \le m_n$ leads to the expression

$$
\begin{aligned}
&V_n(x_{\text{ini}}, 0) - \int_0^T b(x_n(t), u_n(t))\, dt \\
&= \sum_{k=0}^{m^\dagger - 1} \left(I_{1,n}^k + I_{2,n}^k + I_{3,n}^k + I_{4,n}^k\right) \\
&\quad + \underbrace{\mathbb{E}_{x(\cdot)\sim p_n(x_n(t_n^{m^\dagger}))}\left[\int_{t_n^{m^\dagger}}^{t_n^{m^\dagger+1}} b(x(t), u(t))\, dt\right] - \int_{t_n^{m^\dagger}}^{t_n^{m^\dagger+1}} b(x_n(t), u_n(t))\, dt}_{I_{2,n}^{m^\dagger} + I_{4,n}^{m^\dagger}} \\
&\quad + \underbrace{\mathbb{E}_{x\sim p_n(x_n(t_n^{m^\dagger}), \Delta_{n,m^\dagger})} V_n(x, t_n^{m^\dagger+1}) - \mathbb{E}_{x\sim p_n^*(x_n(t_n^{m^\dagger}), \Delta_{n,m^\dagger})} V_n(x, t_n^{m^\dagger+1})}_{I_{3,n}^{m^\dagger}} \\
&\quad + \underbrace{\mathbb{E}_{x\sim p_n^*(x_n(t_n^{m^\dagger}), \Delta_{n,m^\dagger})} V_n(x, t_n^{m^\dagger+1}) - V_n(x_n(t_n^{m^\dagger+1}), t_n^{m^\dagger+1})}_{I_{1,n}^{m^\dagger}} \\
&\quad + V_n(x_n(t_n^{m^\dagger+1}), t_n^{m^\dagger+1}) - \int_{t_n^{m^\dagger+1}}^{T} b(x_n(t), u_n(t))\, dt. \qquad \text{(B.3)}
\end{aligned}
$$

Applying equation B.3 with $m^\dagger = m_n - 1$ and noting that $t_n^{m_n} = T$ and $V_n(\cdot, T) = 0$, we obtain

$$
V_n(x_{\text{ini}}, 0) - \int_0^T b(x_n(t), u_n(t))\, dt = \sum_{k=0}^{m_n - 1} \left(I_{1,n}^k + I_{2,n}^k + I_{3,n}^k + I_{4,n}^k\right),
$$

which completes the proof. □

Lemma B.9. Let $I_{0,n}, I_{1,n}^k, \ldots, I_{3,n}^k$ be terms introduced in Lemma B.8. Let $\widetilde{\mathcal{N}} \subseteq [N]$ be an episode index set satisfying $\widetilde{\mathcal{N}} \subseteq [N] \setminus \mathcal{N}$ and satisfying $n \in \widetilde{\mathcal{N}}$ is a stopping time. Then under event $\mathcal{E}_{B.6}$, with probability at least $1 - 4\delta$, the following bounds hold:

$$
\begin{aligned}
\sum_{n\in\widetilde{\mathcal{N}}} I_{0,n} &\lesssim \sqrt{\log(1/\delta) \sum_{n=1}^{N} \mathrm{Var}_{f^*,g^*}^{u_n} + \log(1/\delta)}, \\
\sum_{n\in\widetilde{\mathcal{N}}} \sum_{k=0}^{m_n-1} I_{1,n}^k &\lesssim \sqrt{\sum_{n\in\widetilde{\mathcal{N}}} \sum_{k=0}^{m_n-1} \mathbb{V}_{x\sim p_n^*(x_n(t_n^k), \Delta_{n,k})}\left[V_n(x, t_n^{k+1})\right] \cdot \log(1/\delta) + \log(1/\delta)}, \\
\sum_{n\in\widetilde{\mathcal{N}}} \sum_{k=0}^{m_n-1} I_{2,n}^k &\lesssim \sqrt{\sum_{n=1}^{N} \boldsymbol{\Delta}_n^2 \log(1/\delta)}, \\
\sum_{n\in\widetilde{\mathcal{N}}} \sum_{k=0}^{m-1} |I_{3,n}^k| &\lesssim \sqrt{d_{\boldsymbol{m}_N} \beta \log(\boldsymbol{m}_N) \sum_{n\in\widetilde{\mathcal{N}}} \sum_{k=0}^{m_n-1} \left[\mathbb{V}_{x\sim p_n^*(x_n(t_n^k), \Delta_{n,k})} V_n(x, t_n^{k+1})\right]} + d_{\boldsymbol{m}_N} \beta \log(\boldsymbol{m}_N).
\end{aligned}
$$

Proof. First, by Azuma-Bernstein inequality, with probability at least $1 - \delta$, we have

$$
\begin{aligned}
\sum_{n\in\widetilde{\mathcal{N}}} I_{0,n} &= \sum_{n=1}^{N} \mathbb{1}(n \in \widetilde{\mathcal{N}}) \left(\int_{t=0}^{T} b(x_n(t), u_n(t)) dt - \mathbb{E}_{x(\cdot)\sim p_n^*(x_{\text{ini}})}\left[\int_{t=0}^{T} b(x(t), u(t)) dt\right]\right) \\
&\lesssim \sqrt{\log(1/\delta) \sum_{n=1}^{N} \mathrm{Var}_{f^*,g^*}^{u_n} + \log(1/\delta)}.
\end{aligned}
$$

Next, by definition,

$$I_{1,n}^k := \mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}\left[V_n(x,t_n^{k+1})\right] - V_n(x_n(t_n^{k+1}),t_n^{k+1}).$$

Each $I_{1,n}^k$ is a zero-mean random variable whose variance is:

$$\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}\left[V_n(x,t_n^{k+1})\right] \le 1,$$

since $V_n \in [0,1]$.

We apply Bernstein's inequality for zero-mean, bounded (≤ 1) random variables. We have with probability at least $1-\delta$,

$$\begin{aligned}
\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1} I_{1,n}^k &= \sum_{n=1}^{N}\mathbb{1}(n\in\widetilde{\mathcal{N}})\sum_{k=0}^{m_n-1} I_{1,n}^k \\
&\lesssim \sqrt{\sum_{n=1}^{N}\mathbb{1}(n\in\widetilde{\mathcal{N}})\sum_{k=0}^{m_n-1}\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}\left[V_n(x,t_n^{k+1})\right]\cdot\log(1/\delta)} + \log(1/\delta). \\
&= \sqrt{\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}\left[V_n(x,t_n^{k+1})\right]\cdot\log(1/\delta)} + \log(1/\delta).
\end{aligned}$$

Next, we recall the definition:

$$I_{2,n}^k := \mathbb{E}_{x(t)\sim p_n^*(x_n(t_n^k))}\left[\int_0^{\Delta_{n,k}} b(x(t),u_n(t))dt\right] - \int_{t_n^k}^{t_n^{k+1}} b(x_n(t),u_n(t))dt.$$

Each $I_{2,n}^k$ is a martingale difference and satisfies $|I_{2,n}^k| \le 2\Delta_{n,k}$, since $b(x,u)\le 1$ by Assumption 5.1.

We apply the Azuma-Hoeffding inequality for bounded martingale differences. With probability at least $1-\delta$:

$$\begin{aligned}
\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1} I_{2,n}^k &= \sum_{n=1}^{N}\mathbb{1}(n\in\widetilde{\mathcal{N}})\sum_{k=0}^{m_n-1} I_{2,n}^k \\
&\lesssim \sqrt{\sum_{n=1}^{N}\mathbb{1}(n\in\widetilde{\mathcal{N}})\sum_{k=0}^{m_n-1}\Delta_{n,k}^2\log(1/\delta)} \\
&\le \sqrt{\sum_{n=1}^{N}\boldsymbol{\Delta}_n^2\log(1/\delta)}.
\end{aligned}$$

Finally, by Assumption 5.1, the stage-wise reward is bounded in $[0,1]$. Leveraging Lemma B.1, we can bound

$$|\mathbb{E}_{x\sim p_n(x_n(t_n^k),\Delta_{n,k})}\left[V_n(x,t_n^{k+1})\right] - \mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}\left[V_n(x,t_n^{k+1})\right]|$$

in terms of the corresponding variance and squared Hellinger distance:

$$\begin{aligned}
&\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}|I_{3,n}^k| \\
&= \sum_{n=1}^{N}\mathbb{1}(n\in\widetilde{\mathcal{N}})\sum_{k=0}^{m_n-1}|\mathbb{E}_{x\sim p_n(x_n(t_n^k),\Delta_{n,k})}\left[V_n(x,t_n^{k+1})\right] - \mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}\left[V_n(x,t_n^{k+1})\right]| \\
&\lesssim \sum_{n=1}^{N}\mathbb{1}(n\in\widetilde{\mathcal{N}})\sum_{k=0}^{m_n-1}\left[\sqrt{\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n(x,t_n^{k+1})\cdot\mathbb{H}^2\Big(p_n^*(x_n(t_n^k),\Delta_{n,k})\|p_n(x_n(t_n^k),\Delta_{n,k})\Big)}\right]
\end{aligned}$$

$$+\sum_{n=1}^{N}\mathbb{1}(n\in\widetilde{\mathcal{N}})\sum_{k=0}^{m_n-1}\left[\mathbb{H}^2\Big(p_n^*(x_n(t_n^k),\Delta_{n,k})\|p_n(x_n(t_n^k),\Delta_{n,k})\Big)\right]. \tag{B.4}$$

