EXPRESSIVITY AND STATISTICAL TRADE-OFFS IN DIFFUSION POLICY LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based policies have recently emerged as powerful policy parameterizations for reinforcement learning, representing state-conditioned action distributions as terminal laws of diffusion processes with parameterized drifts. This terminal-law representation has shown substantial expressive flexibility in practice, enabling diffusion policies to model complex, multimodal, and highly non-Gaussian action distributions; however, it remains unclear what mathematically drives this expressivity and how to fully exploit it when the policy is learned from finite data. In this paper, we identify the drift Lipschitz budget K as a central quantity governing the expressivity and statistical behavior of diffusion policies. We quantify expressivity through approximation: diffusion policies with K-Lipschitz drifts can concentrate near optimal deterministic policies and achieve value approximation error of order $1/K$; moreover, we prove a matching lower bound under nondegenerate diffusion noise. This increased expressivity comes with a statistical cost. When the drift is parameterized by neural networks, increasing K improves approximation but increases statistical complexity. Balancing these two terms yields a finite-sample performance gap of order $\tilde{\mathcal{O}}(n^{-2/(m+6)})$ for generic neural-network drifts, and a sharper rate $\tilde{\mathcal{O}}(n^{-2/(m+4)})$ for one-sided dissipative drift classes, where n is the sample size and m is the dimension of the state space. We also discuss how additional smoothness and low-dimensional structure of the optimal policy can further sharpen these rates. Numerical experiments provide empirical evidence for the sample-dependent trade-off in K, supporting both theoretical regimes. Our framework also suggests a practical implementation principle: choose the diffusion budget K according to the available sample size, and then select a neural-network architecture with the corresponding fixed Lipschitz coefficient.

1 INTRODUCTION

Many reinforcement learning (RL) problems concern learning a stochastic policy that optimizes an expected objective through repeated interaction with the environment (Sutton et al., 1998). Policy-gradient and its variants are among the most widely used approaches for this task (Haarnoja et al., 2018; Kakade, 2001; Schulman et al., 2015; 2017; Sutton et al., 1999). Their performance, however, depends critically on the choice of policy parameterization. An effective policy class must satisfy three requirements: (i) it should be expressive enough to approximate good, or even optimal, policies; (ii) it should allow efficient sampling of actions during interaction with the environment; and (iii) it should support tractable policy-gradient estimation (e.g., Sutton et al., 1999).

Designing such policy classes becomes especially challenging beyond finite action spaces. In continuous-action RL, classical methods often rely on tractable stochastic policy families, such as Gaussian, Beta, Gaussian-mixture, or exponential-family policies, because these distributions are easy to sample from and have tractable likelihoods (Haarnoja et al., 2018; Hansen-Estruch et al., 2023; Schulman et al., 2015; 2017). This computational convenience, however, comes with limited expressivity. Gaussian policies may be poorly matched to bounded or constrained action spaces, log-concave families cannot represent many multimodal action distributions, and mixture models can introduce numerical and optimization difficulties. As a result, these simple policy classes can struggle to represent the complex, multimodal, or strongly non-Gaussian action distributions that

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

arise in modern RL tasks, such as multi-stage manipulation (Chi et al., 2025; Ke et al., 2024; Scheikl et al., 2024) and locomotion (Huang et al., 2024).

Diffusion-based policies provide a more expressive alternative by representing state-conditioned action distributions as terminal laws of diffusion processes with parameterized drifts. The key idea is to transform a simple reference distribution into a complex action law through stochastic dynamics. This approach has shown strong expressive power in generative modeling, including image and video generation (Ho et al., 2020; Sohl-Dickstein et al., 2015; Song & Ermon, 2019). Building on these advances, recent works have used diffusion models as policy parameterizations in control and RL (Balim et al., 2025; Chi et al., 2025; Ding et al., 2024; Ke et al., 2024; Ma et al., 2025; Psenka et al., 2024; Ren et al., 2025; Scheikl et al., 2024; Wang et al., 2024). These methods represent the action distribution at each state through a parameterized diffusion process. This allows the policy to capture rich and potentially multimodal action distributions that are difficult to express with standard policy families, while still allowing actions to be sampled by simulating the diffusion process.

Why diffusion policies in classical RL? Even when an optimal policy is deterministic, stochastic policies can explore alternative actions and retain multiple promising behaviors during learning. Diffusion policies combine this multimodal flexibility with sampling that avoids terminal-density evaluation. Diffusion Policy Policy Optimization (DPPO) reports exploration aligned with demonstration data, stable training through iterative denoising, and robustness to perturbations in the dynamics and initial conditions (Ren et al., 2025). Approaching deterministic optimal behavior, however, requires the drift to concentrate actions despite nondegenerate noise. Our analysis asks how much drift complexity this concentration requires and quantifies its finite-sample cost through the drift Lipschitz budget K.

Despite this empirical success, the theoretical understanding of diffusion-based policies in RL remains limited. Existing theory for diffusion models mainly focuses on generative modeling and does not directly explain what drives the expressivity of diffusion policies for decision-making objectives. In particular, it remains unclear which structural quantity controls their ability to approximate optimal policies, and how to fully exploit this expressive power when the policy is learned from finite data. This leads to two fundamental questions:

1. **Expressivity of diffusion policies**: What mathematically governs the ability of diffusion policies to approximate optimal policies for a given RL objective?
2. **Impact on learning and statistical guarantees:** How does the increased expressivity of diffusion policies affect the finite-sample statistical error of policy learning?

Answering these questions is challenging because the expressivity of diffusion policies is not merely a qualitative property of the representation. It depends on the Markov decision process (MDP) structure, the regularity of the optimal policy, and the complexity of the diffusion dynamics. Moreover, increasing expressivity may improve approximation while enlarging the statistical complexity of the learned policy class. This calls for a finite-sample theory that identifies the structural quantity driving expressivity and quantifies its role in the resulting approximation–estimation trade-off.

1.1 Our contributions

In this paper, we identify the drift Lipschitz budget K as a central quantity governing the expressivity and statistical behavior of diffusion policies. We consider diffusion policies defined by state-conditioned stochastic differential equations with parameterized drifts and given volatility, and measure the complexity of the policy class through the drift Lipschitz constant K. Our main contributions are four-fold.

First, we identify an intrinsic $1/K$ limit on squared localization around finite target sets, uniformly over admissible K-Lipschitz drifts and diffusion horizons under our noise and initialization assumptions (Proposition 3.3). Neither a richer drift architecture nor a different horizon can overcome this limit. A mean-reverting construction attains this order at a fixed positive horizon (Lemma 3.1). We further establish convergence to optimal value under local continuity (Theorem 3.2) and a $1/K$ value-gap lower bound under the stated quadratic Bellman-gap condition (Theorem 3.4).

Second, we show that this increased expressivity comes with a statistical cost when the drift is parameterized by ReLU neural networks. For generic K-Lipschitz drifts, the error bound combines

diffusion approximation $1/K$, neural-network realization $s^{-2/m}$, and statistical estimation $\sqrt{sK/n}$, where s is the network class size, n is the sample size, and m is the state dimension. Balancing these terms yields a value gap of order $\widetilde{\mathcal{O}}(n^{-2/(m+6)})$ (Theorem 4.1). Under one-sided dissipativity, improved stability reduces the statistical term to $\sqrt{s/n}$, up to logarithmic factors, yielding the sharper rate $\widetilde{\mathcal{O}}(n^{-2/(m+4)})$ (Remark 4.1). We further discuss how additional smoothness and lower effective dimension of $g^\star$ improve the network approximation term and yield sharper rates for suitably chosen drift classes (Remark 4.2).

Third, we derive a Girsanov-based policy-gradient formula that uses action values without their derivatives and avoids evaluating the terminal action density. For continuous-time diffusion policies, the formula replaces the generally intractable policy score with a stochastic integral involving the drift sensitivity (Proposition 5.1). This identity provides the basis for the training used in our numerical experiments.

Fourth, our numerical experiments illustrate the practical consequences of the two statistical regimes. Training a generic neural drift with our policy-gradient estimator in a nonlinear multi-step MDP shows a sample-dependent preferred drift budget: larger training batches support larger K. A controlled contextual-bandit experiment with mean-reverting, hence dissipative, drifts isolates diffusion error from the center-estimation floor. It illustrates a minimum sufficient budget, beyond which policy regret plateaus. Together, these experiments suggest a practical *implementation principle*: choose the diffusion budget and network capacity jointly according to the available data and the stability of the drift class.

Some of the proofs are highly nontrivial and we summarize the technical highlights in Appendix A.

1.2 Related literature

Our work is connected to literature on diffusion models, RL, and continuous-time stochastic analysis. The most related line studies diffusion models as flexible policy parameterizations for continuous-action RL. We also discuss diffusion-based planning and imitation learning, theory of score-based generative models, and approximation, statistical learning, and PDE/SDE tools used in our analysis.

Diffusion policy in RL. One line of work directly represents the policy by a conditional diffusion model and trains or optimizes the parameters of the diffusion model (Ding et al., 2024; Fang et al., 2025; Ma et al., 2025; Psenka et al., 2024; Ren et al., 2025; Wang et al., 2024; 2023; Yang et al., 2023). Recent extensions further develop likelihood-based on-policy diffusion policy optimization, continuous-time Q-score matching, and maximum-entropy diffusion-policy training via adjoint matching (Ding et al., 2025; Hua et al., 2025; Thilges et al., 2026). A second line uses diffusion models primarily as samplers, generating actions from a given intermediate policy while training a classical RL algorithm (Chen et al., 2024a; 2023a; 2024b; Hansen-Estruch et al., 2023; Lu et al., 2023; Mao et al., 2024). Both lines of work are largely empirical, but they consistently demonstrate the strong expressive power of diffusion policies on large-scale RL tasks. In contrast, we provide theoretical foundations for this empirical observation by studying how the diffusion Lipschitz budget controls both expressivity and finite-sample statistical error in diffusion policy learning. To the best of our knowledge, this is among the first theoretical results of this kind in the literature.

Theory of diffusion and score-based generative models. Our analysis is also related to the theory of diffusion and score-based generative models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021; Tzen & Raginsky, 2019; De Bortoli, 2022; Lee et al., 2022; Chen et al., 2023c;b). Existing theory in this area primarily studies sampling, score estimation, and distribution recovery. In contrast, we connect the diffusion drift budget to RL performance by showing that K-Lipschitz diffusion policies achieve, and in general cannot improve upon, a $1/K$ value-approximation rate near deterministic optima.

Flow-based policy parameterizations. A closely related recent direction replaces stochastic diffusion sampling by ODE or flow-based generative policies, motivated in part by faster sampling and more direct pathwise or likelihood-ratio training (Lyu et al., 2025; Doo et al., 2026; Shin et al., 2026; Hu et al., 2026; Dong et al., 2026; Kim et al., 2026; Zhou et al., 2026; Zhong et al., 2026). These methods are complementary to ours: Pathwise training typically relies on action-value gradients, while likelihood-ratio training for generic flows requires inversion and Jacobian control or reversible

constructions. Our Girsanov-based gradient for diffusion policies uses action values without their derivatives and avoids inverting the sampling map or evaluating terminal densities.

Additional discussion of planning, imitation learning and related technical tools is in Appendix E.

Roadmap. Section 2 introduces the RL setting and diffusion policies. Section 3 establishes upper and lower value-approximation bounds for diffusion policy, while Section 4 develops finite-sample statistical error for ReLU drift. Section 5 presents policy-gradient implementation and experiments illustrating the generic $n^{-2/(m+6)}$ and dissipative $n^{-2/(m+4)}$ regimes. We conclude in Section 6.

Throughout, $\|\cdot\|$ denotes the Euclidean norm on finite-dimensional vector spaces. For a map $h : E \to \mathbb{R}^q$, we define $\|h\|_\infty := \sup_{x\in E} \|h(x)\|$ and $\mathrm{Lip}(h) := \sup_{x\neq x'} \|h(x) - h(x')\|/\|x - x'\|$. We use $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_{\mathrm{HS}}$ for the operator and Hilbert–Schmidt norms, respectively, and $\|\cdot\|_{L^p}$ for function-space L^p norms. The absolute constant $C > 0$ is generic and may vary line by line.

2 Preliminaries on MDP and diffusion policy

Markov decision process. Consider an infinite-horizon MDP $(\mathcal{X}, \mathcal{A}, P, r, \gamma)$ with measurable state space $\mathcal{X}$, action space $\mathcal{A} = \mathbb{R}^d$, transition kernel $P \in \mathcal{P}(\mathcal{X}|\mathcal{X} \times \mathcal{A})$, bounded continuous reward $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, and discount factor $\gamma \in [0, 1)$. A Markov stochastic policy $\pi \in \mathcal{P}(\mathcal{A}|\mathcal{X})$ assigns an action distribution $\pi(\cdot|x)$ to each state $x \in \mathcal{X}$. Starting from the initial state x_0 sampled from $\rho \in \mathcal{P}(\mathcal{X})$, actions and states evolve as $a_h \sim \pi(\cdot|x_h)$ and $x_{h+1} \sim P(\cdot|x_h, a_h)$ for $h \geq 0$. We consider the following maximization problem:

$$\sup_{\pi\in\mathcal{P}(\mathcal{A}|\mathcal{X})} V^\pi(\rho), \quad \text{with} \quad V^\pi(\rho) := \mathbb{E}^\pi_\rho\left[\sum_{h=0}^{\infty} \gamma^h r(x_h, a_h)\right], \tag{2.1}$$

where $\mathbb{E}^\pi_\rho$ denotes expectation under the state-action law induced by ρ and π. We write $V^\pi(x) = V^\pi(\delta_x)$, where δ_x is the Dirac measure at x.

A deterministic policy takes the form $\pi(\cdot|x) = \delta_{g(x)}$ for a measurable action selector $g : \mathcal{X} \to \mathcal{A}$. Under mild continuity conditions (Hernández-Lerma & Lasserre, 2012, Theorem 4.2.3), optimal policies can be chosen to be deterministic. We assume that the optimal value function $V^\star : \mathcal{X} \to \mathbb{R}$ is well-defined and measurable, and define its corresponding optimal Q-function by $Q^\star(x, a) := r(x, a) + \gamma \int_{\mathcal{X}} V^\star(x')P(\mathrm{d}x'|x, a)$. We further assume that there exists a measurable selector $g^\star : \mathcal{X} \to \mathcal{A}$ satisfying

$$g^\star(x) \in A^\star(x), \quad \text{with} \quad A^\star(x) := \arg\max_{a\in\mathbb{R}^d} Q^\star(x, a), \quad \forall x \in \mathcal{X}. \tag{2.2}$$

Then $g^\star$ is an optimal deterministic policy.