Applying the Cauchy–Schwarz inequality to equation B.4 yields

$$\begin{aligned}
&\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}|I_{3,n}^k|\\
&\le\sqrt{\sum_{n=1}^{N}\mathbb{1}(n\in\widetilde{\mathcal{N}})\sum_{k=0}^{m_n-1}\left[\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n(x,t_n^{k+1})\right]}\\
&\quad\cdot\sqrt{\sum_{n=1}^{N}\mathbb{1}(n\in\widetilde{\mathcal{N}})\sum_{k=0}^{m_n-1}\left[\mathbb{H}^2\Big(p_n^*(x_n(t_n^k),\Delta_{n,k})\|p_n(x_n(t_n^k),\Delta_{n,k})\Big)\right]}\\
&\quad+\sum_{n=1}^{N}\mathbb{1}(n\in\widetilde{\mathcal{N}})\sum_{k=0}^{m_n-1}\left[\mathbb{H}^2\Big(p_n^*(x_n(t_n^k),\Delta_{n,k})\|p_n(x_n(t_n^k),\Delta_{n,k})\Big)\right]\\
&\le\sqrt{d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N)\sum_{n=1}^{N}\mathbb{1}(n\in\widetilde{\mathcal{N}})\sum_{k=0}^{m_n-1}\left[\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n(x,t_n^{k+1})\right]}+d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N)\\
&=\sqrt{d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N)\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}\left[\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n(x,t_n^{k+1})\right]}+d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N),
\end{aligned}$$

where the final inequality follows directly from Lemma B.7. □

B.4 Lemmas to control total variance

The following lemma provides an upper bound on the cumulative variance of the value function estimates $V_n(x,t_n^{k+1})$ in terms of the variances of their reference optimal values $V_n^*(x,t_n^{k+1})$, an eluder-dimension-dependent complexity term, and an additional error term.

Lemma B.10. Let $\widetilde{\mathcal{N}}\subseteq[N]$ be the episode index set defined in Lemma B.9. Under the event $\mathcal{E}_{B.9}$, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n(x,t_n^{k+1})\\
&\lesssim\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n^*(x,t_n^{k+1})+d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N)+\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}|I_{4,n}^k|.
\end{aligned}$$

Proof. First we have

$$\begin{aligned}
&\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n(x,t_n^{k+1})\\
&\le 2\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n^*(x,t_n^{k+1})\\
&\quad+2\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}\widehat{V}_n(x,t_n^{k+1}),
\end{aligned}\tag{B.5}$$

where $\widehat{V}_n(x,t) := V_n(x,t) - V_n^*(x,t)$. Next we focus on bound the second term. We introduce a more general high order momentum $C_i, i = 0, \ldots, \log(\boldsymbol{m}_N)$, where

$$C_i := \sum_{n\in\widetilde{\mathcal{N}}} \sum_{k=0}^{m_n-1} \mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} \widehat{V}_n^{2^i}(x, t_n^{k+1}).$$

Then we have

$$\begin{aligned}
C_i &= \sum_{n\in\widetilde{\mathcal{N}}} \sum_{k=0}^{m_n-1} \mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} \widehat{V}_n^{2^{i+1}}(x, t_n^{k+1}) - [\mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} \widehat{V}_n^{2^i}(x, t_n^{k+1})]^2 \\
&= \sum_{n\in\widetilde{\mathcal{N}}} \sum_{k=0}^{m_n-1} \mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} \widehat{V}_n^{2^{i+1}}(x, t_n^{k+1}) - \widehat{V}_n^{2^{i+1}}(x_n(t_n^{k+1}), t_n^{k+1}) \\
&\quad - [\mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} \widehat{V}_n^{2^i}(x, t_n^{k+1})]^2 + \widehat{V}_n^{2^{i+1}}(x_n(t_n^{k+1}), t_n^{k+1}) \\
&\le \sum_{n\in\widetilde{\mathcal{N}}} \sum_{k=0}^{m_n-1} \underbrace{\mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} \widehat{V}_n^{2^{i+1}}(x, t_n^{k+1}) - \widehat{V}_n^{2^{i+1}}(x_n(t_n^{k+1}), t_n^{k+1})}_{J_{1,n,i}^k} \\
&\quad \underbrace{-[\mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} \widehat{V}_n^{2^i}(x, t_n^{k+1})]^2 + \widehat{V}_n^{2^{i+1}}(x_n(t_n^k), t_n^k)}_{J_{2,n,i}^k},
\end{aligned} \tag{B.6}$$

where the last line holds since we move the index one step earlier and we use the fact $\widehat{V}_n(x, t_n^m) = 0$.

Next, for $J_{1,n,i}^k$, by Azuma-Bernsetin inequality, we have with probability at least $1-\delta$ for all $i = 0, \ldots, \log(\boldsymbol{m}_N)$,

$$\begin{aligned}
&\sum_{n\in\widetilde{\mathcal{N}}} \sum_{k=0}^{m_n-1} J_{1,n,i}^k \\
&= \sum_{n=1}^{N} \mathbb{1}(n\in\widetilde{\mathcal{N}}) \sum_{k=0}^{m_n-1} J_{1,n,i}^k \\
&\lesssim \sqrt{\sum_{n\in\widetilde{\mathcal{N}}} \sum_{k=0}^{m_n-1} \mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} \widehat{V}_n^{2^{i+1}}(x, t_n^{k+1}) \log(\log(\boldsymbol{m}_N)/\delta)} + \log(\log(\boldsymbol{m}_N)/\delta) \\
&\le \sqrt{C_{i+1}\log(\log(\boldsymbol{m}_N)/\delta)} + \log(\log(\boldsymbol{m}_N)/\delta).
\end{aligned} \tag{B.7}$$

For $J_{2,n,i}^k$, we have

$$\begin{aligned}
&J_{2,n,i}^k \\
&= [\widehat{V}_n^{2^i}(x_n(t_n^k), t_n^k) - \mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} \widehat{V}_n^{2^i}(x, t_n^{k+1})][\widehat{V}_n^{2^i}(x_n(t_n^k), t_n^k) + \mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} \widehat{V}_n^{2^i}(x, t_n^{k+1})] \\
&\le [\widehat{V}_n^{2^i}(x_n(t_n^k), t_n^k) - [\mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} \widehat{V}_n^{2^{i-1}}(x, t_n^{k+1})]^2][\widehat{V}_n^{2^i}(x_n(t_n^k), t_n^k) + \mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} \widehat{V}_n^{2^i}(x, t_n^{k+1})] \\
&\le \prod_{j=0}^{i} [\widehat{V}_n^{2^i}(x_n(t_n^k), t_n^k) + \mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} \widehat{V}_n(x, t_n^{k+1})] \cdot |\widehat{V}_n(x_n(t_n^k), t_n^k) - \mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} \widehat{V}_n(x, t_n^{k+1})| \\
&\le 2^{i+1} |\widehat{V}_n(x_n(t_n^k), t_n^k) - \mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} \widehat{V}_n(x, t_n^{k+1})|,
\end{aligned} \tag{B.8}$$

where we use the fact that $\mathbb{E}X^2 \ge [\mathbb{E}X]^2$. Then taking summation of equation B.8 over $n \in \widetilde{\mathcal{N}}$ and k, we have

$$2^{-(i+1)} \cdot \sum_{n\in\widetilde{\mathcal{N}}} \sum_{k=0}^{m_n-1} J_{2,n,i}^k$$

$$
\begin{aligned}
&\leq \sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1} |\widehat{V}_n(x_n(t_n^k),t_n^k) - \mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}\widehat{V}_n(x,t_n^{k+1})| \\
&= \sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1} |V_n(x_n(t_n^k),t_n^k) - V_n^*(x_n(t_n^k),t_n^k) - \mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n(x,t_n^{k+1}) \\
&\quad + \mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n^*(x,t_n^{k+1})| \\
&= \sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1} \Bigg| \underbrace{\mathbb{E}_{x\sim p_n(x_n(t_n^k),\Delta_{n,k})}V_n(x,t_n^{k+1}) - \mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n(x,t_n^{k+1})}_{I_{3,n}^k} \\
&\quad + \underbrace{\mathbb{E}_{x(t)\sim p_n(x_n(t_n^k))}\left[\int_{t=0}^{\Delta_{n,k}} b(x(t),u_n(t))dt\right] - \mathbb{E}_{x(t)\sim p_n^*(x_n(t_n^k))}\left[\int_{t=0}^{\Delta_{n,k}} b(x(t),u_n(t))dt\right]}_{I_{4,n}^k}\Bigg| \\
&\leq \sqrt{d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N)\Bigg(\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}\left[\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n(x,t_n^{k+1})\right]\Bigg)} \\
&\quad + d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N) + \sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}|I_{4,n}^k|, \qquad \text{(B.9)}
\end{aligned}
$$

where the last inequality holds due to the upper bounds of $|I_{3,n}^k|$ obtained in Lemma B.9. Combining equation B.6, equation B.7 and equation B.9, we have $a < G/2$, $C_i \leq 2^i G + \sqrt{aC_{i+1}} + a$ and $C_i \leq H = \boldsymbol{m}_N$, where

$$
\begin{aligned}
G &:= \sqrt{d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N)\Bigg(\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}\left[\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n(x,t_n^{k+1})\right]\Bigg)} \\
&\quad + d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N) + \sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}|I_{4,n}^k|, \\
a &:= \log(\log(\boldsymbol{m}_N)/\delta).
\end{aligned}
$$

Therefore, by Lemma B.4, we have

$$
C_0 \lesssim G. \qquad \text{(B.10)}
$$

Finally, substituting equation B.10 back to equation B.5, we have

$$
\begin{aligned}
&\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n(x,t_n^{k+1}) \\
&\lesssim \sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n^*(x,t_n^{k+1}) + d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N) + \sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}|I_{4,n}^k| \\
&\quad + \sqrt{d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N)\Bigg(\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}\left[\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n(x,t_n^{k+1})\right]\Bigg)}.
\end{aligned}
$$

Using the fact that $x \lesssim \sqrt{ax} + b \Rightarrow x \lesssim a + b$, we obtain our final bound. □

The following lemma bounds the cumulative variance of the optimal value function V_n^* by the measurement gaps.

Lemma B.11. With probability at least $1-\delta$, for all $n \in [N]$, we have

$$
\sum_{k=0}^{m_n-1}\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n^*(x,t_n^{k+1}) \lesssim \log(N/\delta) + \sqrt{\log(N/\delta)\max_{1\leq n\leq N}\boldsymbol{\Delta}_n^2}. \qquad \text{(B.11)}
$$

Proof. Fix any $n \in [N]$. For simplicity, define $J_n := \sum_{k=0}^{m_n-1} \mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} V_n^*(x, t_n^{k+1})$. We begin by expanding the variance:

$$\begin{aligned} J_n &= \sum_{k=0}^{m_n-1} \Big[\mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} V_n^*(x,t_n^{k+1})^2 - \big(\mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} V_n^*(x,t_n^{k+1})\big)^2\Big] \\ &\le \sum_{k=0}^{m_n-1} \Big\{ \underbrace{\mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} V_n^*(x,t_n^{k+1})^2 - V_n^*(x_n(t_n^{k+1}),t_n^{k+1})^2}_{J_{1,n}^k} \\ &\quad + \underbrace{V_n^*(x_n(t_n^k),t_n^k)^2 - \big(\mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} V_n^*(x,t_n^{k+1})\big)^2}_{J_{2,n}^k} \Big\}, \end{aligned} \tag{B.12}$$

where the inequality uses the monotonicity $V_n^*(x_n(t_n^{k+1}),t_n^{k+1}) \le V_n^*(x_n(t_n^k),t_n^k)$.

By the Azuma–Bernstein inequality, with probability at least $1-\delta/N$,

$$\begin{aligned} \sum_{k=0}^{m_n-1} J_{1,n}^k &\lesssim \sqrt{\sum_{k=0}^{m_n-1} \mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} V_n^*(x,t_n^{k+1})^2 \cdot \log(N/\delta)} + \log(N/\delta) \\ &\le 2\sqrt{\sum_{k=0}^{m_n-1} \mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} V_n^*(x,t_n^{k+1}) \cdot \log(N/\delta)} + \log(N/\delta) \\ &= 2\sqrt{J_n \log(N/\delta)} + \log(N/\delta), \end{aligned} \tag{B.13}$$

where the second inequality follows from Lemma B.5.

For $J_{2,n}^k$, using equation 4.1 and the Markov property of Itô's SDE, we write

$$\begin{aligned} J_{2,n}^k &= \big[V_n^*(x_n(t_n^k),t_n^k) - \mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} V_n^*(x,t_n^{k+1})\big] \\ &\quad \cdot \big[V_n^*(x_n(t_n^k),t_n^k) + \mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} V_n^*(x,t_n^{k+1})\big] \\ &= \left[\mathbb{E}_{x(\cdot)\sim p_n^*(x_n(t_n^k))} \int_{t=0}^{\Delta_{n,k}} b(x(t),u(t))dt\right] \\ &\quad \cdot \big[V_n^*(x_n(t_n^k),t_n^k) + \mathbb{E}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} V_n^*(x,t_n^{k+1})\big] \\ &\lesssim \mathbb{E}_{x(\cdot)\sim p_n^*(x_n(t_n^k))} \int_{t=0}^{\Delta_{n,k}} b(x(t),u(t))dt, \end{aligned} \tag{B.14}$$

since $V_n^* \le 1$. Hence, with probability at least $1-\delta/N$, we have

$$\begin{aligned} \sum_{k=0}^{m_n-1} J_{2,n}^k &\lesssim \sum_{k=0}^{m_n-1} \left\{ \int_{t_n^k}^{t_n^{k+1}} b(x_n(t),u_n(t))dt \right. \\ &\qquad \left. + \left[\mathbb{E}_{x(\cdot)\sim p_n^*(x_n(t_n^k))} \int_0^{\Delta_{n,k}} b(x(t),u(t))dt - \int_{t_n^k}^{t_n^{k+1}} b(x_n(t),u_n(t))dt\right]\right\} \\ &\le 1 + \sum_{k=0}^{m_n-1} \left[\mathbb{E}_{x(\cdot)\sim p_n^*(x_n(t_n^k))} \int_0^{\Delta_{n,k}} b(x(t),u(t))dt - \int_{t_n^k}^{t_n^{k+1}} b(x_n(t),u_n(t))dt\right] \\ &\lesssim 1 + \sqrt{\boldsymbol{\Delta}_n^2 \log(N/\delta)}, \end{aligned} \tag{B.15}$$

where the second inequality follows from Assumption 5.1, and the third comes from Azuma-Hoeffding inequality using the bound $\int_{t_n^k}^{t_n^{k+1}} b(x(t),u(t))dt \le \Delta_{n,k}$.

Substituting equation B.13 and equation B.15 into equation B.12, and replacing each individual confidence level $1-\delta/N$ in equation B.13 and equation B.15 with $1-\delta/(2N)$ (which does not

affect the order of the bounds), we can apply a union bound to obtain an overall high-probability guarantee of $1-\delta$. Consequently, with probability at least $1-\delta$, for all $n \in [N]$,

$$J_n \lesssim \sqrt{J_n \log(N/\delta)} + 1 + \sqrt{\mathbf{\Delta}_n^2 \log(N/\delta)}$$
$$\Rightarrow J_n \lesssim \log(N/\delta) + \sqrt{\mathbf{\Delta}_n^2 \log(N/\delta)} \leq \log(N/\delta) + \sqrt{\log(N/\delta) \max_{1\leq n\leq N} \mathbf{\Delta}_n^2}.$$

□

The following lemma provides a global bound on the cumulative variance of the optimal value functions V_n^* over all episodes. It shows that this quantity is controlled by the total variance and measurement gaps.