Diffusion policy parametrization. Although an optimal deterministic policy exists, differentiable stochastic policy classes support exploration during learning and likelihood-ratio policy-gradient estimation (Schulman et al., 2015; 2017; Sutton et al., 1999). Our parameterization utilizes stochasticity as part of the policy parameterization while retaining the classical expected-return objective in (2.1). Motivated by the need for expressive stochastic policies that remain compatible with sampling and gradient-based optimization, we consider diffusion-model-based policies, which define $\pi(\cdot|x)$ as the terminal law of a diffusion process. Specifically, let $T > 0$ be a fixed time horizon, let $(\Omega, \mathcal{F}, \mathbb{P}_B)$ be a probability space supporting a d-dimensional Brownian motion, and let $\mathbb{F}$ be the natural filtration of B and the initial action a_0. For each $x \in \mathcal{X}$, consider the following dynamics in $\mathcal{A} = \mathbb{R}^d$:

$$d\bar{a}^x_t = f(t, \bar{a}^x_t, x)\, dt + \sigma(t, \bar{a}^x_t, x)\, dB_t, \qquad \bar{a}^x_0 \sim \mu_0. \tag{2.3}$$

where $f : [0, T] \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}^d$ and $\sigma : [0, T] \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}^{d\times d}$ are some measurable functions such that (2.3) admits a unique weak solution $(\bar{a}^x_t)_{t\geq 0}$. The initial distribution μ_0 is assumed to be absolutely continuous w.r.t. the Lebesgue measure and have finite second moment, denoted by $m_2(\mu_0) := \int \|a\|^2\mu_0(da) < \infty$. In practice, a Gaussian initial distribution μ_0 is a common choice.

Remark 2.1 (Non-Gaussian terminal marginals)**.** *With state-dependent volatility, meaning that σ depends on the evolving diffusion state $\bar{a}^x_t$, the terminal marginal distribution* $\mathrm{Law}(\bar{a}^x_T)$ *is generally non-Gaussian and can exhibit nontrivial structure, even when the initialization is Gaussian and the drift is affine in the action variable.*

Throughout we assume that

Assumption 2.1 (Volatility assumption). *There exist absolute constants $0 < \kappa \le \Lambda < \infty$ such that $\kappa I_d \preceq \sigma(t,a,x)\sigma(t,a,x)^\top \preceq \Lambda I_d$, for all $t \in [0,T]$, $a \in \mathcal{A}$ and $x \in \mathcal{X}$.*

Define the corresponding policy by

$$\pi(\cdot \mid x) := \mathrm{Law}_{\mathbb{P}_B}\big(\bar{a}_T^x\big). \tag{2.4}$$

The collection of all such policies, obtained by varying f, defines the diffusion policy class, which arises as the natural continuous-time limit of the diffusion-based policies used in Balim et al. (2025); Ding et al. (2024); Ma et al. (2025); Psenka et al. (2024); Ren et al. (2025); Wang et al. (2024). In the sequel, we measure the complexity of a diffusion policy through the Lipschitz size of its drift in the action variable. Unless otherwise stated, K denotes a uniform Lipschitz bound in a, namely

$$\|f(t,a,x) - f(t,a',x)\| \le K\|a - a'\|, \qquad t \in [0,T],\ x \in \mathcal{X}.$$

We refer to K as the drift Lipschitz budget.

3 Expressivity of diffusion policy

We now turn to the approximation power of diffusion policies. The goal of this section is to show how the drift Lipschitz budget K controls their ability to concentrate near deterministic optimal actions, and how this translates into value-function approximation.

3.1 Upper bound on value function approximation with K-Lipschitz drifts

In this section, we show that diffusion policies can approximate deterministic policies arbitrarily well as the Lipschitz constant of the drift increases. This is relevant because a larger Lipschitz constant allows the drift to represent more flexible neural networks. Under Assumption 2.1 and a fixed initialization, we construct a K-Lipschitz drift whose expressivity gap vanishes as $K \to \infty$. The key step is the following lemma. Although the volatility σ depends on the state $\bar{a}_t^x$, so that the terminal law is generally not Gaussian, we can still obtain the following concentration estimate.

Lemma 3.1 (Concentration with mean-reverting drift). *Assume Assumption 2.1 holds. Fix $\tau > 0$, $K > 0$, $c \in \mathbb{R}^d$, and $x \in \mathcal{X}$. Consider a linear drift $f(s,a,x) = K(c/(1-e^{-K\tau}) - a)$ in* (2.3)*, and the following diffusion process $d\bar{a}_s^x = K\left(\frac{c}{1-e^{-K\tau}} - \bar{a}_s^x\right)ds + \sigma(s,\bar{a}_s^x,x)\,dB_s$ with $\quad s \in [0,\tau]$, where $\bar{a}_0^x \sim \mu_0$ is independent of B . Then*

$$\mathbb{E}\|\bar{a}_\tau^x - c\|^2 \le e^{-2K\tau} m_2(\mu_0) + \frac{d\Lambda}{2K}\big(1 - e^{-2K\tau}\big),$$

where $m_2(\mu_0)$ is the second moment of μ_0.

Consequently, for fixed $\tau > 0$, c, and Λ independent of K, we have $\mathbb{E}\|\bar{a}_\tau - c\|^2 \lesssim 1/K$, with the omitted constant only depending on $\tau, d, m_2(\mu_0), \Lambda$. The proof is given in Appendix B.2. This lemma shows that $\bar{a}_\tau^x$ can approximate c at any time τ, as long as K is sufficiently large. Taking $c = g^\star(x)$ therefore yields a K-Lipschitz drift that concentrates near an optimal action.

We next translate this localization estimate into value-function approximation. This requires the reward and transition kernel to be continuous near the optimal action $g^\star(x)$, as formalized below.

Assumption 3.1 (Moduli of continuity). *For every state x, there exists $R_x > 0$ and moduli of continuity $\omega_{r,x}(z), \omega_{P,x}(z) \downarrow 0$ as $z \downarrow 0$ such that for all actions a satisfying $\|a - g^\star(x)\| \le R_x$,*

$$\begin{aligned} |r(x,a) - r(x,g^\star(x))| &\le \omega_{r,x}(\|a - g^\star(x)\|) \\ \mathrm{TV}\left(P(\cdot|x,a), P(\cdot|x,g^\star(x))\right) &\le \omega_{P,x}(\|a - g^\star(x)\|) \end{aligned}$$

This is a mild local condition: it requires continuity only near $g^\star(x)$, rather than global smoothness assumptions on the MDP. With this assumption, we can achieve value function convergence.

Theorem 3.2 (Value approximation by mean-reverting diffusion policies). *Assume Assumptions 2.1 and 3.1 hold. For each $K \ge 1$, define*

$$c_K(x) := \frac{g^\star(x)}{1 - \exp(-KT)}. \tag{3.1}$$

Suppose that, for each $K \geq 1$ and $x \in \mathcal{X}$, the mean-reverting diffusion

$$d\bar{a}_t^{K,x} = K\big(c_K(x) - \bar{a}_t^{K,x}\big)\,dt + \sigma(t, \bar{a}_t^{K,x}, x)\,dB_t, \qquad \bar{a}_0^{K,x} \sim \mu_0, \tag{3.2}$$

admits a weak solution on $[0,T]$, where $\bar{a}_0^{K,x}$ is independent of B and μ_0 has finite second moment. Assume moreover that the terminal laws define a measurable stochastic policy $\pi_K(\cdot \mid x) := \mathrm{Law}(\bar{a}_T^{K,x})$. Then, for every initial state law ρ,

$$\lim_{K\to\infty} |V^{\pi_K}(\rho) - V^\star(\rho)| = 0.$$

The proof is in Appendix B.3. We remark that the same argument can yield uniform convergence of the value function when the radius and moduli of continuity are uniform over x.

3.2 Lower bound on value function approximation for K-Lipschitz drifts

The upper-bound result shows that a K-Lipschitz mean-reverting drift can concentrate a diffusion policy near an optimal deterministic action with squared error of order $1/K$. We now show that this rate is intrinsic: under nondegenerate diffusion noise, no K-Lipschitz drift can localize the terminal action around a finite set of target actions at a faster rate.

Proposition 3.3 (Matching lower bound for finite-set localization)**.** *Let $\mathcal{Z} = \{z_1, \dots, z_J\} \subset \mathbb{R}^d$ be a finite set and fix $R > 0$. Let $(\bar{a}_t^x)_{t\geq 0}$ follow (2.3), with $\bar{a}_0^x$ independent of B. Suppose that $f(t,\cdot,x)$ is K-Lipschitz uniformly in t, x and $f(\cdot, 0, x) \in L^1_{loc}$. Additionally, assume either Assumption 2.1 or*

(A2) (State-independent volatility) $\sigma(t,a,x) \equiv \sigma(t)$ is deterministic and locally square-integrable, and $\sigma(t)\sigma(t)^\top \succeq \kappa I_d$, where $\kappa > 0$ is independent of K.

Then there exist $c_R > 0$ and $K_R < \infty$, depending on J, d, κ, μ_0, R, and additionally on Λ in the state-dependent case, such that, for every $K \geq K_R$,

$$\inf_{f,\sigma,x,T\geq 0} \mathbb{E}\left[\mathsf{dist}(\bar{a}_T^x, \mathcal{Z})^2 \wedge R^2\right] \geq \frac{c_R}{K}.$$

Here $\mathsf{dist}(a, \mathcal{Z}) := \inf_{z\in\mathcal{Z}} \|a - z\|$. The constant c_R does not depend on the locations of the points in $\mathcal{Z}$, nor on the minimum distance between distinct points of $\mathcal{Z}$.

The proof is novel with a phase transition that separates horizons shorter than K^{-1} from longer horizons. In the latter case, noise accumulated over the final interval of length K^{-1} leaves fluctuations of order $K^{-1/2}$, yielding the squared-distance lower bound. Details are in Appendices B.5 and B.9.

Remark 3.1 (Sharpness of the localization rate)**.** *Proposition 3.3 shows that the $1/K$ localization rate in Lemma 3.1 is sharp, even when there are several isolated optimal actions. In RL, taking $\mathcal{Z} = A^\star(x)$ therefore identifies K as a limit on how closely the policy can concentrate near the optimal action set.*

Proposition 3.3 immediately yields a value-function lower bound. The only additional assumption is that the Bellman gap grows at least quadratically locally near the finite set of optimal actions.

Theorem 3.4 (Finite-well diffusion lower bound)**.** *For each $K \geq 1$ and $T > 0$, let $\pi_{K,T}$ be the Markov policy whose action law at state x is the terminal law at time T of (2.3). Assume that the conditions in Proposition 3.3 hold.*

Fix $M \geq 1$ and $R > 0$. Suppose that, for every $x \in \mathcal{X}$, $|A^\star(x)| \leq M$, and that there exists a constant $\lambda_{M,R} > 0$, independent of x, such that for all $(x,a) \in \mathcal{X} \times \mathcal{A}$,

$$V^\star(x) - Q^\star(x,a) \geq \lambda_{M,R}\left(\mathsf{dist}(a, A^\star(x))^2 \wedge R^2\right). \tag{3.3}$$

Then there exist constants $c_{M,R} > 0$ and $K_0 < \infty$, depending only on $M, R, d, \kappa, \Lambda, \mu_0$, such that

$$\inf_{T>0}\left(V^\star(\rho) - V^{\pi_{K,T}}(\rho)\right) \geq \frac{c_{M,R}\lambda_{M,R}}{K}, \quad \forall K \geq K_0.$$

Condition (3.3) allows several isolated optimal actions and requires a positive suboptimality margin away from them, commonly assumed in the RL literature (Wang et al., 2026; Simchowitz & Jamieson,

2019; Yang et al., 2021). Appendix B.1 discusses this condition and the extension to arbitrary initialization after a short burn-in period. The proof is given in Appendix B.6.

Our results identify the drift Lipschitz budget K as an *intrinsic complexity parameter* of diffusion policies. The matching bounds establish $1/K$ as the best achievable squared-localization scale around an action, independent of the drift architecture or training algorithm. Thus, K controls how finely a policy can concentrate its actions while retaining nondegenerate stochasticity. Our finite-sample analysis in Section 4 further quantifies how this approximation benefit interacts with statistical error.

Beyond deterministic targets, the Lipschitz budget also reflects the geometry of smooth distributions. For instance, with fixed isotropic noise, the canonical Langevin drift has a Lipschitz constant proportional to the maximum curvature of the target's log-density, linking the budget to peak widths and log-density curvature between modes.

4 Approximation–estimation trade-off in K

We now study the finite-sample cost of learning a K-Lipschitz diffusion policy with a ReLU drift. We take $\mathcal{X} = [0,1]^m$ for convenience and impose the following assumptions.

Assumption 4.1 (Bounded and Lipschitz optimal deterministic policy)**.** *The optimal selector $g^\star$, defined in* (2.2)*, satisfies $\|g^\star\|_\infty \le R_0$ and $\mathrm{Lip}(g^\star) \le L_{g^\star}$.*

Assumption 4.2 (MDP environment and diffusion policy volatility)**.** *The reward satisfies*

$$|r(x,a) - r(x',a')| \le L_r^x \|x - x'\| + L_r^a \|a - a'\|.$$

The transition admits a measurable generator $\Phi_P(x,a,U) \sim P(\cdot \mid x,a)$, such that

$$\big(\mathbb{E}_U \|\Phi_P(x,a,U) - \Phi_P(x',a',U)\|^2\big)^{1/2} \le L_P^x \|x - x'\| + L_P^a \|a - a'\|. \tag{4.1}$$

where $U \sim \mathrm{Unif}[0,1]$. Moreover,

$$\|\sigma(t,a,x) - \sigma(t,a',x')\|_{\mathrm{HS}} \le L_\sigma\big(\|x - x'\| + \|a - a'\|\big). \tag{4.2}$$

Note that such a measurable generator Φ_P always exists on standard Borel state spaces (Kallenberg, 1997, Lemma 3.22).

Drift class and empirical value. We use a ReLU drift class $\mathcal{F}_{K,s}$, where K controls the action Lipschitz bound and s controls the network size. The architecture and parameter bounds are specified in Appendix C.1. The class contains a mean-reverting drift centered at a ReLU approximation h_s of $g^\star$, with squared uniform error of order $s^{-2/m}$.

Let $V(f) := V^{\pi_f}(\rho)$ denote the value of the terminal-law policy induced by f. Draw n independent collections of simulator randomness, each comprising an initial state, diffusion initializations, Brownian paths, and uniform random variables for the transition generator. Use the same collections for every f to generate trajectories $(X_{i,j}^f, A_{i,j}^f)$, and define the following unbiased estimator for each f:

$$\widehat{V}_n(f) := \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{\infty} \gamma^j r(X_{i,j}^f, A_{i,j}^f), \quad \widehat{f}_{n,K,s} \in \operatorname*{arg\,max}_{f \in \mathcal{F}_{K,s}} \widehat{V}_n(f). \tag{4.3}$$

Theorem 4.1 (Finite-sample trade-off for Lipschitz drifts)**.** *Suppose Assumptions 2.1, 4.1, and 4.2 hold, and the Bellman gap satisfies*

$$0 \le V^\star(x) - Q^\star(x,a) \le L_Q \|a - g^\star(x)\|^2, \qquad x \in [0,1]^m,\ a \in \mathbb{R}^d. \tag{4.4}$$

For every $n \ge 1$, $K, s \ge 1$, and $\delta \in (0,1)$, with probability at least $1-\delta$,

$$V^\star(\rho) - V(\widehat{f}_{n,K,s}) \le \widetilde{O}\Big(\ \frac{1}{K} + s^{-2/m} + \sqrt{\frac{sK}{n}} + \frac{sK}{n}\ \Big). \tag{4.5}$$

Here $\widetilde{O}$ suppresses logarithmic factors in $(n, K, s, 1/\delta)$. In particular, $K \asymp n^{2/(m+6)}$ and $s \asymp n^{m/(m+6)}$ give a value error of order $\widetilde{O}(n^{-2/(m+6)})$.