Lemma B.12. Under event $\mathcal{E}_{B.11}$, with probability at least $1-\delta$, we have

$$\sum_{n=1}^{N} \sum_{k=0}^{m_n-1} \mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} V_n^*(x, t_n^{k+1}) \lesssim \log^2(N/\delta)\left(1 + \sum_{n=1}^{N} \mathrm{Var}^{u_n} + \sum_{n=1}^{N} \mathbf{\Delta}_n^2\right).$$

Proof. We define $J_n := \sum_{k=0}^{m_n-1} \mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})} V_n^*(x, t_n^{k+1})$ following Lemma B.11. Then by equation B.11 we have

$$J_n \lesssim \log(N/\delta) + \sqrt{\log(N/\delta) \max_{1\leq n\leq N} \mathbf{\Delta}_n^2}.$$

Next we prove that the conditional expectation of J_n can be bounded. First, following equation 4.1,

$$\begin{aligned} V_n^*(x,t) &= \mathbb{E}_{x(\cdot)\sim p_n^*(x)}\left[\int_t^T b(x(t),u(t))dt\right] \\ &= \mathbb{E}_{x(\cdot)\sim p_n^*(x)}\left[\int_t^{t+\Delta} b(x(t),u(t))dt\right] + \mathbb{E}_{x'\sim p_n^*(x,\Delta)} V_n^*(x', t+\Delta) \\ &= \mathbb{E}_{x(\cdot)\sim p_n^*(x)}\left[\int_0^{\Delta} b(x(t),u(t))dt\right] + \mathbb{E}_{x'\sim p_n^*(x,\Delta)} V_n^*(x', t+\Delta). \end{aligned} \tag{B.16}$$

Then, we have

$$\begin{aligned} \mathrm{Var}^{u_n} &= \mathbb{E}_{x(\cdot)\sim p_n^*(x_{\text{ini}})}\left[\sum_{k=0}^{m_n-1} \int_{t_n^k}^{t_n^{k+1}} b(x(t),u(t))dt - V_n^*(x_{\text{ini}},0)\right]^2 \\ &= \mathbb{E}_{x(\cdot)\sim p_n^*(x_{\text{ini}})}\Bigg[\sum_{k=0}^{m_n-1} \underbrace{\int_{t_n^k}^{t_n^{k+1}} b(x(t),u(t))dt + V_n^*(x_n(t_n^{k+1}),t_n^{k+1}) - V_n^*(x_n(t_n^k),t_n^k)}_{J_{n,k}}\Bigg]^2 \\ &= \mathbb{V}_{x(\cdot)\sim p_n^*(x_{\text{ini}})}\left[\sum_{k=0}^{m_n-1} J_{n,k}\right] \\ &= \sum_{k=0}^{m_n-1} \mathbb{V}_{x(\cdot)\sim p_n^*(x_{\text{ini}})}\left[J_{n,k}\right]. \end{aligned} \tag{B.17}$$

The first equality follows immediately from the definition of variance, and the second comes from equation B.16. Next, on each subinterval $[t_n^k, t_n^{k+1}]$ we introduce the *temporal increment* $J_{n,k}$, for which, by construction, $\mathbb{E}[J_{n,k}] = \mathbb{E}[J_{n,k}|x_n(t_n^k)] = 0$, yielding the third equality. Then, $\{J_{n,k}\}_{k=0}^{m_n-1}$ is a martingale-difference sequence with respect to the natural filtration $\mathcal{F}_k = \sigma\big(x_n(t_n^0), \ldots, x_n(t_n^k)\big)$, so orthogonality implies

$$\mathbb{V}_{x(\cdot)\sim p_n^*(x_{\text{ini}})}\left[\sum_{k=0}^{m_n-1} J_{n,k}\right] = \sum_{k=0}^{m_n-1} \mathbb{V}_{x(\cdot)\sim p_n^*(x_{\text{ini}})}[J_{n,k}].$$

Moreover, by the law of total variance, together with $\mathbb{E}[J_{n,k}|x_n(t_n^k)] = 0$, we have

$$\begin{aligned}
&\mathbb{V}_{x(\cdot)\sim p_n^*(x_{\text{ini}})}\left[J_{n,k}\right] \\
&= \mathbb{E}_{x\sim p_n^*(x_{\text{ini}},t_n^k)}\left[\mathbb{V}_{x(\cdot)\sim p_n^*(x)}\left[J_{n,k}|x\right]\right] + \mathbb{V}_{x(\cdot)\sim p_n^*(x)}\left[\mathbb{E}_{x\sim p_n^*(x_{\text{ini}},t_n^k)}\left[J_{n,k}|x\right]\right] \\
&= \mathbb{E}_{x\sim p_n^*(x_{\text{ini}},t_n^k)}\left[\mathbb{V}_{x(\cdot)\sim p_n^*(x)}\left[J_{n,k}|x\right]\right] \\
&= \mathbb{E}_{x\sim p_n^*(x_{\text{ini}},t_n^k)}\left[\mathbb{V}_{\substack{x(\cdot)\sim p_n^*(x),\\ x'\sim p_n^*(x,\Delta_{n,k})}}\left[\int_0^{\Delta_{n,k}} b(x(t),u(t))dt + V_n^*(x',t_n^{k+1})\right]\right]. \qquad \text{(B.18)}
\end{aligned}$$

Furthermore, by Assumption 5.1 we have $\int_0^{\Delta_{n,k}} b(x_n(t),u_n(t))dt \le \Delta_{n,k}$ for each $\Delta_{n,k}$. Thus,

$$\begin{aligned}
&\mathbb{E}_{x\sim p_n^*(x_{\text{ini}},t_n^k)}\left[\mathbb{V}_{\substack{x(\cdot)\sim p_n^*(x),\\ x'\sim p_n^*(x,\Delta_{n,k})}}\left[\int_0^{\Delta_{n,k}} b(x(t),u(t))dt + V_n^*(x',t_n^{k+1})\right]\right] \\
&\le 2\mathbb{E}_{x\sim p_n^*(x_{\text{ini}},t_n^k)}\left[\mathbb{V}_{x'\sim p_n^*(x,\Delta_{n,k})}\left[V_n^*(x',t_n^{k+1})\right]\right] \\
&\quad + 2\mathbb{E}_{x\sim p_n^*(x_{\text{ini}},t_n^k)}\left[\mathbb{V}_{x(\cdot)\sim p_n^*(x)}\left[\int_0^{\Delta_{n,k}} b(x(t),u(t))dt\right]\right] \qquad \text{(B.19)} \\
&\le 2\mathbb{E}_{x\sim p_n^*(x_{\text{ini}},t_n^k)}\left[\mathbb{V}_{x'\sim p_n^*(x,\Delta_{n,k})}\left[V_n^*(x',t_n^{k+1})\right]\right] + 2\Delta_{n,k}^2. \qquad \text{(B.20)}
\end{aligned}$$

Here, equation B.19 follows from the fact that $\mathrm{Var}(a+b) = \mathrm{Var}(a) + \mathrm{Var}(b) + 2\mathrm{Cov}(a,b) \le \mathrm{Var}(a) + \mathrm{Var}(b) + 2\sqrt{\mathrm{Var}(a)\cdot\mathrm{Var}(b)} \le 2\mathrm{Var}(a) + 2\mathrm{Var}(b)$. Summing equation B.20 over $k = 0,\dots,m_n-1$.

Thus we have, for each n,

$$\mathbb{E}[J_n|J_{n-1},\dots,J_1] \lesssim \mathrm{Var}^{u_n} + \boldsymbol{\Delta}_n^2. \qquad \text{(B.21)}$$

Applying Lemma B.3 to equation B.21 then yields

$$\begin{aligned}
\sum_{n=1}^N \min\{J_n, y\} &\lesssim y\log(1/\delta) + \log(1/\delta)\sum_{n=1}^N \mathbb{E}[J_n|J_{n-1},\dots,J_1] \\
&\lesssim y\log(1/\delta) + \log(1/\delta)\sum_{n=1}^N \mathrm{Var}^{u_n} + \log(1/\delta)\sum_{n=1}^N \boldsymbol{\Delta}_n^2. \qquad \text{(B.22)}
\end{aligned}$$

Finally, we plug y as the upper bound of J_n in equation B.11 in equation B.22, leading to

$$\begin{aligned}
\sum_{n=1}^N J_n &\lesssim \log^2(N/\delta)\left(1 + \sqrt{\max_{1\le n\le N}\boldsymbol{\Delta}_n^2} + \sum_{n=1}^N \mathrm{Var}^{u_n} + \sum_{n=1}^N \boldsymbol{\Delta}_n^2\right) \\
&\lesssim \log^2(N/\delta)\left(1 + \sum_{n=1}^N \mathrm{Var}^{u_n} + \sum_{n=1}^N \boldsymbol{\Delta}_n^2\right),
\end{aligned}$$

where for the second inequality we use the fact $\sqrt{x} \le 1 + x$, thus completing the proof. □

The following lemma gives the final high-probability upper bound on the cumulative regret in terms of decomposition results established in previous lemmas.