The first two terms in (4.5) are the diffusion and network approximation errors. The statistical terms increase with both s and K, giving a sample-dependent diffusion budget. The dimension-explicit bound and proof appear in Appendices C.1 and C.3.

Remark 4.1 (Dissipative drift classes and better rate)**.** *If the drift class additionally satisfies*

$$\langle a - a', f(t,a,x) - f(t,a',x)\rangle \le -\lambda K\|a-a'\|^2 \tag{4.6}$$

for some $\lambda > 0$ independent of (K, s), the leading statistical term improves to $\widetilde{O}(\sqrt{s/n})$. The mean-reverting comparison drift satisfies this condition with $\lambda = 1$. Balancing network approximation and statistical error gives $s \asymp n^{m/(m+4)}$ and the improved rate $\widetilde{O}(n^{-2/(m+4)})$, provided $K \gtrsim n^{2/(m+4)}$, This is a minimum sufficient scale for K: larger values attain the same leading error floor. The full comparison and a supporting experiment are given in Appendices C.1 and G.

Remark 4.2 (Sharper rates under smoothness and intrinsic dimension.)**.** *Additional structure of $g^\star$ can sharpen the approximation–estimation trade-off. For example, suppose $g^\star$ has intrinsic dimension $r < m$ and Hölder smoothness $\beta > 1$ in the intrinsic variables. Standard ReLU approximation results yield squared uniform error $\widetilde{\mathcal{O}}(s^{-2\beta/r})$ using $\widetilde{\mathcal{O}}(s)$ parameters (Schmidt-Hieber, 2020). Replacing $s^{-2/m}$ by $s^{-2\beta/r}$ in the oracle inequality yields the rate $\widetilde{\mathcal{O}}(n^{-2\beta/(r+6\beta)})$. If the class additionally satisfies one-sided dissipativity, the statistical term becomes independent of K, up to logarithmic factors, giving the sharper rate $\widetilde{\mathcal{O}}(n^{-2\beta/(r+4\beta)})$. As $\beta \to \infty$, or that the optimal action is smooth with bounded gradients, the rate approaches $n^{-1/2}$ with dissipative architectures. With intrinsic dimension $r < m$, the rate can approach $n^{-1/2}$ faster compared to m.*

5 Numerical experiments

We derive a policy-gradient formula to facilitate fast gradient-based implementation; its convergence analysis is *not* the focus of this paper. Two experiments illustrate the finite-sample regimes in Section 4. The first uses generic neural drifts in a multi-step MDP to illustrate the sample-dependent budget $K \asymp n^{2/(m+6)}$ in Theorem 4.1. The second, in Appendix G, uses mean-reverting, one-sided dissipative drifts and varies K with $s = s(n) \asymp (n/\log n)^{m/(m+4)}$ fixed, following Remark 4.1.

5.1 Policy gradient expression

We derive a policy-gradient formula for the above diffusion-policy classes. Let the drift in (2.3) be parameterized by f_θ, with $\theta \in \mathbb{R}^k$, and denote the corresponding policy by π_θ. We require continuous differentiability in θ, and regularity and integrability conditions in Assumption D.1 of Appendix D.

Proposition 5.1 (Policy gradient theorem)**.** *Suppose Assumption D.1, Assumption 2.1 and* (4.2) *hold. Then the gradient of the value function with respect to the parameter θ follows*

$$\nabla_\theta V^{\pi_\theta}(\rho) = \frac{1}{1-\gamma}\int_{\mathcal{X}} \mathbb{E}\big[S_\theta(x) Q^{\pi_\theta}(x, \bar{a}_T^x)\big]\, d_\rho^{\pi_\theta}(\mathrm{d}x),$$

where $Q^\pi(x,a) := r(x,a) + \gamma\int_{\mathcal{X}} V^\pi(x')P(\mathrm{d}x'|x,a)$, d_ρ^π is the normalized discounted state distribution under π with initial distribution ρ, and $S_\theta(x) := \int_0^T \nabla_\theta f_\theta(t, \bar{a}_t^x, x)\big(\sigma^\top(t,\bar{a}_t^x,x)\big)^{-1} dB_t$.

This identity is in the same spirit as the classical policy-gradient theorem (Sutton et al., 1999), but tailored towards diffusion policy with stochastic integral S_θ. The derivation and comparison with discrete-time diffusion-policy gradients are given in Appendix D.

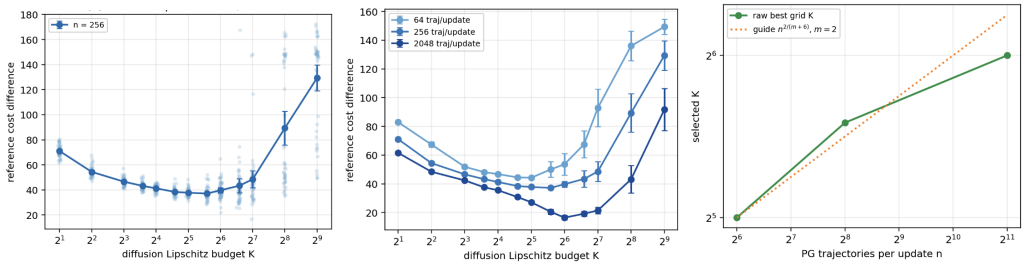
5.2 Experiment with generic drifts

This experiment tests the generic neural-drift regime of Theorem 4.1. The drift is represented by a generic neural network and trained directly by policy gradient. Another set of experiments with dissipative drift is deferred to Appendix G.

Setup. We consider a two-dimensional stochastic MDP with state $X_h = (X_h^{(1)}, X_h^{(2)}) \in \mathbb{R}^2$ and scalar bounded action $A_h \in [-R_a, R_a]$. The reward and transition specifications are given in Appendix F. The reference feedback g_{ref}, unknown to the algorithm, is a smooth saturated nonlinear feedback map. It is used to define a nontrivial reward landscape and to compute the reference cost.

Implementation details. At each MDP step, the policy samples the action from the terminal value of the internal action diffusion $d\bar{a}_t = f_{\theta,K}(t, \bar{a}_t, X_h)\,dt + \sigma\,dB_t$, where $T = 1$, $\sigma = 0.70$, and $\bar{a}_0 \sim N(0, 1.10^2)$. The drift is represented by a two-hidden-layer ReLU network, $f_{\theta,K}(t, a, x) = K\,\tanh(\mathrm{MLP}_\theta(t, a, x))$, with the network width chosen as $w(K) = \min\{256, \max\{16, \mathrm{round}(16+2K)\}\}$, so the width grows approximately as $K^{m/2}$ on the main part of the grid.

We train the actor using the path-likelihood score estimator from the Euler-discretized action SDE, together with a learned critic baseline evaluated at the pre-action state X_h. The optimizer settings and the grid of K are given in Appendix F. We evaluate $K \in \{2, 4, 8, 12, 16, 24, 32, 48, 64, 96, 128, 256, 512\}$, use $n \in \{64, 256, 2048\}$ trajectories per policy-gradient update, average over 50 seeds, and evaluate each trained policy using 1024 held-out episodes. We report held-out discounted cost relative to the deterministic reference feedback g_{ref}. The reference feedback is used only for constructing cost and for evaluation, not for actor training.



(a) Trade-off at $n = 256$. (b) Reference cost difference vs.K. (c) Empirical optimal $K(n)$ wrt n.

Figure 1: Empirical approximation–estimation trade-off in K. Panel (a) fixes $n = 256$ trajectories per policy-gradient update and shows a U-shaped dependence on the diffusion Lipschitz budget K. Panel (b) compares different values of n, showing that larger policy-gradient batches tolerate larger K. Panel (c) reports the best grid value of K, together with the generic guide $K(n) \asymp n^{2/(m+6)}$ for $m = 2$. Markers show averages over 50 seeds and error bars indicate 95% CIs.

Results. Figure 1(a) shows the fixed-sample trade-off at $n = 256$. The reference cost difference decreases from about 71.06 at $K = 2$ to 37.14 at $K = 48$. It then increases for larger budgets, reaching 89.36 at $K = 256$ and 129.36 at $K = 512$. Thus, increasing K is initially beneficial, but overly large K makes policy-gradient learning unstable under a fixed sample budget. Figure 1(b) shows that increasing the number of trajectories per update shifts the stable range of K to the right. With $n = 64$, the best grid value is $K = 32$, with mean reference cost difference 44.24. With $n = 256$, the best grid value is $K = 48$, with mean reference cost difference 37.14. With $n = 2048$, the best grid value is $K = 64$, with mean reference cost difference 16.42. The best-K summary in Figure 1(c) gives an increasing sequence $K = 32, 48, 64$ over $n = 64, 256, 2048$. The selected K increases with n, consistent with the predicted sample-dependent budget.

In the dissipative experiment in Appendix G, mean reversion stabilizes the policy class: with $s = s(n)$ fixed, increasing K reduces diffusion error until regret reaches the center-estimation floor. Thus, $K \asymp (n/\log n)^{2/(m+4)}$ is a minimum sufficient scale. Dissipativity holds by construction in this example, while enforcing it through parameter projection in general cases may be costly.

Overall, both sets of experiments support the two theoretical regimes: a U-shaped trade-off for generic neural drifts and a plateau under dissipativity.

6 Conclusion

We develop a theoretical framework for the expressivity and finite-sample behavior of diffusion policies in reinforcement learning. The drift Lipschitz budget K governs a sharp approximation limit, while stability determines its statistical cost. For ReLU drift classes, we obtain an explicit finite-sample oracle inequality and a sample-dependent choice of K, qualitatively supported by our experiments. Future directions include sharpening the finite-sample bound in Section 4 using Bernstein-type concentration under stronger variance or stability assumptions, and incorporating optimization error from practical policy-gradient training.

AI USE STATEMENT

Generative AI tools are used to suggest clearer wording for the interpretations in Sections 1 and 3, paraphrase to shorten paragraphs, and provide initial drafts for the first two sections from which we iterate and modify. They are also used to find literature for related work and references for analysis results that we employ in our work.

REPRODUCIBILITY STATEMENT

Statements and proofs for theoretical results are provided in the main text and Appendix, respectively. Code for the experiments is attached in the .zip file for this submission.

REFERENCES

Anurag Ajay, Yilun Du, Abhi Gupta, Joshua B. Tenenbaum, Tommi S. Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making? In *International Conference on Learning Representations*, 2023.

Donald G. Aronson. Bounds for the fundamental solution of a parabolic equation. *Bulletin of the American Mathematical Society*, 73(6):890–896, 1967.

Haldun Balim, Na Li, and Yilun Du. Model-based diffusion sampling for predictive control in offline decision making. *arXiv preprint arXiv:2512.08280*, 2025.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Vladimir I. Bogachev, Nicolai V. Krylov, Michael Röckner, and Stanislav V. Shaposhnikov. *Fokker–Planck–Kolmogorov Equations*, volume 207 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2015.

René Carmona. *Lectures on BSDEs, stochastic control, and stochastic differential games with financial applications*. SIAM, 2016.

Huayu Chen, Cheng Lu, Chengyang Ying, Hang Su, and Jun Zhu. Offline reinforcement learning via high-fidelity generative behavior modeling. In *International Conference on Learning Representations*, 2023a.

Huayu Chen, Cheng Lu, Zhengyi Wang, Hang Su, and Jun Zhu. Score regularized policy optimization through diffusion behavior. In *International Conference on Learning Representations*, 2024a.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4672–4712. PMLR, 2023b.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023c.

Tianyu Chen, Zhendong Wang, and Mingyuan Zhou. Diffusion policies creating a trust region for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024b.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

Shutong Ding, Ke Hu, Zhenhao Zhang, Kan Ren, Weinan Zhang, Jingyi Yu, Jingya Wang, and Ye Shi. Diffusion-based reinforcement learning via q-weighted variational policy optimization. In *Advances in Neural Information Processing Systems*, 2024.

Shutong Ding, Ke Hu, Shan Zhong, Haoyang Luo, Weinan Zhang, Jingya Wang, Jun Wang, and Ye Shi. GenPO: Generative diffusion models meet on-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Xiaoyi Dong, Xi Sheryl Zhang, and Jian Cheng. Mean flow policy optimization. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

JaeHyeok Doo, Byeongguk Jeon, Seonghyeon Ye, Kimin Lee, and Minjoon Seo. Q-Flow: Stable and expressive reinforcement learning with flow-based policy. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Linjiajie Fang, Ruoxue Liu, Jing Zhang, Wenjia Wang, and Bing-Yi Jing. Diffusion actor-critic: Formulating constrained policy iteration as diffusion noise regression for offline reinforcement learning. In *International Conference on Learning Representations*, 2025.

Bronius Grigelionis and Vigirdas Mackevičius. The finiteness of moments of a stochastic exponential. *Statistics & probability letters*, 64(3):243–248, 2003.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Bruce Hajek. Mean stochastic comparison of diffusions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 68(3):315–329, 1985.

Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. Idql: Implicit q-learning as an actor-critic method with diffusion policies. *arXiv preprint arXiv:2304.10573*, 2023.

Onésimo Hernández-Lerma and Jean B Lasserre. *Discrete-time Markov control processes: basic optimality criteria*, volume 30. Springer Science & Business Media, 2012.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ke Hu, Shutong Ding, Panxin Tao, Jingya Wang, and Ye Shi. GenPO++: Generative policy optimization with jacobian-free likelihood ratios. *arXiv preprint arXiv:2606.06967*, 2026.

Chengxiu Hua, Jiawen Gu, and Yushun Tang. Continuous Q-score matching: Diffusion guided reinforcement learning for continuous-time control. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Xiaoyu Huang, Yufeng Chi, Ruofeng Wang, Zhongyu Li, Xue Bin Peng, Sophia Shao, Borivoje Nikolic, and Koushil Sreenath. Diffuseloco: Real-time legged locomotion control with diffusion from offline datasets. *arXiv preprint arXiv:2404.19264*, 2024.

Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9902–9915. PMLR, 2022.

Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Olav Kallenberg. *Foundations of modern probability*. Springer, 1997.

Ioannis Karatzas and Steven Shreve. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 1991.

Norihiko Kazamaki. *Continuous exponential martingales and BMO*. Springer, 2006.

Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. *arXiv preprint arXiv:2402.10885*, 2024.

Bekzhan Kerimkulov, James-Michael Leahy, David Siska, Lukasz Szpruch, and Yufei Zhang. A fisher–rao gradient flow for entropy-regularised markov decision processes in polish spaces: B. kerimkulov et al. *Foundations of Computational Mathematics*, pp. 1–75, 2025.

Kyungyoon Kim, Donghyeon Ki, Hee-Jun Ahn, and Byung-Jun Lee. Score-based one-step MeanFlow policy optimization. *arXiv preprint arXiv:2605.23365*, 2026.

N. V. Krylov. On diffusion processes with drift in L_{d+1}. *arXiv preprint arXiv:2102.11465*, 2021.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22870–22882, 2022.

Wenhao Li, Xiangfeng Wang, Bo Jin, and Hongyuan Zha. Hierarchical diffusion for offline decision making. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20035–20064. PMLR, 2023.

Cheng Lu, Huayu Chen, Jianfei Chen, Hang Su, Chongxuan Li, and Jun Zhu. Contrastive energy prediction for exact energy-guided diffusion sampling in offline reinforcement learning. In *International Conference on Machine Learning*, 2023.