Lemma B.13. Let $\widetilde{\mathcal{N}} \subseteq [N]$ be the episode index set defined in Lemma B.9. Under events $\mathcal{E}_{B.6}, \mathcal{E}_{B.9}, \mathcal{E}_{B.10}, \mathcal{E}_{B.11}, \mathcal{E}_{B.12}$, we have

$$\begin{aligned}
\mathrm{Regret}(N) \lesssim \log(N/\delta)\Bigg(&\sqrt{d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N)\left(\sum_{n=1}^N \mathrm{Var}_{f^*,g^*}^{u_n} + \sum_{n=1}^N \boldsymbol{\Delta}_n^2\right)} \\
&+ N - |\widetilde{\mathcal{N}}| + d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N) + \sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}|I_{4,n}^k|\Bigg).
\end{aligned}$$

Proof. By Lemma B.6, we have $R_{f^*,g^*}(u_n) \le R_{f_n,g_n}(u_n)$. For any $n \in \widetilde{\mathcal{N}}$, by Lemma B.8, we have

$$\begin{aligned} R_{f^*,g^*}(u_n) - R_{f^*,g^*}(u_n) &\le R_{f_n,g_n}(u_n) - R_{f^*,g^*}(u_n) \\ &\le \min\left\{1, \sum_{k=0}^{m-1}(I_{1,n}^k + I_{2,n}^k + I_{3,n}^k + I_{4,n}^k) + I_{0,n}\right\}. \end{aligned}$$

Then we can bound the regret as

$$\text{Regret}(N) \lesssim N - |\widetilde{\mathcal{N}}| + \sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m-1}(I_{1,n}^k + I_{2,n}^k + I_{3,n}^k + I_{4,n}^k) + I_{0,n}. \tag{B.23}$$

From Lemma B.9, we have

$$\begin{aligned} &\sum_{n\in\widetilde{\mathcal{N}}}\left(I_{0,n} + \sum_{k=0}^{m-1} I_{1,n}^k + I_{2,n}^k + I_{3,n}^k + I_{4,n}^k\right) \\ &\lesssim d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N) + \sqrt{d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N)\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m-1}\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta)}\left[V_n(x,t_n^{k+1})\right]} \\ &\quad + \log(1/\delta)\left(\sqrt{\sum_{n=1}^{N}\text{Var}_{f^*,g^*}^{u_n}} + \sqrt{\sum_{n=1}^{N}\boldsymbol{\Delta}_n^2}\right) + \sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}|I_{4,n}^k|. \end{aligned} \tag{B.24}$$

From Lemma B.10, we have

$$\begin{aligned} &\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n(x,t_n^{k+1}) \\ &\lesssim \sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),\Delta_{n,k})}V_n^*(x,t_n^{k+1}) + d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N) + \sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}|I_{4,n}^k| \\ &\lesssim \log^2(N/\delta)\left(1 + \sum_{n=1}^{N}\text{Var}_{f^*,g^*}^{u_n} + \sum_{n=1}^{N}\boldsymbol{\Delta}_n^2\right) + d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N) + \sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}|I_{4,n}^k|, \end{aligned} \tag{B.25}$$

where the second inequality holds due to Lemma B.12. Substituting equation B.25 into equation B.24, then substituting them into equation B.23, we have our final regret bound. □

B.5 Proof of Theorem 5.10

We first have our concentration lemma.

Lemma B.14. With probability at least $1-\delta$, we have for all $n \in [N]$, $(f^*, g^*) \in \mathcal{P}_n$, and

$$\sum_{i=1}^{n-1}\sum_{k=0}^{m_i-1}\mathbb{H}^2(p_{f_n,g_n}(u_i, x_i(t_i^k), \widehat{\Delta}_{i,k})\|p_{f^*,g^*}(u_i, x_i(t_i^k), \widehat{\Delta}_{i,k})) \le \beta, \tag{B.26}$$

where $\beta = \log(2|\mathcal{F}||\mathcal{G}|N/\delta)$.

Proof. We apply Theorem E.4 in Wang et al. (2023) with $D_{i,k}$ being the delta distribution at $(u_i, x_i(t_i^k), \widehat{\Delta}_{i,k})$ guarantees $(f^*, g^*) \in \widehat{\mathcal{P}}_n$ and equation B.26 holds with probability at least $1-\delta/2$. Taking a union bound over the two events, we conclude that with probability at least $1-\delta$, both statements hold simultaneously. □

Next we have the following lemma.

Lemma B.15. Let the event $\mathcal{E}_{B.6}$ be the event of Lemma B.6. Then under event $\mathcal{E}_{B.6}$, there exists a set $\mathcal{N}_1 \subseteq [N]$ such that

- We have $|\mathcal{N}_1| \leq 13\log^2(4\beta \boldsymbol{m}_N) \cdot d_{8\beta \boldsymbol{m}_N}$.
- For any $n \in [N]$, $n \in \mathcal{N}_1$ is a stopping time.
- We have

$$\sum_{i\in[N]\setminus\mathcal{N}_1}\sum_{k=0}^{m_i-1}\mathbb{H}^2\left(p_{f_i,g_i}(u_i,x_i(t_i^k),\widehat{\Delta}_{i,k})\|p_{f^*,g^*}(u_i,x_i(t_i^k),\widehat{\Delta}_{i,k})\right) \leq 3d_{\boldsymbol{m}_N} + 7d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N).$$

Proof. We apply Lemma 6 in Wang et al. (2024b) here with the distribution class $p_{f,g}$, input space $\Pi \times \mathcal{X} \times [T]$ and function class Ψ. □

Next is our key lemma.

Lemma B.16. Let $\widetilde{\mathcal{N}} \subseteq [N]$ be an episode index set satisfying $\widetilde{\mathcal{N}} \subseteq [N] \setminus \mathcal{N}_1$. Under event $\mathcal{E}_{B.6}$, with probability at least $1-\delta$, the quantities $I_{4,n}^k$ introduced in introduced in Lemma B.8 satisfy

$$\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m-1}|I_{4,n}^k| \lesssim \sqrt{d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N)\sum_{n=1}^{N}\boldsymbol{\Delta}_n^2}.$$

Proof. Fix $n \in \widetilde{\mathcal{N}}$ and $0 \leq k < m_n$. We have

$$\begin{aligned}
&|I_{4,n}^k| \\
&= \left|\mathbb{E}_{x(\cdot)\sim p_n(x_n(t_n^k))}\left[\int_{t=0}^{\Delta_{n,k}} b(x(t),u(t))dt\right] - \mathbb{E}_{x(\cdot)\sim p_n^*(x_n(t_n^k))}\left[\int_{t=0}^{\Delta_{n,k}} b(x(t),u(t))dt\right]\right| \\
&\leq \int_{t=0}^{\Delta_{n,k}}\left|\mathbb{E}_{x\sim p_n^*(x_n(t_n^k),t)}b(x,u) - \mathbb{E}_{x\sim p_n(x_n(t_n^k),t)}b(x,u)\right|dt \\
&\lesssim \int_{t=0}^{\Delta_{n,k}}\sqrt{\mathbb{V}_{x\sim p_n^*(x_n(t_n^k),t)}b(x,u)\mathbb{H}^2(p_n^*(x_n(t_n^k),t)\|p_n(x_n(t_n^k),t))} \\
&\quad + \mathbb{H}^2(p_n^*(x_n(t_n^k),t)\|p_n(x_n(t_n^k),t))dt \\
&\lesssim \int_{t=0}^{\Delta_{n,k}}\mathbb{H}(p_n^*(x_n(t_n^k),t)\|p_n(x_n(t_n^k),t))dt \\
&= \underbrace{\Delta_{n,k}\cdot\mathbb{H}(p_n^*(x_n(t_n^k),\widehat{\Delta}_{n,k})\|p_n(x_n(t_n^k),\widehat{\Delta}_{n,k}))}_{J_{1,n}^k} \\
&\quad + \underbrace{\int_{t=0}^{\Delta_{n,k}}\mathbb{H}(p_n^*(x_n(t_n^k),t)\|p_n(x_n(t_n^k),t))dt - \Delta_{n,k}\mathbb{H}(p_n^*(x_n(t_n^k),\widehat{\Delta}_{n,k})\|p_n(x_n(t_n^k),\widehat{\Delta}_{n,k}))}_{J_{2,n}^k},
\end{aligned}$$

where we use the fact that $b \leq 1$ and $\mathbb{H} \leq 1$. For $J_{1,n}^k$, we have:

$$\begin{aligned}
\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}J_{1,n}^k &\leq \sqrt{\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}\Delta_{n,k}^2}\cdot\sqrt{\sum_{n\in\widetilde{\mathcal{N}}}\sum_{k=0}^{m_n-1}\mathbb{H}^2(p_n^*(x_n(t_n^k),\widehat{\Delta}_{n,k})\|p_n(x_n(t_n^k),\widehat{\Delta}_{n,k}))} \\
&\leq \sqrt{d_{\boldsymbol{m}_N}\beta\log(\boldsymbol{m}_N)\sum_{n=1}^{N}\boldsymbol{\Delta}_n^2}, \qquad \text{(B.27)}
\end{aligned}$$

where the first inequality is by Cauchy-Schrawz inequality and the last one holds due to Lemma B.7.