Lei Lyu, Yunfei Li, Yu Luo, Fuchun Sun, Tao Kong, Jiafeng Xu, and Xiao Ma. Flow-based policy for online reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Haitong Ma, Tianyi Chen, Kai Wang, Na Li, and Bo Dai. Efficient online reinforcement learning for diffusion policy. *arXiv preprint arXiv:2502.00361*, 2025.

Liyuan Mao, Haoran Xu, Xianyuan Zhan, Weinan Zhang, and Amy Zhang. Diffusion-dice: In-sample diffusion guidance for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024.

Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer, Berlin, 6 edition, 2003.

Weigutian Ou and Helmut Bölcskei. Covering numbers for deep ReLU networks with applications to function approximation and nonparametric regression. *arXiv preprint arXiv:2410.06378*, 2024.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.

Michael Psenka, Alejandro Escontrela, Pieter Abbeel, and Yi Ma. Learning a diffusion model policy from rewards via q-score matching. In *International Conference on Machine Learning*, 2024.

Allen Z. Ren, Justin Lidard, Lars Lien Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. Diffusion policy policy optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=mEpqHvbD2h.

L. C. G. Rogers and David Williams. *Diffusions, Markov Processes and Martingales. Volume 2: Itô Calculus*. Cambridge University Press, 2 edition, 2000.

Paul Maria Scheikl, Nicolas Schreiber, Christoph Haas, Niklas Freymuth, Gerhard Neumann, Rudolf Lioutikov, and Franziska Mathis-Ullrich. Movement primitive diffusion: Learning gentle robotic manipulation of deformable objects. *IEEE Robotics and Automation Letters*, 9(6):5338–5345, 2024.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Yongjae Shin, Jongseong Chae, Jongeui Park, and Youngchul Sung. Flow matching with injected noise for offline-to-online reinforcement learning. In *International Conference on Learning Representations*, 2026.

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

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Daniel W. Stroock and S. R. S. Varadhan. *Multidimensional Diffusion Processes*, volume 233 of *Grundlehren der Mathematischen Wissenschaften*. Springer, Berlin, 1979.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Serge Thilges, Onur Celik, Denis Blessing, Emiliyan Gospodinov, and Gerhard Neumann. Scalable maximum entropy reinforcement learning for diffusion policies via adjoint matching. *arXiv preprint arXiv:2606.22630*, 2026.

Belinda Tzen and Maxim Raginsky. Theoretical guarantees for sampling and inference in generative models with latent diffusions. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 3084–3114. PMLR, 2019.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Shengbo Wang, Jose Blanchet, and Peter Glynn. Fast convergence of policy regret in learning stochastic optimal control. *arXiv preprint arXiv:2605.26361*, 2026.

Yinuo Wang, Likun Wang, Yuxuan Jiang, Wenjun Zou, Tong Liu, Xujie Song, Wenxuan Wang, Liming Xiao, Jiang Wu, Jingliang Duan, et al. Diffusion actor-critic with entropy regulator. *Advances in Neural Information Processing Systems*, 37:54183–54204, 2024.

Zhendong Wang, Jonathan J. Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. In *International Conference on Learning Representations*, 2023.

Kunhe Yang, Lin Yang, and Simon Du. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pp. 1576–1584. PMLR, 2021.

Long Yang, Zhixiong Huang, Fenghao Lei, Yucun Zhong, Yiming Yang, Cong Fang, Shiting Wen, Binbin Zhou, and Zhouchen Lin. Policy representation via diffusion probability model for reinforcement learning. *arXiv preprint arXiv:2305.13122*, 2023.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94: 103–114, 2017. doi: 10.1016/j.neunet.2017.07.002. URL https://doi.org/10.1016/j.neunet.2017.07.002.

Hai Zhong, Zhuoran Li, Xun Wang, and Longbo Huang. Reparameterization flow policy optimization. *arXiv preprint arXiv:2602.03501*, 2026.

Xubin Zhou, Yipeng Yang, and Zhan Li. Truncated rectified flow policy for reinforcement learning with one-step sampling. *arXiv preprint arXiv:2604.09159*, 2026.

A TECHNICAL HIGHLIGHTS

We summarize the proof approaches in the paper.

- A key technical contribution of the paper is to turn the informal intuition that "larger drift yields sharper diffusion policies" into a sharp approximation–estimation theory.
  - The upper-bound argument is constructive: we use a linear mean-reverting drift with Lipschitz budget K to pull the terminal action toward an optimal selector $g^\star(x)$, and Itô's formula together with Gronwall's inequality yields an $\mathcal{O}(K^{-1})$ terminal mean-square localization error even under bounded state-dependent volatility. This localization estimate is then lifted to RL performance by comparing the one-step reward and transition kernels induced by the diffusion policy with those of the deterministic optimal policy, and propagating this comparison through the discounted Bellman recursion.
  - The lower-bound argument shows that the same rate is intrinsic. It splits time at the natural scale K^{-1}: on shorter horizons, absolute continuity of the initialization and volume control prevent concentration near finitely many target actions; on longer horizons, the final K^{-1} window is rescaled by $K^{1/2}$, yielding a unit-time residual diffusion with uniformly nondegenerate noise. A density bound in the state-independent case, and a local Krylov lower estimate plus a pigeonhole argument in the state-dependent case, imply that this residual fluctuation remains at distance $\Omega(K^{-1/2})$ from any finite optimal-action set with constant probability, giving the matching $\Omega(K^{-1})$ squared-localization and value lower bound.
- The finite-sample theorem then combines this sharp diffusion approximation error with two learning errors: ReLU center networks approximate Lipschitz optimal selectors at uniform rate $s^{-1/m}$, which becomes $s^{-2/m}$ under a quadratic Bellman gap, while terminal-map stability and covering-number bounds yield statistical error of order $\sqrt{sK/n}$, up to logarithmic factors. Balancing these terms gives the rate $\widetilde{\mathcal{O}}(n^{-2/(m+6)})$ with $K \asymp n^{2/(m+6)}$. Under one-sided dissipativity, the leading statistical term improves to $\sqrt{s/n}$, yielding the sharper rate $\widetilde{\mathcal{O}}(n^{-2/(m+4)})$. In this regime, $K \gtrsim n^{2/(m+4)}$ suffices to attain the same leading error floor, up to logarithmic factors.

B PROOFS FOR SECTION 3

This appendix contains the proofs for the expressivity results in Section 3. We first prove the constructive upper bound showing that mean-reverting diffusion policies concentrate near deterministic selectors, then prove the matching lower bound showing that the resulting K^{-1} rate is intrinsic under nondegenerate diffusion noise.

B.1 FURTHER DISCUSSION OF THE LOWER BOUND

The proofs are given in Appendix B.5 for Assumption A2, and Appendix B.9 for Assumption 2.1. The key observation, which appears to be new, is to split at the intrinsic time K^{-1}. On time scales shorter than K^{-1}, a K-Lipschitz drift cannot move an absolutely continuous initial law into arbitrarily small neighborhoods of finitely many target points. On longer horizons, one can look only at the final window of length K^{-1}. After rescaling this window by

$$\bar{a}_{T-1/K+\ell/K} = \phi_\ell + K^{-1/2} Y_\ell, \qquad 0 \le \ell \le 1,$$

the residual process Y_ℓ has $\mathcal{O}(1)$-Lipschitz drift and uniformly nondegenerate volatility.

As a result, Y_1 has a nondegenerate density uniformly in K, hence having distance at least c from any finite set with constant probability. Scaling back, $\bar{a}_T$ is at distance at least $cK^{-1/2}$ from $\mathcal{Z}$ with constant probability, yielding the c/K lower bound after squaring.

Remark B.1. *We give a few remarks about Proposition 3.3.* ***(i) Role of bounded volatility.*** *For state-independent volatility, uniform ellipticity is enough. For state-dependent volatility, the upper bound in Assumption 2.1 is essential: Proposition B.5 shows that, if this upper bound is removed,*

the $1/K$ lower bound can fail even when the volatility remains uniformly elliptic from below. ***(ii) Sharpness.*** *The lower bound matches the $\mathcal{O}(1/K)$ upper bound from the mean-reverting construction in Lemma 3.1. Indeed, with finite second moment initialization and uniformly bounded volatility, a linear mean-reverting drift can achieve $\mathbb{E}\|\bar{a}_T - z\|^2 \lesssim K^{-1}$ for any fixed $z \in \mathbb{R}^d$, and hence also $\mathbb{E}[\mathsf{dist}(\bar{a}_T, \mathcal{Z})^2] \lesssim K^{-1}$ for any finite set $\mathcal{Z}$. Proposition 3.3 shows that this rate cannot be improved in general.* ***(iii) Relevance for RL.*** *In RL, $\mathcal{Z}$ should be viewed as the Bellman-optimal action set $A^\star(x)$. Although the upper bound only needs one selector $g^\star(x) \in A^\star(x)$, the lower bound covers the case where $A^\star(x)$ contains several isolated optimal actions. Thus, near deterministic optima, the drift Lipschitz budget K is the expressivity bottleneck for diffusion policies.*

Remark B.2 (Bellman gap assumption in (3.3))**.** *Condition* (3.3) *is a mild finite-well condition on the Bellman gap. The clipping makes the assumption local in nature: near $A^\star(x)$, it requires the Bellman gap to grow at least quadratically with the distance to the optimal action set, while away from $A^\star(x)$, it only requires a positive suboptimality margin. It does not require concavity, uniqueness of the optimizer, or growth at infinity. The quadratic exponent is chosen to match the squared-distance lower bound in Proposition 3.3. More generally, a clipped p-th order gap would yield the corresponding $K^{-p/2}$ rate; for example, quadratic wells give K^{-1}, while quartic wells give K^{-2}.*

Remark B.3 (A note on policy initialization)**.** *We note that the lower bound over all $t > 0$ in Theorem 3.4 cannot hold under arbitrary action initialization μ_0. A counterexample is given in Appendix B.7. However, if we allow for burn-in time $t_0 \gtrsim 1/K$, the conclusion of Theorem 3.4 would still hold regardless of initialization. We state this extension in Corollary B.3 at Appendix B.7.*

B.2 Proof of Lemma 3.1

Let $c \in \mathbb{R}^d$ denote the target action and define, as shorthand,

$$z_K := z_K(c, \tau) = \frac{c}{1 - e^{-K\tau}}.$$

as c, τ do not change in this proof. Consider

$$d\bar{a}_s^x = K(z_K - \bar{a}_s^x)\, ds + \sigma(s, \bar{a}_s^x, x)\, dB_s, \qquad \bar{a}_0^x \sim \mu_0.$$

Define the moving center

$$m_s := (1 - e^{-Ks}) z_K, \qquad 0 \le s \le \tau.$$

Then $m_0 = 0$, $m_\tau = c$, and $\dot{m}_s = K(z_K - m_s)$. Set $Y_s := \bar{a}_s^x - m_s$, so that

$$\begin{aligned} dY_s &= d\bar{a}_s^x - \dot{m}_s\, ds \\ &= K(z_K - \bar{a}_s^x)\, ds - K(z_K - m_s)\, ds + \sigma(s, \bar{a}_s^x, x)\, dB_s \\ &= -K(\bar{a}_s^x - m_s)\, ds + \sigma(s, \bar{a}_s^x, x)\, dB_s \\ &= -KY_s\, ds + \sigma(s, \bar{a}_s^x, x)\, dB_s. \end{aligned}$$

Moreover, $Y_0 = \bar{a}_0^x$ and $Y_\tau = \bar{a}_\tau^x - c$. By Itô's formula,

$$d\|Y_s\|^2 = \big(-2K\|Y_s\|^2 + \mathrm{tr}\big(\sigma(s, \bar{a}_s^x, x)\sigma(s, \bar{a}_s^x, x)^\top\big)\big)\, ds + 2\langle Y_s, \sigma(s, \bar{a}_s^x, x)\, dB_s\rangle.$$

Under Assumption 2.1,

$$\mathrm{tr}\big(\sigma(s, \bar{a}_s^x, x)\sigma(s, \bar{a}_s^x, x)^\top\big) \le d\Lambda.$$

Taking expectations gives

$$\frac{d}{ds}\mathbb{E}\|Y_s\|^2 \le -2K\mathbb{E}\|Y_s\|^2 + d\Lambda.$$

Gronwall's inequality yields

$$\begin{aligned} \mathbb{E}\|Y_\tau\|^2 &\le e^{-2K\tau}\mathbb{E}\|Y_0\|^2 + d\Lambda \int_0^\tau e^{-2K(\tau - r)}\, dr \\ &= e^{-2K\tau} m_2(\mu_0) + \frac{d\Lambda}{2K}\big(1 - e^{-2K\tau}\big). \end{aligned}$$

Since $Y_\tau = \bar{a}_\tau^x - c$, we obtain

$$\mathbb{E}\|\bar{a}_\tau^x - c\|^2 \le e^{-2K\tau} m_2(\mu_0) + \frac{d\Lambda}{2K}\big(1 - e^{-2K\tau}\big).$$

This proves the claim.

B.3 Proof of Theorem 3.2

We now prove the value-approximation theorem stated in Theorem 3.2. The proof lifts the terminal-action concentration from Lemma 3.1 to convergence of the induced one-step reward and transition kernel, and then propagates this convergence through the discounted Bellman recursion.

Proof. Below, we allow the volatility to depend on K and write it as σ_K; if the volatility is fixed, simply read $\sigma_K = \sigma$. We prove the result by converting the controlled MDP into a Markov reward process. The main point is that, when the volatility is state-dependent, the terminal noise is no longer independent of x. We therefore use only the concentration of the terminal action around $g^\star(x)$.

From Lemma 3.1, we have

$$\mathbb{E}\|\bar{a}_T^{K,x} - g^\star(x)\|^2 \le e^{-2KT} m_2(\mu_0) + \frac{d\Lambda}{2K}\big(1 - e^{-2KT}\big) =: \varepsilon_K.$$

Since $\varepsilon_K \to 0$, we have, for every fixed x,

$$\Delta_K(x) := \|\bar{a}_T^{K,x} - g^\star(x)\| \longrightarrow 0 \qquad \text{in } L^2 \text{ and hence in probability.}$$

Define the averaged reward and transition kernel induced by the diffusion policy:

$$r_K(x) := \mathbb{E}\left[r(x, \bar{a}_T^{K,x})\right],$$

and

$$P_K(A \mid x) := \mathbb{E}\left[P(A \mid x, \bar{a}_T^{K,x})\right], \qquad A \in \mathcal{B}(\mathcal{X}).$$

For the deterministic policy $\pi^\star(\cdot \mid x) = \delta_{g^\star(x)}$, define

$$r^\star(x) := r(x, g^\star(x)), \qquad P^\star(A \mid x) := P(A \mid x, g^\star(x)).$$

We first show pointwise convergence of the one-step reward and transition kernel. Fix $x \in \mathcal{X}$, and let

$$A_{K,x} := \{\Delta_K(x) \le R_x\},$$

where R_x is the radius from Assumption 3.1. Recall $r_{\max} := \sup_{x,a} |r(x,a)| < \infty$. On $A_{K,x}$, Assumption 3.1 gives

$$|r(x, \bar{a}_T^{K,x}) - r(x, g^\star(x))| \le \omega_{r,x}(\Delta_K(x)).$$

On $A_{K,x}^c$, we use the trivial bound $|r(x, \bar{a}_T^{K,x}) - r(x, g^\star(x))| \le 2r_{\max}$. Hence

$$|r_K(x) - r^\star(x)| \le \mathbb{E}\big[\omega_{r,x}(\Delta_K(x))\mathbf{1}_{A_{K,x}}\big] + 2r_{\max}\mathbb{P}(A_{K,x}^c).$$

Since $\Delta_K(x) \to 0$ in probability, $\mathbb{P}(A_{K,x}^c) = \mathbb{P}(\Delta_K(x) > R_x) \to 0$. Moreover, taking the modulus to be nondecreasing, for any $0 < \eta < R_x$,

$$\mathbb{E}\big[\omega_{r,x}(\Delta_K(x))\mathbf{1}_{A_{K,x}}\big] \le \omega_{r,x}(\eta) + \omega_{r,x}(R_x)\mathbb{P}(\Delta_K(x) > \eta).$$

Letting $K \to \infty$ and then $\eta \downarrow 0$ gives $r_K(x) \to r^\star(x)$.