Algorithm 3 Lagrangian CT-MLE

Require: Episode number N, policy class Π, initial state x_{ini}, drift class $\mathcal{F}$, diffusion class $\mathcal{G}$, reward function b, planning horizon T, parameter η.

1: For each $n \in [N]$, determine a fixed measurement time sequence $0 = t_n^0 < \cdots < t_n^{m_n} = T$. For any $0 \le k < m_n$, denote measurement gaps $\Delta_{n,k} := t_n^{k+1} - t_n^k$, randomized measurement gap $\widehat{\Delta}_{n,k} \sim \text{Unif}(0, \Delta_{n,k})$.

2: **for** episode $n = 1, \ldots, N$ **do**

3: Solve (f_n, g_n) via

$$f_n, g_n = \underset{(f,g)\in\mathcal{F}\times\mathcal{G}}{\arg\max} \Bigg\{ R_{f,g}(u) + \eta_n \cdot \Bigg(\sum_{i=1}^{n-1} \sum_{k=0}^{m_i-1} \log p_{f,g}(x_i(t_i^{k+1})|u_i, x_i(t_i^k), \Delta_{i,k}) + \sum_{i=1}^{n-1} \sum_{k=0}^{m_i-1} \log p_{f,g}(x_i(t_i^k + \widehat{\Delta}_{i,k})|u_i, x_i(t_i^k), \widehat{\Delta}_{i,k}) \Bigg) \Bigg\},$$

4: Set policy u_n as $u_n = \arg\max_{u\in\Pi} R_{f_n,g_n}(u)$.

5: Execute the n-th episode and observe $x_n(t_n^0), x_n(t_n^0 + \widehat{\Delta}_{n,0}) \ldots, x_n(t_n^{m_n-1} + \widehat{\Delta}_{n,m_n-1}), x_n(t_n^{m_n})$.

6: **end for**

7: **return** Randomly pick an $n \in [N]$ uniformly and output $\widehat{u}$ as u_n.

For $\{J_{2,n}^k\}_{n,k}$, because $\widehat{\Delta}_{n,k}$ is sampled uniformly from $[0, \Delta_{n,k}]$, the sequence $\{J_{2,n}^k\}_{n,k}$ forms a martingale difference sequence (MDS). Noting $|J_{2,n}^k| \le 2\Delta_n^k$ we can apply Azuma-Hoeffding inequality to $J_{2,n}^k$, which infers that with probability at least $1-\delta$,

$$\sum_{n\in\widetilde{\mathcal{N}}} \sum_{k=0}^{m_n-1} J_{2,n}^k = \sum_{n=1}^{N} \mathbb{1}(n \in \widetilde{\mathcal{N}}) \sum_{k=0}^{m_n-1} J_{2,n}^k \lesssim \sqrt{\sum_{n\in\widetilde{\mathcal{N}}} \sum_{k=0}^{m_n-1} \Delta_{n,k}^2 \log(1/\delta)} \le \sqrt{\sum_{n=1}^{N} \boldsymbol{\Delta}_n^2 \log(1/\delta)}. \tag{B.28}$$

Therefore, from equation B.27 and equation B.28, we obtain our bound. □

Then we have our final proof of Theorem 5.10.

Proof of Theorem 5.10. We set $\widetilde{\mathcal{N}} = [N] \setminus (\mathcal{N} \cup \mathcal{N}_1)$. Since both $n \in \mathcal{N}, n \in \mathcal{N}_1$ are stopping time, then $\widetilde{\mathcal{N}}$ is also a stopping time. Clearly we have $\widetilde{\mathcal{N}} \subseteq [N] \setminus \mathcal{N}$ and $\widetilde{\mathcal{N}} \subseteq [N] \setminus \mathcal{N}_1$, thus we can apply both Lemma B.13 and B.16. Then substituting the bound of $\sum_{n\in\widetilde{\mathcal{N}}} \sum_{k=0}^{m_n-1} |I_{4,n}^k|$ from Lemma B.16 into Lemma B.13 and using the fact that

$$N - |\widetilde{\mathcal{N}}| \le |\mathcal{N}| + |\mathcal{N}_1| \le 26 \log^2(4\beta \boldsymbol{m}_N) \cdot d_{8\beta\boldsymbol{m}_N}$$

concludes our proof. Here, the second inequality holds due to the bounds of $|\mathcal{N}|$ in Lemma B.6 and $|\mathcal{N}_1|$ in Lemma B.14. □

C Numerical Experiments

Algorithm 1 (CT-MLE) is theoretically clean and analysis-friendly, but its direct use is computationally prohibitive. The core difficulty is that it optimizes a reward $R_{f,g}(u)$ over parameters (f, g) subject to *two* confidence constraints, i.e., membership in the intersection $\mathcal{P}_n \cap \widehat{\mathcal{P}}_n$. This yields a constrained program with set intersections defined by likelihood inequalities, which is generally intractable at scale.

Let

$$\mathcal{L}_{f,g}^{(n)} := \sum_{i=1}^{n-1} \sum_{k=0}^{m_i-1} \log p_{f,g}(x_i(t_i^k + \Delta_{i,k})|u_i, x_i(t_i^k), \Delta_{i,k}) \tag{C.1}$$

$$\widehat{\mathcal{L}}_{f,g}^{(n)} := \sum_{i=1}^{n-1} \sum_{k=0}^{m_i-1} \log p_{f,g}(x_i(t_i^k + \widehat{\Delta}_{i,k}) | u_i, x_i(t_i^k), \widehat{\Delta}_{i,k}). \tag{C.2}$$

The CT-MLE solves

$$\max_{(f,g)\in\mathcal{F}\times\mathcal{G}} \quad R_{f,g}(u) \tag{C.3}$$

$$\text{s.t.} \quad \mathcal{L}_{f,g}^{(n)} \geq \max_{(f',g')\in\mathcal{F}\times\mathcal{G}} \mathcal{L}_{f',g'}^{(n)} - \beta, \tag{C.4}$$

$$\widehat{\mathcal{L}}_{f,g}^{(n)} \geq \max_{(f',g')\in\mathcal{F}\times\mathcal{G}} \widehat{\mathcal{L}}_{f',g'}^{(n)} - \beta, \tag{C.5}$$

i.e., (f, g) must lie in the β-near-optimal regions of both likelihoods.

To make the problem implementable, we replace the hard constraints by penalties via standard Lagrangian relaxation. Introducing multipliers $\eta_n, \widehat{\eta}_n \geq 0$, we obtain the unconstrained surrogate

$$\max_{(f,g)\in\mathcal{F}\times\mathcal{G}} \Big\{ R_{f,g}(u) + \eta_n \big(\mathcal{L}_{f,g}^{(n)} - \max \mathcal{L}^{(n)} + \beta\big) + \widehat{\eta}_n \big(\widehat{\mathcal{L}}_{f,g}^{(n)} - \max \widehat{\mathcal{L}}^{(n)} + \beta\big) \Big\}. \tag{C.6}$$

Since $\max \mathcal{L}^{(n)}$, $\max \widehat{\mathcal{L}}^{(n)}$, and β are constants with respect to (f, g), they do not affect the maximizer and can be dropped. For simplicity we tie the multipliers, $\eta_n = \widehat{\eta}_n$, yielding the implementation-friendly objective used in Algorithm 3:

$$\max_{(f,g)\in\mathcal{F}\times\mathcal{G}} \Big\{ R_{f,g}(u) \; + \; \eta_n \big(\mathcal{L}_{f,g}^{(n)} + \widehat{\mathcal{L}}_{f,g}^{(n)}\big) \Big\}. \tag{C.7}$$

The coefficient η_n governs the trade-off between the task reward and adherence to high-likelihood regions defined by both data fidelities ($\mathcal{L}^{(n)}$ and $\widehat{\mathcal{L}}^{(n)}$). In effect, the relaxation converts the intractable set intersection into a soft regularizer that is straightforward to optimize with standard gradient-based methods over parameterized (f, g). This surrogate serves as the entry point to our experiments, enabling a scalable approximation to CT-MLE while preserving the original constraints.

C.1 Implementation Details

We address several practical implementation challenges for Algorithm 3. The primary challenge is computing the conditional probability density function $p_{f,g}(x_i(t_i^{k+1}) \mid u_i, x_i(t_i^k), \Delta_{i,k})$, where $\Delta_{i,k} = t_i^{k+1} - t_i^k$. Since direct maximization of the conditional log-likelihood is infeasible due to the unknown normalizing constant of the SDE transition density, we employ continuous-time score matching (Hyvärinen & Dayan, 2005). This approach eliminates the intractable normalization term by minimizing the Fisher divergence between the model score and the data score, providing a tractable and computationally efficient surrogate for MLE (Pabbaraju et al., 2023). Following Song et al. (2020), we adopt the sliced formulation to obtain unbiased and computationally efficient estimators for the drift and diffusion parameters (f_θ, g_θ) used in Algorithm 3.