Similarly, by convexity of total variation under mixtures,

$$\begin{aligned}\mathrm{TV}(P_K(\cdot \mid x), P^\star(\cdot \mid x)) &\le \mathbb{E}\left[\mathrm{TV}\Big(P(\cdot \mid x, \bar{a}_T^{K,x}), P(\cdot \mid x, g^\star(x))\Big)\right] \\ &\le \mathbb{E}\big[\omega_{P,x}(\Delta_K(x))\mathbf{1}_{A_{K,x}}\big] + 2\mathbb{P}(A_{K,x}^c).\end{aligned}$$

The same argument as above yields

$$\mathrm{TV}(P_K(\cdot \mid x), P^\star(\cdot \mid x)) \to 0.$$

Now define the finite-horizon value functions for the two Markov reward processes. Set $V_0^K = V_0^\star = 0$, and recursively define

$$V_{n+1}^K(x) = r_K(x) + \gamma \int_{\mathcal{X}} V_n^K(y) P_K(dy \mid x),$$

and

$$V_{n+1}^\star(x) = r^\star(x) + \gamma \int_{\mathcal{X}} V_n^\star(y) P^\star(dy \mid x).$$

We prove by induction that, for every fixed $n \geq 0$ and every fixed $x \in \mathcal{X}$,

$$V_n^K(x) \to V_n^\star(x).$$

The case $n = 0$ is immediate. Suppose the claim holds for n. Since rewards are bounded,

$$\sup_K \|V_n^K\|_\infty \vee \|V_n^\star\|_\infty \leq \frac{r_{\max}}{1-\gamma}.$$

For fixed x,

$$\begin{aligned} & \left| \int V_n^K(y) P_K(dy \mid x) - \int V_n^\star(y) P^\star(dy \mid x) \right| \\ \leq \; & \left| \int \big(V_n^K(y) - V_n^\star(y)\big) P^\star(dy \mid x) \right| + 2\|V_n^K\|_\infty \, \mathrm{TV}(P_K(\cdot \mid x), P^\star(\cdot \mid x)) \, . \end{aligned}$$

The first term goes to zero by dominated convergence under the fixed measure $P^\star(\cdot \mid x)$. The second term goes to zero by the total-variation convergence shown above. Together with $r_K(x) \to r^\star(x)$, this gives

$$V_{n+1}^K(x) \to V_{n+1}^\star(x).$$

The induction is complete.

It remains to pass from finite horizon to infinite horizon. Let $V^K := V^{\pi_K}$ and $V^\star := V^{\pi^\star}$. For every $n \geq 0$,

$$\|V^K - V_n^K\|_\infty \vee \|V^\star - V_n^\star\|_\infty \leq \frac{\gamma^n r_{\max}}{1-\gamma}.$$

Therefore, for every fixed x,

$$\begin{aligned} \limsup_{K\to\infty} |V^K(x) - V^\star(x)| &\leq \frac{2\gamma^n r_{\max}}{1-\gamma} + \limsup_{K\to\infty} |V_n^K(x) - V_n^\star(x)| \\ &= \frac{2\gamma^n r_{\max}}{1-\gamma}. \end{aligned}$$

Letting $n \to \infty$ yields

$$V^K(x) \to V^\star(x) \qquad \text{for every } x \in \mathcal{X}.$$

Finally,

$$|V^K(x) - V^\star(x)| \leq \frac{2r_{\max}}{1-\gamma},$$

so dominated convergence with respect to the initial state law ρ gives

$$V^{\pi_K}(\rho) = \int_{\mathcal{X}} V^K(x)\rho(dx) \longrightarrow \int_{\mathcal{X}} V^\star(x)\rho(dx) = V^{\pi^\star}(\rho).$$

Since $\pi^\star(\cdot \mid x) = \delta_{g^\star(x)}$ is an optimal deterministic policy, $V^{\pi^\star}(\rho) = V^\star(\rho)$. This proves the theorem. □

B.4 Smoothing and anti-concentration for state-independent volatility

This subsection proves the smoothing estimate needed for the final-window lower-bound argument in Proposition 3.3. The goal is to show that, after rescaling the last time interval of length K^{-1}, the residual diffusion has an endpoint density uniformly bounded from above. This prevents the terminal action from concentrating too much near any finite set of target actions.

We proceed in two steps. Lemma B.1 first gives an a priori L^2 estimate for solutions of the Fokker-Planck equation. Lemma B.2 then combines this estimate with Nash's inequality and a duality argument to obtain a uniform L^∞ bound on the density at the terminal time. This L^∞ bound implies the anti-concentration estimate used in Appendix B.5.

Lemma B.1 (A priori L^2 finiteness for SDE marginal density under state-independent volatility)**.** *Fix $0 \le r < t \le 1$. Let*

$$\Gamma(s) := \sigma(s)\sigma(s)^\top, \qquad \Gamma_{ij}(s) = [\Gamma(s)]_{ij}.$$

Assume that $\beta(s,\cdot)$ is smooth and 1-Lipschitz uniformly in s, that $\beta(\cdot,0) \in L^1([r,t])$, and that $\Gamma(s)$ is deterministic, positive semidefinite, and bounded on $[r,t]$. Let p_s be a nonnegative probability density solving

$$\partial_s p_s(x) = -\sum_{i=1}^d \partial_i\big(\beta_i(s,x)p_s(x)\big) + \frac{1}{2}\sum_{i,j=1}^d \Gamma_{ij}(s)\partial_{ij}p_s(x), \qquad p_r = \phi,$$

where $\phi \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ is a probability density. Then, for every $u \in [r,t]$, it holds that

$$\|p_u\|_2^2 \le e^{d(u-r)}\|\phi\|_2^2.$$

In particular, $p_u \in L^2(\mathbb{R}^d)$ for every $u \in [r,t]$.

Proof. For $M > 0$, define the truncated quadratic function

$$\Psi_M(z) = \begin{cases} z^2, & 0 \le z \le M, \\ 2Mz - M^2, & z > M. \end{cases}$$

Then Ψ_M is convex, C^1, and satisfies $0 \le \Psi_M(z) \le 2Mz$, $\Psi_M(z) \le z^2$, and $\Psi_M(z) \uparrow z^2$ as $M \to \infty$.

Also define

$$H_M(z) := z\Psi_M'(z) - \Psi_M(z).$$

A direct computation gives

$$H_M(z) = z^2 \wedge M^2, \qquad 0 \le H_M(z) \le \Psi_M(z).$$

Let $\chi_R \in C_c^\infty(\mathbb{R}^d)$ be a standard cutoff such that

$$0 \le \chi_R \le 1, \qquad \chi_R = 1 \text{ on } B_R, \qquad \operatorname{supp}\chi_R \subset B_{2R},$$

and

$$|\nabla\chi_R| \le \frac{C}{R}, \qquad |D^2\chi_R| \le \frac{C}{R^2}.$$

Set

$$F_{M,R}(s) := \int_{\mathbb{R}^d} \Psi_M(p_s(x))\chi_R(x)\,dx.$$

This quantity is finite because $\Psi_M(p_s) \le 2Mp_s$ and p_s has total mass one.

We differentiate $F_{M,R}(s)$ and use the Fokker–Planck equation:

$$\frac{d}{ds}F_{M,R}(s) = \int \Psi_M'(p_s)\chi_R\,\partial_s p_s.$$

For the drift term,

$$-\int \Psi_M'(p_s)\chi_R \sum_{i=1}^d \partial_i(\beta_i p_s) = -\int (\nabla\cdot\beta_s)H_M(p_s)\chi_R + \int \Psi_M(p_s)\beta_s\cdot\nabla\chi_R.$$

For the diffusion term, since $\Gamma_{ij}(s)$ does not depend on x,

$$\begin{aligned} &\frac{1}{2}\int \Psi_M'(p_s)\chi_R \sum_{i,j=1}^d \Gamma_{ij}(s)\partial_{ij}p_s \\ &\quad = -\frac{1}{2}\int \Psi_M''(p_s)\chi_R \sum_{i,j=1}^d \Gamma_{ij}(s)\partial_i p_s\,\partial_j p_s \\ &\qquad + \frac{1}{2}\int \Psi_M(p_s)\sum_{i,j=1}^d \Gamma_{ij}(s)\partial_{ij}\chi_R. \end{aligned}$$

The first diffusion term is nonpositive because Ψ_M is convex and $\Gamma(s)$ is positive semidefinite. Since β_s is 1-Lipschitz,

$$-\nabla \cdot \beta_s \le d.$$

Therefore,

$$\frac{d}{ds}F_{M,R}(s) \le dF_{M,R}(s) + E_{M,R}(s),$$

where

$$E_{M,R}(s) := \int \Psi_M(p_s)\beta_s \cdot \nabla\chi_R + \frac{1}{2}\int \Psi_M(p_s)\sum_{i,j=1}^{d}\Gamma_{ij}(s)\partial_{ij}\chi_R.$$

We now send $R \to \infty$. Let

$$A_R := \{x \in \mathbb{R}^d : R \le |x| \le 2R\}.$$

The support of $\nabla\chi_R$ and $D^2\chi_R$ is contained in A_R. Since β_s is 1-Lipschitz,

$$|\beta(s,x)| \le |\beta(s,0)| + |x|.$$

Using $\Psi_M(p_s) \le 2Mp_s$, the cutoff bounds, and boundedness of $\Gamma(s)$, we get

$$|E_{M,R}(s)| \le CM\big(1 + |\beta(s,0)| + \|\Gamma(s)\|_{\mathrm{op}}\big)\int_{A_R} p_s(x)\,dx.$$

For each fixed s, $\int_{A_R} p_s(x)\,dx \to 0$ as $R \to \infty$, because p_s is a probability density. The right-hand side is dominated by an integrable function of s on $[r,t]$. Hence, by dominated convergence,

$$\int_r^u |E_{M,R}(s)|\,ds \to 0 \qquad \text{as } R \to \infty.$$

Integrating the differential inequality from r to u gives

$$F_{M,R}(u) \le F_{M,R}(r) + d\int_r^u F_{M,R}(s)\,ds + \int_r^u E_{M,R}(s)\,ds.$$

Letting $R \to \infty$ yields

$$F_M(u) \le F_M(r) + d\int_r^u F_M(s)\,ds,$$

where

$$F_M(s) := \int_{\mathbb{R}^d} \Psi_M(p_s(x))\,dx.$$

By Gronwall's inequality,

$$F_M(u) \le e^{d(u-r)}F_M(r).$$

At the initial time, $p_r = \phi$, and $\Psi_M(\phi) \le \phi^2$. Therefore,

$$F_M(u) \le e^{d(u-r)}\|\phi\|_2^2.$$

Finally, since $\Psi_M(z) \uparrow z^2$, monotone convergence gives

$$\int_{\mathbb{R}^d} p_u(x)^2\,dx = \lim_{M\to\infty}\int_{\mathbb{R}^d}\Psi_M(p_u(x))\,dx \le e^{d(u-r)}\|\phi\|_2^2.$$

This proves the claim.

The computation with Ψ_M can be made fully classical by first replacing Ψ_M by smooth convex approximations with the same bounds and then letting the approximation parameter tend to zero. □

Lemma B.2. *Let $(Y_t)_{0\le t\le 1}$ solve*

$$dY_t = \beta(t,Y_t)\,dt + \sigma(t)\,dW_t, \qquad Y_0 = y \in \mathbb{R}^d.$$

Assume that $\beta(t,\cdot)$ is 1-Lipschitz uniformly in t, $\beta(\cdot,0) \in L^1([0,1])$, and that $\sigma(t)$ is deterministic, state-independent, and square-integrable on $[0,1]$. Assume also that there is $\kappa > 0$ such that

$$\sigma(t)\sigma(t)^\top \succeq \kappa I_d$$

for a.e. $t \in [0,1]$.

Then the law of Y_1 *has a density* p_1^y *with respect to Lebesgue measure, and there is a constant* $C_{d,\kappa} < \infty$*, depending only on* d *and* κ*, such that*

$$\sup_{y \in \mathbb{R}^d} \|p_1^y\|_\infty \le C_{d,\kappa}.$$

Consequently, for any $z_1, \dots, z_m \in \mathbb{R}^d$ *and any* $\rho > 0$*,*

$$\mathbb{P}_y \left(Y_1 \in \bigcup_{j=1}^m B(z_j, \rho) \right) \le C_{d,\kappa} m v_d \rho^d,$$

where $v_d = \lambda_d(B(0,1))$*, with* λ_d *the Lebesgue measure in* $\mathbb{R}^d$.

Proof. We first prove the result under the additional assumptions that β is smooth in the spatial variable, σ is bounded, and the initial law has a smooth probability density. We then remove these assumptions by approximation. All the integration by parts in Steps 1-2 can be justified by inserting standard cut-offs χ_R as in Lemma B.1 and then let $R \to \infty$.

Step 1: forward $L^1 \to L^2$ smoothing for smooth probability densities.

Fix $0 \le r < t \le 1$. Suppose the law of Y_r has smooth (compactly supported) probability density φ. Let p_t be the density of Y_t. We write

$$p_t = P_{r,t}^* \varphi,$$

where $P_{r,t}^*$ denotes the forward Fokker–Planck evolution of densities. Since φ is a probability density, p_t is also a probability density:

$$p_t \ge 0, \qquad \int_{\mathbb{R}^d} p_t(x)\, dx = 1.$$

Furthermore, Lemma B.1 ensures that $p_t \in L^2(\mathbb{R}^d)$ for $t > r$. The Fokker–Planck equation is

$$\partial_t p_t = -\nabla \cdot (\beta_t p_t) + \frac{1}{2} \sum_{i,j=1}^d [\sigma(t)\sigma(t)^\top]_{ij} \partial_{ij} p_t,$$

where $\beta_t(x) = \beta(t,x)$.

We compute the evolution of

$$\|p_t\|_2^2 = \int_{\mathbb{R}^d} p_t(x)^2\, dx.$$

Multiplying the Fokker–Planck equation by $2p_t$ and integrating gives

$$\frac{d}{dt}\|p_t\|_2^2 = -2 \int p_t \nabla \cdot (\beta_t p_t) + \int p_t \sum_{i,j=1}^d [\sigma(t)\sigma(t)^\top]_{ij} \partial_{ij} p_t.$$

For the drift term, use

$$\nabla \cdot (\beta_t p_t) = (\nabla \cdot \beta_t) p_t + \beta_t \cdot \nabla p_t.$$

Then

$$-2 \int p_t \nabla \cdot (\beta_t p_t) = -2 \int (\nabla \cdot \beta_t) p_t^2 - 2 \int p_t \beta_t \cdot \nabla p_t.$$

Since $2p_t \beta_t \cdot \nabla p_t = \beta_t \cdot \nabla (p_t^2)$, integration by parts gives

$$-2 \int p_t \beta_t \cdot \nabla p_t = -\int \beta_t \cdot \nabla (p_t^2) = \int (\nabla \cdot \beta_t) p_t^2.$$

Therefore

$$-2 \int p_t \nabla \cdot (\beta_t p_t) = -\int (\nabla \cdot \beta_t) p_t^2.$$

For the diffusion term, $\sigma(t)\sigma(t)^\top$ is independent of x, so integration by parts gives

$$\int p_t \sum_{i,j=1}^{d} [\sigma(t)\sigma(t)^\top]_{ij} \partial_{ij} p_t = -\int \nabla p_t^\top \sigma(t)\sigma(t)^\top \nabla p_t.$$

Combining the two identities,

$$\frac{d}{dt}\|p_t\|_2^2 = -\int (\nabla \cdot \beta_t) p_t^2 - \int \nabla p_t^\top \sigma(t)\sigma(t)^\top \nabla p_t.$$

Since β_t is 1-Lipschitz, we have $-\nabla \cdot \beta_t \le d$. And since $\sigma(t)\sigma(t)^\top \succeq \kappa I_d$, we have

$$\nabla p_t^\top \sigma(t)\sigma(t)^\top \nabla p_t \ge \kappa \|\nabla p_t\|^2.$$

Hence

$$\frac{d}{dt}\|p_t\|_2^2 \le d\|p_t\|_2^2 - \kappa\|\nabla p_t\|_2^2.$$

Define $q_t := e^{-d(t-r)/2} p_t$. Then $\|q_t\|_2^2 = e^{-d(t-r)}\|p_t\|_2^2$. Differentiating it gives

$$\frac{d}{dt}\|q_t\|_2^2 = -de^{-d(t-r)}\|p_t\|_2^2 + e^{-d(t-r)}\frac{d}{dt}\|p_t\|_2^2.$$

Using the preceding inequality,

$$\frac{d}{dt}\|q_t\|_2^2 \le -de^{-d(t-r)}\|p_t\|_2^2 + e^{-d(t-r)}\left(d\|p_t\|_2^2 - \kappa\|\nabla p_t\|_2^2\right).$$

The two terms involving $d\|p_t\|_2^2$ cancel, so

$$\frac{d}{dt}\|q_t\|_2^2 \le -\kappa e^{-d(t-r)}\|\nabla p_t\|_2^2.$$

Because $q_t = e^{-d(t-r)/2} p_t$ and the exponential factor is independent of x, we have $\nabla q_t = e^{-d(t-r)/2}\nabla p_t$, and therefore

$$\|\nabla q_t\|_2^2 = e^{-d(t-r)}\|\nabla p_t\|_2^2.$$

Thus

$$\frac{d}{dt}\|q_t\|_2^2 \le -\kappa\|\nabla q_t\|_2^2.$$

Since p_t is a probability density,

$$\|q_t\|_1 = e^{-d(t-r)/2}\|p_t\|_1 = e^{-d(t-r)/2} \le 1.$$

We now apply Nash's inequality on $\mathbb{R}^d$: there exists $N_d < \infty$, depending only on d, such that for every $u \in L^1(\mathbb{R}^d) \cap H^1(\mathbb{R}^d)$,

$$\|u\|_2^{2+4/d} \le N_d \|\nabla u\|_2^2 \|u\|_1^{4/d}.$$

Applying this to $u = q_t$ gives

$$\|q_t\|_2^{2+4/d} \le N_d\|\nabla q_t\|_2^2\|q_t\|_1^{4/d} \le N_d\|\nabla q_t\|_2^2.$$

Therefore

$$\|\nabla q_t\|_2^2 \ge \frac{1}{N_d}\|q_t\|_2^{2+4/d}.$$

Combining with the energy inequality,

$$\frac{d}{dt}\|q_t\|_2^2 \le -\frac{\kappa}{N_d}\|q_t\|_2^{2+4/d}.$$

Set $F(t) := \|q_t\|_2^2$. Then $\|q_t\|_2^{2+4/d} = F(t)^{1+2/d}$. Thus

$$F'(t) \le -\frac{\kappa}{N_d}F(t)^{1+2/d}.$$

Let $\theta = 2/d$ and $c_{d,\kappa} = \kappa/N_d$, where N_d is the coefficient in Nash's inequality. Then

$$F'(t) \le -c_{d,\kappa} F(t)^{1+\theta}.$$

If $F(t) = 0$, the desired bound is trivial. Otherwise,

$$\frac{d}{dt} F(t)^{-\theta} = -\theta F(t)^{-\theta-1} F'(t) \ge \theta c_{d,\kappa}.$$

Integrating from r to t gives

$$F(t)^{-\theta} - F(r)^{-\theta} \ge \theta c_{d,\kappa}(t-r).$$

Since $F(r)^{-\theta} \ge 0$, we have $F(t)^{-\theta} \ge \theta c_{d,\kappa}(t-r)$. Therefore

$$F(t) \le (\theta c_{d,\kappa}(t-r))^{-1/\theta}.$$

Substituting $\theta = 2/d$ and $c_{d,\kappa} = \kappa/N_d$, we have $F(t) \le \left(\frac{2\kappa}{d\,N_d}(t-r)\right)^{-d/2}$. Hence

$$\|q_t\|_2 \le \left(\frac{d\,N_d}{2\kappa}\right)^{d/4} (t-r)^{-d/4}.$$

Since $p_t = e^{d(t-r)/2} q_t$ and $t - r \le 1$,

$$\|p_t\|_2 \le C_{d,\kappa}(t-r)^{-d/4}.$$

This proves the forward smoothing estimate.

Step 2: a backward estimate and the L^∞ bound.

We now obtain an $L^2 \to L^\infty$ estimate for the forward evolution. Let $P_{r,t}$ denote the backward Markov semigroup:

$$(P_{r,t} g)(x) = \mathbb{E}[g(Y_t) \mid Y_r = x].$$

For this step, g is a test function. It need not be a density.

Fix $0 \le r < t \le 1$ and first take $g \ge 0$. Define

$$w_s(x) = P_{t-s,t} g(x), \qquad 0 \le s \le t - r.$$

Then $w_0 = g$, and w_s satisfies the backward equation, written forward in the variable s:

$$\partial_s w_s = \beta(t-s,x) \cdot \nabla w_s + \frac{1}{2} \sum_{i,j=1}^{d} [\sigma(t-s)\sigma(t-s)^\top]_{ij} \partial_{ij} w_s.$$

Here

$$\beta(t-s,x) \cdot \nabla w_s(x) = \sum_{i=1}^{d} \beta_i(t-s,x) \partial_{x_i} w_s(x).$$

Repeating the same L^2 integration by parts as in Step 1 gives

$$\frac{d}{ds} \|w_s\|_2^2 \le d\|w_s\|_2^2 - \kappa \|\nabla w_s\|_2^2.$$

We also need an L^1 bound on w_s. Since $g \ge 0$ and $P_{t-s,t}$ preserves nonnegativity, we have $w_s \ge 0$. Hence $\|w_s\|_1 = \int w_s$. Integrating the backward equation in x gives

$$\frac{d}{ds} \|w_s\|_1 = \int \beta(t-s,x) \cdot \nabla w_s(x)\, dx.$$

The second-order term integrates to zero. By integration by parts,

$$\int \beta(t-s,x) \cdot \nabla w_s(x)\, dx = -\int (\nabla \cdot \beta)(t-s,x) w_s(x)\, dx.$$

Since $-\nabla \cdot \beta \le d$ and $w_s \ge 0$,

$$\frac{d}{ds}\|w_s\|_1 \le d\|w_s\|_1.$$

By Gronwall's inequality,

$$\|w_s\|_1 \le e^{ds}\|w_0\|_1 = e^{ds}\|g\|_1.$$

Define $v_s = e^{-ds/2}w_s$. Exactly as in Step 1,

$$\frac{d}{ds}\|v_s\|_2^2 \le -\kappa\|\nabla v_s\|_2^2.$$

Moreover,

$$\|v_s\|_1 = e^{-ds/2}\|w_s\|_1 \le e^{-ds/2}e^{ds}\|g\|_1 = e^{ds/2}\|g\|_1 \le e^{d/2}\|g\|_1.$$

Applying Nash's inequality to v_s gives

$$\|v_s\|_2^{2+4/d} \le N_d\|\nabla v_s\|_2^2\|v_s\|_1^{4/d} \le N_d e^2\|\nabla v_s\|_2^2\|g\|_1^{4/d}.$$

Therefore,

$$\|\nabla v_s\|_2^2 \ge \frac{1}{N_d e^2\|g\|_1^{4/d}}\|v_s\|_2^{2+4/d}.$$

Repeating the differential inequality argument from Step 1 yields

$$\|v_s\|_2 \le C_{d,\kappa}s^{-d/4}\|g\|_1.$$

Since $w_s = e^{ds/2}v_s$ and $s \le 1$,

$$\|w_s\|_2 \le C_{d,\kappa}s^{-d/4}\|g\|_1.$$

Taking $s = t - r$ gives

$$\|P_{r,t}g\|_2 \le C_{d,\kappa}(t-r)^{-d/4}\|g\|_1.$$

For signed g, apply this estimate to g_+ and g_- and use linearity. Thus the same estimate holds for all $g \in L^1$.

Now use duality. For $h \in L^2(\mathbb{R}^d)$ and $g \in C_c^\infty(\mathbb{R}^d)$ with $\|g\|_1 \le 1$, adjointness gives

$$\left|\int_{\mathbb{R}^d} P_{r,t}^* h(x)g(x)\,dx\right| = \left|\int_{\mathbb{R}^d} h(x)P_{r,t}g(x)\,dx\right|.$$

Hence, by Cauchy–Schwarz and the backward estimate,

$$\left|\int P_{r,t}^* h\, g\right| \le \|h\|_2\|P_{r,t}g\|_2 \le C_{d,\kappa}(t-r)^{-d/4}\|h\|_2\|g\|_1.$$

Taking the supremum over $\|g\|_1 \le 1$ gives

$$\|P_{r,t}^* h\|_\infty \le C_{d,\kappa}(t-r)^{-d/4}\|h\|_2.$$

Apply this over two half intervals. If the initial law at time 0 has smooth probability density φ, then

$$p_1 = P_{1/2,1}^* P_{0,1/2}^* \varphi.$$

By Step 1,

$$\|P_{0,1/2}^*\varphi\|_2 \le C_{d,\kappa}.$$

By the $L^2 \to L^\infty$ estimate,

$$\|p_1\|_\infty = \|P_{1/2,1}^*(P_{0,1/2}^*\varphi)\|_\infty \le C_{d,\kappa}\|P_{0,1/2}^*\varphi\|_2 \le C_{d,\kappa}.$$

Step 3: point-mass initial condition.

We now keep the smooth drift and bounded volatility assumptions, but start from $Y_0 = y$.

Let $\eta \in C_c^\infty(\mathbb{R}^d)$ be such that $\eta \ge 0$ and $\int \eta = 1$. Define

$$\varphi_n(x) = n^d\eta(n(x-y)).$$

Then each φ_n is a smooth probability density and $\varphi_n(x)\,dx \Rightarrow \delta_y$ weakly.

Let X_n have density φ_n. We may write $X_n = y + Z/n$, where Z has density η, so $X_n \to y$ almost surely. Let $Y_t^{X_n}$ and Y_t^y be solutions driven by the same Brownian motion, started from X_n and y. Because the volatility is state-independent,

$$d(Y_t^{X_n} - Y_t^y) = \Big[\beta(t, Y_t^{X_n}) - \beta(t, Y_t^y)\Big]\,dt.$$

Since $\beta(t,\cdot)$ is 1-Lipschitz,

$$|Y_1^{X_n} - Y_1^y| \le e|X_n - y|.$$

Hence $Y_1^{X_n} \to Y_1^y$ almost surely, and therefore in distribution.

For each n, Step 2 gives

$$\mathbb{P}(Y_1^{X_n} \in E) \le C_{d,\kappa}\lambda_d(E)$$

for every Borel set E. Let O be open. By the portmanteau theorem,

$$\mathbb{P}(Y_1^y \in O) \le \liminf_{n\to\infty} \mathbb{P}(Y_1^{X_n} \in O) \le C_{d,\kappa}\lambda_d(O).$$

For a Borel set E, choose open $O \supset E$ with $\lambda_d(O) \le \lambda_d(E) + \varepsilon$. Then

$$\mathbb{P}(Y_1^y \in E) \le \mathbb{P}(Y_1^y \in O) \le C_{d,\kappa}(\lambda_d(E) + \varepsilon).$$

Letting $\varepsilon \downarrow 0$ gives

$$\mathbb{P}(Y_1^y \in E) \le C_{d,\kappa}\lambda_d(E).$$

Thus the law of Y_1^y has density bounded by $C_{d,\kappa}$.

Step 4: remove smoothness of β and boundedness of σ.

We now remove the two auxiliary assumptions used in Steps 1–3. The final statement only assumes that $\beta(t,\cdot)$ is 1-Lipschitz and that $\sigma(t)\sigma(t)^\top \succeq \kappa I_d$.

Step 4 (a): remove smoothness of β, keeping σ bounded.

Let $\eta \in C_c^\infty(\mathbb{R}^d)$ be a standard mollifier: $\eta \ge 0$, $\int_{\mathbb{R}^d} \eta(z)\,dz = 1$, and $\operatorname{supp}\eta \subset B(0,1)$. Define

$$\eta_n(z) = n^d \eta(nz).$$

For each n, define the spatial mollification of β by

$$\beta_n(t,x) = (\eta_n *_x \beta(t,\cdot))(x) = \int_{\mathbb{R}^d} \eta_n(z)\beta(t, x-z)\,dz.$$

The convolution is only in the spatial variable x, not in time.

For each fixed t, the function $\beta_n(t,\cdot)$ is smooth. Moreover, $\beta_n(t,\cdot)$ is still 1-Lipschitz. Indeed, for $x, x' \in \mathbb{R}^d$,

$$\begin{aligned}
|\beta_n(t,x) - \beta_n(t,x')| &= \left|\int_{\mathbb{R}^d} \eta_n(z)\,[\beta(t,x-z) - \beta(t,x'-z)]\,dz\right| \\
&\le \int_{\mathbb{R}^d} \eta_n(z)|\beta(t,x-z) - \beta(t,x'-z)|\,dz \\
&\le \int_{\mathbb{R}^d} \eta_n(z)|x - x'|\,dz \\
&= |x - x'|.
\end{aligned}$$

Thus the Lipschitz constant is not enlarged by mollification.