The second challenge involves determining the optimal policy u_n given the estimated drift f and diffusion g terms. Using the learned SDE, we generate model rollouts and implement a continuous-time actor-critic update: the critic V_ξ minimizes the mean-squared temporal difference error, while the actor u_ϕ maximizes discounted n-step returns through stochastic gradient ascent. Our implementation follows deterministic policy gradients (Silver et al., 2014; Lillicrap et al., 2015) but obtains exact gradients by backpropagating through the ODE, similar to neural ODE policy evaluation in continuous time (Chen et al., 2018; Yildiz et al., 2021).

We build upon the continuous-time model-based RL framework of Yildiz et al. (2021), augmenting it with additive Gaussian noise to formulate the environment dynamics as an SDE rather than an ODE. Crucially, we replace the original dynamics learning objective with a continuous-time sliced score matching (SSM) loss (Song et al., 2020). Over each of the N_{dyn} gradient updates, we perform the following steps to minimize the model loss:

$$\mathcal{L}(\theta) = \mathcal{J}_{\text{SSM}}(\theta) - \eta' \mathbb{E}[V_{f_\theta, g_\theta}^{u_\psi}(x)],$$

where $\mathcal{J}_{\text{SSM}}$ is the sliced score-matching objective, the second term biases model learning toward higher policy value, and $\eta' = \frac{1}{\eta_n \kappa}$ with $\kappa > 0$ as scale factor aligning the numerical scales of the SSM loss and the (negative) planning objective.

1. **Data Sampling:** Draw a batch of B_{dyn} subsequences of length H_{dyn} from the training dataset $\mathcal{D}$:
$$\big\{\big(x_i(t_0), u_i(t_0)\big), \ldots, \big(x_i(t_{H_{\text{dyn}}}), u_i(t_{H_{\text{dyn}}})\big)\big\}_{i=1}^{B_{\text{dyn}}} \sim \mathcal{D},$$
where $x_i(t_k)$ denotes the state at measurement time t_k and $u_i(t_k) = u\big(x_i(t_k)\big)$ is the corresponding control input under policy u.
2. **Score Matching Computation:** For each subsequence i and time step $k \in \{0, \ldots, H_{\text{dyn}} - 1\}$:
   (a) Compute the interval length: $\Delta t_i^k = t_i^{k+1} - t_i^k$.
   (b) Compute the conditional mean via ODE integration:
   $$\mu_\theta^{(i,k)} = \text{ODEInt}\big(f_\theta(\cdot, u_i(t_k)), x_i(t_k), [0, \Delta t_i^k]\big),$$
   where $\text{ODEInt}(\cdot)$ denotes a numerical ODE solver (we use the Dormand-Prince RK45 integrator), $f_\theta(\cdot, u_i(t_k))$ is the learned drift network with control input $u_i(t_k)$, and $[0, \Delta t_i^k]$ is the integration interval.
   (c) Evaluate the interval covariance:
   $$\Sigma_\theta^{(i,k)} = \big(g_\theta(x_i(t_k), u_i(t_k))\big)^2 \Delta t_i^k,$$
   where we square the instantaneous noise scale element-wise and multiply by the interval length to obtain the diagonal covariance matrix.
   (d) Compute the model score at the interval endpoint $x_i(t_{k+1})$:
   $$s_\theta^{(i,k)} = -\big(\Sigma_\theta^{(i,k)}\big)^{-1}\big(x_i(t_{k+1}) - \mu_\theta^{(i,k)}\big).$$
   (e) Estimate the sliced score matching loss using M_{proj} random projections. For each Rademacher vector $v_{i,k,m} \in \{\pm 1\}^d$, compute:
   $$\ell_{i,k,m} = \frac{1}{2}\|s_\theta^{(i,k)}\|^2 + v_{i,k,m}^\top \nabla_x \big[v_{i,k,m}^\top s_\theta^{(i,k)}\big]\Big|_{x=x_i(t_{k+1})}, \quad m = 1, \ldots, M_{\text{proj}}.$$
   This provides an unbiased Monte Carlo estimate of the sliced score matching loss, combining the score energy term with its directional derivative.
   (f) Aggregate the batched sliced score matching loss:
   $$\mathcal{J}_{\text{SSM}}(\theta) = \frac{1}{B_{\text{dyn}} \cdot H_{\text{dyn}} \cdot M_{\text{proj}}} \sum_{i=1}^{B_{\text{dyn}}} \sum_{k=0}^{H_{\text{dyn}}-1} \sum_{m=1}^{M_{\text{proj}}} \ell_{i,k,m}.$$
3. **Planning Loss Computation:**
   (a) Estimate the advantage $A(x_i(t_k), u_i(t_k))$ for each state-action pair in the batch using the current critic networks:
   $$\widehat{A}_i = r_i(t_k) + \gamma V_\psi'(x_i(t_{k+1})) - Q_\psi(x_i(t_k), u_i(t_k)),$$
   where Q_ψ is the critic network and V_ψ' is the target value function.
   (b) Compute the gradient of the log-transition probability with respect to the model parameters. For a Gaussian transition model parameterized by $(\mu_\theta, \Sigma_\theta)$:
   $$\nabla_\theta \log P_\theta(x_{k+1}|x_k, u_k) = \nabla_\theta \left[-\frac{1}{2}\log|\Sigma_\theta| - \frac{1}{2}(x_{k+1} - \mu_\theta)^\top \Sigma_\theta^{-1}(x_{k+1} - \mu_\theta)\right].$$
   This gradient is computed efficiently using automatic differentiation on the terms calculated in Step 2(b).
   (c) Form the Monte Carlo estimate of the planning gradient:
   $$\nabla_\theta \mathbb{E}[V] \approx \frac{1}{B_{\text{dyn}}} \sum_{i=1}^{B_{\text{dyn}}} \widehat{A}_i \cdot \nabla_\theta \log P_\theta(x_i(t_{k+1})|x_i(t_k), u_i(t_k)).$$
4. **Combined Model Update:** Update the model parameters via gradient descent:
$$\theta \leftarrow \theta - \alpha_{\text{model}}\big(\nabla_\theta \mathcal{J}_{\text{SSM}} - \eta' \nabla_\theta \mathbb{E}[V]\big),$$
using the AdamW optimizer (Kingma, 2014; Loshchilov & Hutter, 2017).

C.2 MAIN RESULTS.

We evaluate Algorithm 3 on three classic control tasks from the Gymnasium benchmark (Brockman et al., 2016; Towers et al., 2024), comparing against two state-of-the-art continuous-time baselines: ENODE (Yildiz et al., 2021) and SAC-TaCoS (Treven et al., 2024b).

Tasks. We consider three environments of increasing difficulty:

- **Pendulum (Easiest):** The inverted pendulum swing-up problem is a fundamental challenge in control theory. The system consists of a pendulum attached at one end to a fixed pivot, with the other end free to move. Starting from a hanging-down position, the goal is to apply torque to swing the pendulum into an upright position, aligning its center of gravity directly above the pivot. The control space represents the torque applied to the free end, while the state space includes the pendulum's x-y coordinates and angular velocity. This environment is considered the simplest due to its continuous control space and relatively straightforward dynamics with a single degree of freedom.
- **CartPole (Medium Difficulty):** The CartPole system comprises a pole attached via an unactuated joint to a cart that moves along a frictionless track. Initially, the pole is in an upright position, and the objective is to maintain balance by applying forces to the cart in either the left or right direction. The control space determines the direction of the fixed force applied to the cart, while the state space includes the cart's position and velocity, as well as the pole's angle and angular velocity. This environment presents moderate difficulty due to its discrete action space and the need to balance an inherently unstable system with coupled dynamics.
- **Acrobot (Most Difficult):** The Acrobot system consists of two links connected in series, forming a chain with one end fixed. The joint between the two links is actuated, and the goal is to apply torques to this joint to swing the free end above a target height, starting from the initial hanging-down state. We use the fully actuated version of the Acrobot environment, as no method has successfully solved the underactuated balancing problem, consistent with Zhong & Leonard (2020). The control space is discrete and deterministic, representing the torque applied to the actuated joint, while the state space consists of the two rotational joint angles and their angular velocities. This environment is the most challenging due to its complex nonlinear dynamics involving two coupled pendulums, requiring sophisticated control strategies to coordinate the motion of both links.