Also, β_n converges uniformly to β. Since $\beta(t,\cdot)$ is 1-Lipschitz uniformly in t,

$$\begin{aligned}
|\beta_n(t,x) - \beta(t,x)| &= \left|\int_{\mathbb{R}^d} \eta_n(z)\,[\beta(t,x-z) - \beta(t,x)]\,dz\right| \\
&\le \int_{\mathbb{R}^d} \eta_n(z)|z|\,dz.
\end{aligned}$$

Changing variables $u = nz$ gives

$$\int_{\mathbb{R}^d} \eta_n(z)|z|\,dz = \frac{1}{n}\int_{\mathbb{R}^d} \eta(u)|u|\,du.$$

Hence, with

$$C_\eta := \int_{\mathbb{R}^d} \eta(u)|u|\,du,$$

we have

$$\sup_{t\in[0,1]}\sup_{x\in\mathbb{R}^d} |\beta_n(t,x) - \beta(t,x)| \le \frac{C_\eta}{n}.$$

Let $Y^{(n)}$ solve

$$dY_t^{(n)} = \beta_n(t, Y_t^{(n)})\,dt + \sigma(t)\,dW_t, \qquad Y_0^{(n)} = y,$$

and let Y solve

$$dY_t = \beta(t, Y_t)\,dt + \sigma(t)\,dW_t, \qquad Y_0 = y.$$

Drive both equations by the same Brownian motion. Since the volatility is the same and is state-independent,

$$Y_t^{(n)} - Y_t = \int_0^t \Big[\beta_n(s, Y_s^{(n)}) - \beta(s, Y_s)\Big]\,ds.$$

Therefore

$$\begin{aligned}|Y_t^{(n)} - Y_t| &\le \int_0^t |\beta_n(s, Y_s^{(n)}) - \beta_n(s, Y_s)|\,ds + \int_0^t |\beta_n(s, Y_s) - \beta(s, Y_s)|\,ds \\ &\le \int_0^t |Y_s^{(n)} - Y_s|\,ds + \frac{C_\eta t}{n}.\end{aligned}$$

By Gronwall's inequality,

$$\sup_{0\le t\le 1} |Y_t^{(n)} - Y_t| \le \frac{C_\eta e}{n}.$$

In particular, $Y_1^{(n)} \to Y_1$ almost surely, hence also in distribution.

By Steps 1–3, applied to the smooth drift β_n, the endpoint law of $Y_1^{(n)}$ satisfies

$$\mathbb{P}(Y_1^{(n)} \in E) \le C_{d,\kappa}\lambda_d(E)$$

for every Borel set E, with the same constant $C_{d,\kappa}$ for all n. Let $O \subset \mathbb{R}^d$ be open. Since $Y_1^{(n)} \Rightarrow Y_1$, the portmanteau theorem gives

$$\mathbb{P}(Y_1 \in O) \le \liminf_{n\to\infty} \mathbb{P}(Y_1^{(n)} \in O) \le C_{d,\kappa}\lambda_d(O).$$

For a Borel set E, choose open $O \supset E$ with $\lambda_d(O) \le \lambda_d(E) + \varepsilon$. Then

$$\mathbb{P}(Y_1 \in E) \le \mathbb{P}(Y_1 \in O) \le C_{d,\kappa}(\lambda_d(E) + \varepsilon).$$

Letting $\varepsilon \downarrow 0$ yields

$$\mathbb{P}(Y_1 \in E) \le C_{d,\kappa}\lambda_d(E).$$

Thus the density bound holds for Lipschitz β, as long as σ is bounded.

Step 4 (b): remove boundedness of σ.

Now assume only that σ is square-integrable on $[0,1]$ and $\sigma(t)\sigma(t)^\top \succeq \kappa I_d$ for a.e. $t \in [0,1]$.

Set

$$\Gamma(t) := \sigma(t)\sigma(t)^\top.$$

Then $\Gamma(t)$ is symmetric positive definite for a.e. t, and $\Gamma(t) \succeq \kappa I_d$. Define

$$O(t) := \Gamma(t)^{-1/2}\sigma(t).$$

Then $O(t)O(t)^\top = \Gamma(t)^{-1/2}\sigma(t)\sigma(t)^\top\Gamma(t)^{-1/2} = I_d$, so $O(t)$ is orthogonal for a.e. t.

For $N \geq \kappa$, define $\Gamma_N(t)$ by truncating the eigenvalues of $\Gamma(t)$ above at N. More explicitly, if

$$\Gamma(t) = U(t)\,\mathrm{diag}(\lambda_1(t), \ldots, \lambda_d(t))U(t)^\top,$$

then

$$\Gamma_N(t) = U(t)\,\mathrm{diag}(\lambda_1(t) \wedge N, \ldots, \lambda_d(t) \wedge N)U(t)^\top.$$

Because each $\lambda_i(t) \geq \kappa$, we have $\Gamma_N(t) \succeq \kappa I_d$. Define

$$\sigma_N(t) := \Gamma_N(t)^{1/2} O(t).$$

Then $\sigma_N(t)\sigma_N(t)^\top = \Gamma_N(t)^{1/2} O(t) O(t)^\top \Gamma_N(t)^{1/2} = \Gamma_N(t)$. Therefore

$$\sigma_N(t)\sigma_N(t)^\top \succeq \kappa I_d.$$

Also, σ_N is bounded, since all eigenvalues of $\Gamma_N(t)$ are at most N.

We next show that $\sigma_N \to \sigma$ in $L^2([0,1])$. Since $\sigma(t) = \Gamma(t)^{1/2} O(t)$ and $\sigma_N(t) = \Gamma_N(t)^{1/2} O(t)$, orthogonality of $O(t)$ gives

$$\|\sigma_N(t) - \sigma(t)\|_{\mathrm{HS}} = \|\Gamma_N(t)^{1/2} - \Gamma(t)^{1/2}\|_{\mathrm{HS}}.$$

If $\lambda_1(t), \ldots, \lambda_d(t)$ are the eigenvalues of $\Gamma(t)$, then

$$\|\Gamma_N(t)^{1/2} - \Gamma(t)^{1/2}\|_{\mathrm{HS}}^2 = \sum_{i=1}^{d} \left(\sqrt{\lambda_i(t) \wedge N} - \sqrt{\lambda_i(t)}\right)^2.$$

For each fixed t, this converges to 0 as $N \to \infty$. Moreover,

$$\left(\sqrt{\lambda_i(t) \wedge N} - \sqrt{\lambda_i(t)}\right)^2 \leq \lambda_i(t).$$

Hence

$$\|\Gamma_N(t)^{1/2} - \Gamma(t)^{1/2}\|_{\mathrm{HS}}^2 \leq \sum_{i=1}^{d} \lambda_i(t) = \mathrm{tr}(\Gamma(t)) = \|\sigma(t)\|_{\mathrm{HS}}^2.$$

Since σ is square-integrable on $[0,1]$, dominated convergence gives

$$\int_0^1 \|\sigma_N(t) - \sigma(t)\|_{\mathrm{HS}}^2\, dt \to 0.$$

Let $Y^{(N)}$ solve

$$dY_t^{(N)} = \beta(t, Y_t^{(N)})\, dt + \sigma_N(t)\, dW_t, \qquad Y_0^{(N)} = y,$$

and let Y solve the original equation

$$dY_t = \beta(t, Y_t)\, dt + \sigma(t)\, dW_t, \qquad Y_0 = y.$$

Drive both equations by the same Brownian motion. Define

$$M_t^{(N)} := \int_0^t (\sigma_N(s) - \sigma(s))\, dW_s.$$

Then

$$Y_t^{(N)} - Y_t = \int_0^t \left[\beta(s, Y_s^{(N)}) - \beta(s, Y_s)\right] ds + M_t^{(N)}.$$

Using that $\beta(t, \cdot)$ is 1-Lipschitz,

$$|Y_t^{(N)} - Y_t| \leq \int_0^t |Y_s^{(N)} - Y_s|\, ds + \sup_{0 \leq u \leq t} |M_u^{(N)}|.$$

By Gronwall's inequality,

$$\sup_{0 \leq t \leq 1} |Y_t^{(N)} - Y_t| \leq e \sup_{0 \leq t \leq 1} |M_t^{(N)}|.$$

By the Burkholder–Davis–Gundy inequality, or by Doob's inequality and Itô's isometry,

$$\mathbb{E}\left[\sup_{0\le t\le 1}|M_t^{(N)}|^2\right] \le C\int_0^1 \|\sigma_N(t)-\sigma(t)\|_{\mathrm{HS}}^2\,dt \to 0.$$

Therefore $Y_1^{(N)} \to Y_1$ in probability, hence in distribution.

By Step 4a, applied to the bounded volatility σ_N, the endpoint law of $Y_1^{(N)}$ satisfies

$$\mathbb{P}(Y_1^{(N)} \in E) \le C_{d,\kappa}\lambda_d(E)$$

for every Borel set E, with a constant independent of N. Passing to the weak limit exactly as above gives

$$\mathbb{P}(Y_1 \in E) \le C_{d,\kappa}\lambda_d(E)$$

for every Borel set E. Thus the endpoint law of Y_1 has a density bounded by $C_{d,\kappa}$. This completes the removal of the auxiliary assumptions. □

B.5 Proof of Proposition 3.3, Assumption (A2): state-independent volatility

We first prove the matching localization lower bound in the simpler case where the volatility is state-independent. The proof separates two regimes: short horizon, where the K-Lipschitz drift cannot concentrate an absolutely continuous initialization too quickly, and longer horizons, where the final K^{-1}-length time window yields an inevitable $K^{-1/2}$-scale fluctuation.

Proof. All constants below may depend on m, d, κ, μ_0, R, but not on K, f, σ, T, nor on the locations or pairwise separations of the points in $\mathcal{Z}$.

Notational note. Throughout this proof, we will suppress the state argument in the drift $f(t,a,x)$, so that it becomes $f(t,a)$. This is because we do not use the state x in our analysis. Therefore, we clarify any indication of variables $x, z, \cdots \in \mathbb{R}^d$ in this section are actions, not states (d is the action dimension).

Step 1: the case $T \le K^{-1}$.

Under (A2), subtracting the additive noise $M_t := \int_0^t \sigma(s)\,dB_s$ reduces the SDE to a random ODE with globally K-Lipschitz drift; hence the SDE admits a pathwise unique strong realization, and for each deterministic initial action x we may write $\bar{a}_{t,x}$ for the solution driven by the same Brownian path. For two initial actions x, z, state-independence of $\sigma(t)$ gives

$$d(\bar{a}_{t,x}-\bar{a}_{t,z}) = [f(t,\bar{a}_{t,x}) - f(t,\bar{a}_{t,z})]\,dt.$$

Thus $t \mapsto \bar{a}_{t,x}-\bar{a}_{t,z}$ is absolutely continuous, and for a.e. t,

$$\frac{d}{dt}|\bar{a}_{t,x}-\bar{a}_{t,z}|^2 = 2\left\langle \bar{a}_{t,x}-\bar{a}_{t,z}, f(t,\bar{a}_{t,x}) - f(t,\bar{a}_{t,z})\right\rangle.$$

Since $f(t,\cdot)$ is K-Lipschitz,

$$\frac{d}{dt}|\bar{a}_{t,x}-\bar{a}_{t,z}|^2 \ge -2K|\bar{a}_{t,x}-\bar{a}_{t,z}|^2.$$

By Gronwall's inequality, $|\bar{a}_{T,x}-\bar{a}_{T,z}| \ge e^{-KT}|x-z|$. Therefore, conditional on the Brownian path, the map $x \mapsto \bar{a}_{T,x}$ is injective, and its inverse on its image is e^{KT}-Lipschitz.

Hence, for every Borel set $E \subset \mathbb{R}^d$,

$$\lambda_d\left(\{x : \bar{a}_{T,x} \in E\}\right) \le e^{dKT}\lambda_d(E).$$

If $T \le K^{-1}$, then $e^{dKT} \le e^d$, and therefore $\lambda_d\left(\{x : \bar{a}_{T,x} \in E\}\right) \le e^d\lambda_d(E)$. Since $\mu_0 \ll \lambda_d$, there exists $\alpha > 0$ such that

$$\lambda_d(E) \le \alpha \quad\Longrightarrow\quad \mu_0(E) \le \frac{1}{2}.$$

For $\delta > 0$, define

$$U_\delta = \bigcup_{j=1}^{J} B(z_j, \delta).$$

Since $\lambda_d(U_\delta) \le J v_d \delta^d$, choose $\delta > 0$, depending only on J, d, μ_0, such that $e^d J v_d \delta^d \le \alpha$. Then, conditional on the Brownian path, $\lambda_d\left(\{x : \bar{a}_{T,x} \in U_\delta\}\right) \le e^d \lambda_d(U_\delta) \le \alpha$. Therefore

$$\mu_0\left(\{x : \bar{a}_{T,x} \in U_\delta\}\right) \le \frac{1}{2}.$$

Since $\bar{a}_0 \sim \mu_0$ is independent of B, $\mathbb{P}(\bar{a}_T \in U_\delta \mid B) \le 1/2$. Taking expectations, $\mathbb{P}(\bar{a}_T \notin U_\delta) \ge 1/2$. On the event $\{\bar{a}_T \notin U_\delta\}$, we have $\mathsf{dist}(\bar{a}_T, \mathcal{Z}) \ge \delta$, and hence

$$\mathsf{dist}(\bar{a}_T, \mathcal{Z})^2 \wedge R^2 \ge \delta^2 \wedge R^2.$$

Consequently, as $K \ge 1$,

$$\mathbb{E}\left[\mathsf{dist}(\bar{a}_T, \mathcal{Z})^2 \wedge R^2\right] \ge \frac{\delta^2 \wedge R^2}{2} \ge \frac{\delta^2 \wedge R^2}{2K}.$$

Step 2: the case $T > K^{-1}$.

Assume $T > K^{-1}$ and set $s = T - 1/K$. Condition on $\mathcal{F}_s$, and write $y = \bar{a}_s$. Let ϕ_t, for $t \in [s, T]$, solve the deterministic equation

$$\frac{d}{dt}\phi_t = f(t, \phi_t), \qquad \phi_s = y.$$

The existence of (ϕ_t) follows from the condition that $f(\cdot, 0) \in L^1_{\mathrm{loc}}$. For $0 \le t \le 1$, define

$$Y_t = \sqrt{K}\left(\bar{a}_{s+t/K} - \phi_{s+t/K}\right),$$

and

$$W_t = \sqrt{K}\left(B_{s+t/K} - B_s\right).$$

Conditionally on $\mathcal{F}_s$, W is a standard Brownian motion. We compute the equation for Y_t. First,

$$\bar{a}_{s+t/K} = y + \int_s^{s+t/K} f(r, \bar{a}_r)\, dr + \int_s^{s+t/K} \sigma(r)\, dB_r,$$

while

$$\phi_{s+t/K} = y + \int_s^{s+t/K} f(r, \phi_r)\, dr.$$

Subtracting and multiplying by $\sqrt{K}$ gives

$$Y_t = \sqrt{K}\int_s^{s+t/K} \left[f(r, \bar{a}_r) - f(r, \phi_r)\right] dr + \sqrt{K}\int_s^{s+t/K} \sigma(r)\, dB_r.$$

Using the change of variables $r = s + u/K$ and the identity

$$\bar{a}_{s+u/K} = \phi_{s+u/K} + \frac{Y_u}{\sqrt{K}},$$

we obtain

$$dY_t = \beta_K(t, Y_t)\, dt + \sigma(s + t/K)\, dW_t, \qquad Y_0 = 0,$$

where

$$\beta_K(t, u) = \frac{f(s + t/K, \phi_{s+t/K} + u/\sqrt{K}) - f(s + t/K, \phi_{s+t/K})}{\sqrt{K}}.$$

The drift β_K is 1-Lipschitz in u. Indeed,

$$\|\beta_K(t, u) - \beta_K(t, v)\| \le \frac{1}{\sqrt{K}} K \left\|\frac{u - v}{\sqrt{K}}\right\| = \|u - v\|.$$