Baselines. **ENODE** learns dynamics using Bayesian neural ODEs and optimizes a theoretically consistent continuous-time actor-critic, providing uncertainty-aware control without time discretization. However, it was not specifically designed for stochastic environments. **SAC-TaCoS** reformulates the continuous-time SDE control problem as an equivalent discrete-time extended MDP, where policies output both actions and their duration. This enables time-adaptive control using standard algorithms like SAC.

Experimental Setup. We inject Gaussian noise $\mathcal{N}(0, \sigma^2 \cdot I)$ at every time step following Treven et al. (2024b), with $\sigma = 2.0$ across all experiments. The noise perturbs all state components (e.g., angle θ and angular velocity $\dot{\theta}$), transforming these originally deterministic systems into stochastic environments. We evaluate performance after 5, 15, and 15 training episodes for Pendulum, CartPole, and Acrobot, respectively, following standard evaluation protocols.

ENODE uses equidistant time intervals as specified in Yildiz et al. (2021), while SAC-TaCoS adaptively determines intervals following Treven et al. (2024b). For simplicity, our method also employs equidistant intervals. For our method, we apply annealed Lagrange multipliers $\eta_n = \eta_{\text{base}}/n$ with $\eta_{\text{base}} = 4$, and adaptive scaling $\kappa_n \propto$ SSM scale/planning scale to maintain a 10:1 SSM-to-planning ratio for training stability.

Results. Figure 2 presents our main findings. Our CT-MLE algorithm achieves superior asymptotic performance across all three environments, demonstrating effective adaptation to stochastic dynamics. While SAC-TaCoS exhibits faster initial convergence and lower variance, our method ultimately achieves higher cumulative rewards after sufficient training.

ENODE shows consistently poor performance across all tasks, with minimal learning progress even after extended training. This degradation is expected given that ENODE was not designed for

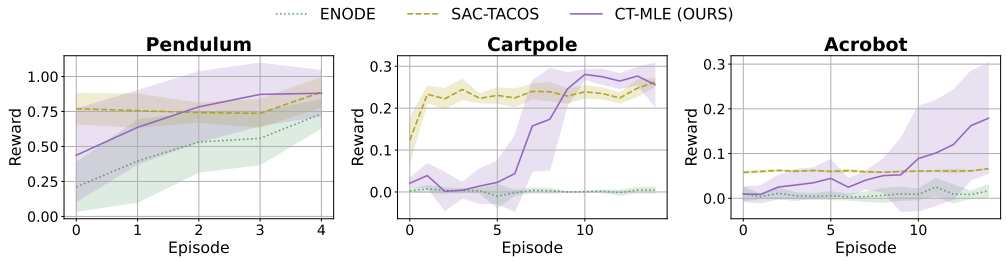

Figure 2: Performance comparison of Algorithm 3, ENODE (Yildiz et al., 2021), and SAC-TaCoS (Treven et al., 2024b) across three environments with noise $\sigma = 2.0$ (± 1 standard error).

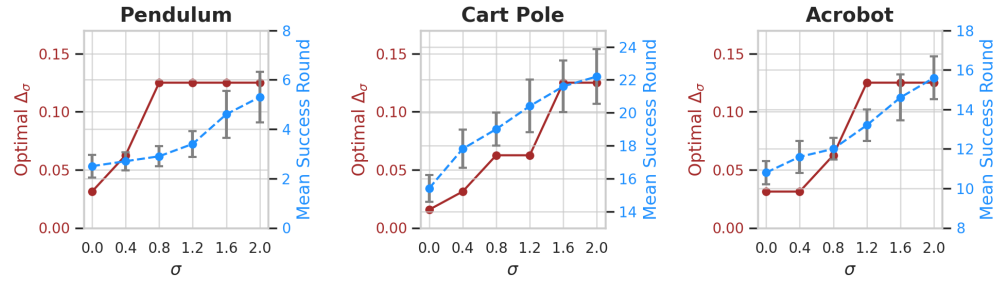

Figure 3: Optimal measurement gap Δ_σ and mean episodes to success (± 1 standard error) under varying environment stochasticity σ. Results averaged over 10 random seeds for Pendulum and 5 seeds for Cart Pole and Acrobot environments.

stochastic environments. The failure is evident in CartPole and Acrobot, where ENODE achieves no meaningful reward improvement.

The performance advantage of our method increases with task complexity. In Acrobot, the most challenging environment with complex nonlinear dynamics, the gap between our approach and the baselines is most pronounced. This suggests that our algorithm's ability to model and adapt to noisy dynamics becomes increasingly valuable as learning difficulty increases, making it particularly well-suited for complex stochastic control problems.

C.3 Ablation Study

Validation of Theoretical Claims. We validate our theoretical claims through systematic numerical experiments. Following Yildiz et al. (2021), we define task success as achieving rewards of at least 0.9 after a warm-up period ($T_{\text{warm-up}} = 3$ seconds) within a planning horizon of $T = 10$ seconds. We set η_n to a large value as it does not influence optimal gap selection. For each volatility level $\sigma \in \{0, 0.4, \ldots, 2.0\}$, we evaluate CT-MLE with equidistant measurement gaps $\Delta = 2^{-i}$ for $i \in \{0, 1, \ldots, 7\}$. The optimal gap Δ_σ is defined as the largest gap achieving the minimum number of episodes to success.

Figure 3 demonstrates that the optimal measurement gap Δ_σ increases monotonically with environment volatility σ, directly validating our theoretical analysis. This empirical observation confirms Remark 5.14, which establishes that the optimal gap for minimizing episode complexity scales proportionally with the total variance: $\Delta \propto \text{Var}^{\Pi}$. Higher volatility induces larger variance, necessitating wider measurement gaps for optimal performance. Notably, in low-stochasticity regimes ($\sigma \in \{0, 0.4\}$), the optimal gap is not the finest resolution tested (2^{-7}), confirming our theoretical prediction that excessive measurement precision yields diminishing returns.

The results further validate our algorithm's instance-dependent complexity guarantees. As shown in Figure 3, the number of episodes required for success increases with σ, confirming that our algorithm correctly identifies harder instances (higher σ) and adaptively allocates more samples. This behavior

aligns with our theoretical framework, where episode complexity directly reflects the total interaction data required for convergence.

Numerical Convergence Rate. We also report the reward error (mean ± 1 standard error over 10 seeds) across training episodes for the Pendulum environment with $\sigma = 2.0$, using the corresponding optimal gap $\Delta_\sigma = 0.125$, as shown in Figure 4. The reward error decreases with the number of episodes N, and the decay trend closely follows an approximate $1/\sqrt{N}$ convergence rate, which aligns well with our theoretical predictions.

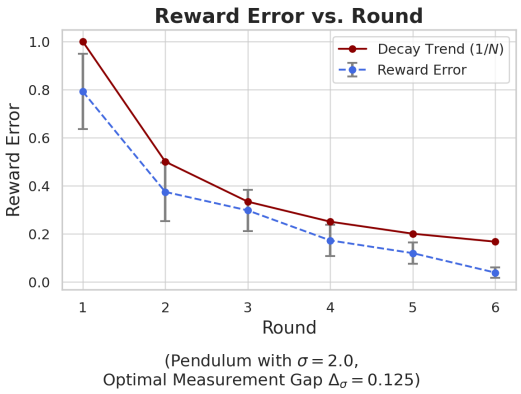

Figure 4: Reward error convergence follows the theoretical $1/\sqrt{N}$ rate (Pendulum, $\sigma = 2.0$).

C.4 Additional Details

All experiments were conducted on a single NVIDIA A6000 GPU. Each 15-episode Pendulum swing-up task required approximately 5 hours to complete; each 30-episode Cart Pole task required approximately 15 hours to complete; and each 25-episode Acrobot task required approximately 12 hours to complete. The peak GPU memory utilization per run ranges from 4GB to 20GB approximately. We listed all the key hyper-parameters used in the numerical experiment in Table 1.

Table 1: Hyper-parameters in numerical experiment

Hyperparameter	Default	Description
η_{base}	4	Base value for Lagrangian Multiplier
N_0	3	Number of trajectories at observation time points in initial data set
H	50	Trajectory length (in seconds) in the data set
N_{inc}	1	Number of trajectories at observation time points added to the data set after each episode
B_{dyn}	5	Batch size of the dynamic learning
N_{dyn}	500	Number of dynamic learning update iterations in each episode
H_{dyn}	5	Length of each subsequence (horizon) in dynamic learning
M_{proj}	1	Rademacher projections per sample in dynamic learning
N_{pol}	250	Number of policy learning update iterations in each episode
H_{pol}	5	Length of each subsequence (horizon) in policy learning
T	10	Length of each test trajectory at the end of every episode
$T_{\text{warm-up}}$	3	The warm-up subsequence of each test trajectory that does not collect rewards and evaluate at observation time points
N_{test}	10	Number of test trajectories at the end of every episode