Moreover, $\sigma(s+t/K)\sigma(s+t/K)^\top \succeq \kappa I_d$. By Lemma B.2, conditionally on $\mathcal{F}_s$, the random variable Y_1 has density bounded by $C_{d,\kappa}$. Therefore, for arbitrary $y_1, \dots, y_J \in \mathbb{R}^d$ and any $\rho > 0$,

$$\mathbb{P}\left(Y_1 \in \bigcup_{j=1}^{J} B(y_j, \rho) \,\middle|\, \mathcal{F}_s \right) \le C_{d,\kappa} J v_d \rho^d.$$

Choose $\rho_0 = (2C_{d,\kappa} J v_d)^{-1/d} \wedge 1$. Then $C_{d,\kappa} J v_d \rho_0^d \le 1/2$. Now set $y_j = \sqrt{K}(z_j - \phi_T)$ for $j = 1, \dots, J$. Since $\bar{a}_T = \phi_T + Y_1/\sqrt{K}$, we have

$$\mathsf{dist}(\bar{a}_T, \mathcal{Z}) \le \frac{\rho_0}{\sqrt{K}} \quad \Longleftrightarrow \quad Y_1 \in \bigcup_{j=1}^{J} B(y_j, \rho_0).$$

Therefore

$$\mathbb{P}\left(\mathsf{dist}(\bar{a}_T, \mathcal{Z}) \le \frac{\rho_0}{\sqrt{K}} \,\middle|\, \mathcal{F}_s \right) \le \frac{1}{2},$$

or equivalently

$$\mathbb{P}\left(\mathsf{dist}(\bar{a}_T, \mathcal{Z}) > \frac{\rho_0}{\sqrt{K}} \,\middle|\, \mathcal{F}_s \right) \ge \frac{1}{2}.$$

If $K \ge \rho_0^2/R^2$, then $\rho_0/\sqrt{K} \le R$. On the event $\mathsf{dist}(\bar{a}_T, \mathcal{Z}) > \rho_0/\sqrt{K}$, we have

$$\mathsf{dist}(\bar{a}_T, \mathcal{Z})^2 \wedge R^2 \ge \frac{\rho_0^2}{K}.$$

Thus

$$\mathbb{E}\left[\mathsf{dist}(\bar{a}_T, \mathcal{Z})^2 \wedge R^2 \,\middle|\, \mathcal{F}_s \right] \ge \frac{\rho_0^2}{K} \mathbb{P}\left(\mathsf{dist}(\bar{a}_T, \mathcal{Z}) > \frac{\rho_0}{\sqrt{K}} \,\middle|\, \mathcal{F}_s \right) \ge \frac{\rho_0^2}{2K}.$$

Taking expectations gives

$$\mathbb{E}\left[\mathsf{dist}(\bar{a}_T, \mathcal{Z})^2 \wedge R^2 \right] \ge \frac{\rho_0^2}{2K}.$$

Step 3: combine the two regimes.

Let

$$K_R = 1 \vee \frac{\rho_0^2}{R^2},$$

and define

$$c_R = \frac{1}{2} \min\left\{ \delta^2 \wedge R^2, \rho_0^2 \right\}.$$

If $K \ge K_R$ and $T \le K^{-1}$, Step 1 gives

$$\mathbb{E}\left[\mathsf{dist}(\bar{a}_T, \mathcal{Z})^2 \wedge R^2 \right] \ge \frac{\delta^2 \wedge R^2}{2K} \ge \frac{c_R}{K}.$$

If $K \ge K_R$ and $T > K^{-1}$, Step 2 gives

$$\mathbb{E}\left[\mathsf{dist}(\bar{a}_T, \mathcal{Z})^2 \wedge R^2 \right] \ge \frac{\rho_0^2}{2K} \ge \frac{c_R}{K}.$$

Thus, for every $T \ge 0$,

$$\mathbb{E}\left[\mathsf{dist}(\bar{a}_T, \mathcal{Z})^2 \wedge R^2 \right] \ge \frac{c_R}{K}.$$

Taking the infimum over all admissible f, σ, T proves

$$\inf_{f,\sigma,T \ge 0} \mathbb{E}\left[\mathsf{dist}(\bar{a}_T, \mathcal{Z})^2 \wedge R^2 \right] \ge \frac{c_R}{K}.$$

The proof is complete. □

B.6 Proof of Theorem 3.4

We next translate the finite-set localization lower bound into an RL value lower bound. The only additional ingredient is the quadratic finite-well Bellman-gap condition, which converts squared distance from optimal action set into loss in value.

Proof. By Proposition 3.3, applied for each cardinality $J = 1, \dots, M$, there exist constants $c_{J,R} > 0$ and $K_{J,R} < \infty$ such that the clipped finite-set localization lower bound holds for every finite set $A \subset \mathbb{R}^d$ with $|A| = m$. Set

$$c_{M,R} := \min_{1 \le m \le M} c_{m,R}, \qquad K_0 := \max_{1 \le m \le M} K_{m,R}.$$

Then, for every finite set $A \subset \mathbb{R}^d$ with $1 \le |A| \le M$, every admissible K-Lipschitz drift, every admissible volatility satisfying the ellipticity bounds, and every terminal time $T > 0$,

$$\mathbb{E}\left[\operatorname{dist}(\bar{a}_T, A)^2 \wedge R^2\right] \ge \frac{c_{M,R}}{K}, \qquad K \ge K_0. \tag{B.1}$$

The constants $c_{M,R}$ and K_0 depend only on $(M, R, d, \kappa, \Lambda, \mu_0)$, and not on the locations or pairwise separations of the points in A.

Fix $K \ge K_0$ and $T > 0$, and write $\pi = \pi_{K,T}$. For each state x,

$$\begin{aligned} V^\star(x) - V^\pi(x) &= V^\star(x) - \int_{\mathbb{R}^d} \left[r(x,a) + \gamma \int_{\mathcal{X}} V^\pi(y) P(dy \mid x, a) \right] \pi(da \mid x) \\ &= \int_{\mathbb{R}^d} \left[V^\star(x) - Q^\star(x,a) \right] \pi(da \mid x) \\ &\quad + \gamma \int_{\mathbb{R}^d} \int_{\mathcal{X}} \left[V^\star(y) - V^\pi(y) \right] P(dy \mid x, a) \pi(da \mid x). \end{aligned}$$

Since $V^\star \ge V^\pi$, the second term is nonnegative. Hence

$$V^\star(x) - V^\pi(x) \ge \int_{\mathbb{R}^d} \left[V^\star(x) - Q^\star(x,a) \right] \pi(da \mid x). \tag{B.2}$$

Integrating over $x \sim \rho$ gives

$$V^\star(\rho) - V^\pi(\rho) \ge \int_{\mathcal{X}} \int_{\mathbb{R}^d} \left[V^\star(x) - Q^\star(x,a) \right] \pi(da \mid x) \rho(dx).$$

By the finite-well Bellman gap assumption in Theorem 3.4, for all (x, a),

$$V^\star(x) - Q^\star(x,a) \ge \lambda_{M,R} \left(\operatorname{dist}(a, A^\star(x))^2 \wedge R^2 \right).$$

Therefore,

$$V^\star(\rho) - V^\pi(\rho) \ge \lambda_{M,R} \int_{\mathcal{X}} \int_{\mathbb{R}^d} \left(\operatorname{dist}(a, A^\star(x))^2 \wedge R^2 \right) \pi(da \mid x) \rho(dx). \tag{B.3}$$

For each x, the set $A^\star(x)$ is nonempty and satisfies $|A^\star(x)| \le M$. Moreover, under $\pi_{K,T}$, the action at state x is the terminal value of the state-conditioned diffusion. Thus, by (B.1) applied with $A = A^\star(x)$,

$$\int_{\mathbb{R}^d} \left(\operatorname{dist}(a, A^\star(x))^2 \wedge R^2 \right) \pi_{K,T}(da \mid x) = \mathbb{E}\left[\operatorname{dist}(\bar{a}_T^x, A^\star(x))^2 \wedge R^2 \right] \ge \frac{c_{M,R}}{K},$$

uniformly in x and $T > 0$. Substituting this into (B.3) yields

$$V^\star(\rho) - V^{\pi_{K,T}}(\rho) \ge \lambda_{M,R} \int_{\mathcal{X}} \frac{c_{M,R}}{K} \rho(dx) = \frac{c_{M,R} \lambda_{M,R}}{K}.$$

The bound is uniform over $T > 0$. Hence

$$\inf_{T > 0} \left(V^\star(\rho) - V^{\pi_{K,T}}(\rho) \right) \ge \frac{c_{M,R} \lambda_{M,R}}{K},$$

which proves the claim. □

B.7 Analog of Theorem 3.4 with arbitrary initialization

The previous lower bound uses the absolute continuity of the initial distribution to control very short horizons. This subsection explains why the statement fails for arbitrary, possibly discrete, initialization. And then we show that the lower bound is restored once a burn-in time of order K^{-1} is imposed.

B.7.1 Counterexample with discrete initialization

Indeed, consider a diffusion from source $\bar{a}_0 = 0$ to the target distribution δ_1 (point mass at 1). Let $f(t, a, x) = M$ constant for all a, t and x, so that

$$\bar{a}_t = Mt + B_t$$

At time $t = 1/M$, we have $\bar{a}_t \sim N(1, 1/M)$, and hence $\mathbb{E}|\bar{a}_t - 1|^2 = 1/M \to 0$. The drift is constant in a, hence K-Lipschitz for every K. Thus no lower bound uniform over all $t > 0$ can hold for arbitrary initialization. The obstruction is that the time horizon can be arbitrarily small, so the accumulated diffusion variance vanishes, while the drift magnitude is not controlled by the Lipschitz budget.

B.7.2 Arbitrary action initialization after burn-in

Corollary B.3 (Arbitrary action initialization after burn-in)**.** *For each $K \geq 1$ and $T > 0$, let $\pi_{K,T}$ be the Markov policy whose action law at state x is the terminal law at time T of Eq.* (2.3)*, with $\bar{a}_0^x \sim P_x$, where P_x is an arbitrary probability law on $\mathbb{R}^d$, possibly depending on x, and $\bar{a}_0^x$ is independent of the Brownian motion. Assume that, apart from the initialization law, the conditions in Proposition 3.3 hold for every x, with (κ, Λ) independent of (t, x).*

Fix $M \geq 1$ and $R > 0$. Suppose that, for every $x \in \mathcal{X}$, $|A^\star(x)| \leq M$. Suppose moreover that there exists a constant $\lambda_{M,R} > 0$, independent of x, such that

$$V^\star(x) - Q^\star(x, a) \geq \lambda_{M,R} \left(\mathsf{dist}(a, A^\star(x))^2 \wedge R^2\right)$$

for all (x, a). Then there exist constants $c_{M,R} > 0$ and $K_0 < \infty$, depending only on M, R, d, κ, Λ, such that for every $K \geq K_0$,

$$\inf_{T \geq K^{-1}} \left(V^\star(\rho) - V^{\pi_{K,T}}(\rho)\right) \geq \frac{c_{M,R}\lambda_{M,R}}{K}.$$

Consequently, for every burn-in sequence $t_0(K)$ satisfying $t_0(K) \geq K^{-1}$ for all sufficiently large K,

$$\inf_{T \geq t_0(K)} \left(V^\star(\rho) - V^{\pi_{K,T}}(\rho)\right) \geq \frac{c_{M,R}\lambda_{M,R}}{K}.$$

Proof. The only difference from Theorem 3.4 is that the action diffusion may start from an arbitrary law P_x, so the short-time part of Proposition 3.3 cannot be used. Instead, we use only the final-window part of the proof of Proposition 3.3.

We first record the resulting burn-in localization bound. There exist constants $c_{M,R} > 0$ and $K_0 < \infty$, depending only on M, R, d, κ, Λ, such that the following holds. For every finite set $A \subset \mathbb{R}^d$ with $1 \leq |A| \leq M$, every admissible K-Lipschitz drift, every admissible volatility satisfying the ellipticity bounds, every initial law, and every terminal time $T \geq K^{-1}$,

$$\mathbb{E}\left[\mathsf{dist}(\bar{a}_T, A)^2 \wedge R^2\right] \geq \frac{c_{M,R}}{K}.$$

Indeed, if $T \geq K^{-1}$, set $s = T - K^{-1}$ and condition on $\mathcal{F}_s$. The rescaling argument in the final window $[s, s + K^{-1}]$ gives, uniformly over the random value $\bar{a}_s$ and hence uniformly over the initial law, a probability bounded below of being at distance at least order $K^{-1/2}$ from any M-point target set. This is exactly the conditional argument in the final-window step of Proposition 3.3. The pigeonhole/Krylov step supplies a constant probability $c_0 > 0$ uniformly over the target locations. Enlarging K_0 so that $r_0/\sqrt{K} \leq R$ gives the clipped lower bound.

Fix $K \geq K_0$ and $T \geq K^{-1}$. Write $\pi = \pi_{K,T}$. For every state x,

$$\begin{aligned} V^\star(x) - V^\pi(x) &= V^\star(x) - \int_{\mathbb{R}^d} \left[r(x,a) + \gamma \int_{\mathcal{X}} V^\pi(y) P(dy \mid x,a) \right] \pi(da \mid x) \\ &= \int_{\mathbb{R}^d} \left[V^\star(x) - Q^\star(x,a) \right] \pi(da \mid x) \\ &\quad + \gamma \int_{\mathbb{R}^d} \int_{\mathcal{X}} \left[V^\star(y) - V^\pi(y) \right] P(dy \mid x,a) \pi(da \mid x). \end{aligned}$$

Since $V^\star \geq V^\pi$, the second term is nonnegative. Hence

$$V^\star(x) - V^\pi(x) \geq \int_{\mathbb{R}^d} \left[V^\star(x) - Q^\star(x,a) \right] \pi(da \mid x).$$

Integrating over $x \sim \rho$ gives

$$V^\star(\rho) - V^\pi(\rho) \geq \int_{\mathcal{X}} \int_{\mathbb{R}^d} \left[V^\star(x) - Q^\star(x,a) \right] \pi(da \mid x) \rho(dx).$$

By the clipped quadratic-gap assumption, there exists $\lambda_{M,R} > 0$, independent of x, such that

$$V^\star(x) - Q^\star(x,a) \geq \lambda_{M,R} \left(\mathsf{dist}(a, A^\star(x))^2 \wedge R^2 \right)$$

for all (x,a). Therefore,

$$V^\star(\rho) - V^\pi(\rho) \geq \lambda_{M,R} \int_{\mathcal{X}} \int_{\mathbb{R}^d} \left(\mathsf{dist}(a, A^\star(x))^2 \wedge R^2 \right) \pi(da \mid x) \rho(dx).$$

For every x, the set $A^\star(x)$ is nonempty and satisfies $|A^\star(x)| \leq M$. Moreover, the action law $\pi(\cdot \mid x)$ is the terminal law at time $T \geq K^{-1}$ of an admissible diffusion started from the arbitrary law P_x. Therefore the burn-in localization bound gives

$$\int_{\mathbb{R}^d} \left(\mathsf{dist}(a, A^\star(x))^2 \wedge R^2 \right) \pi(da \mid x) \geq \frac{c_{M,R}}{K}$$

uniformly in x. Substituting,

$$V^\star(\rho) - V^\pi(\rho) \geq \lambda_{M,R} \int_{\mathcal{X}} \frac{c_{M,R}}{K} \rho(dx) = \frac{c_{M,R} \lambda_{M,R}}{K}.$$

The bound is uniform over $T \geq K^{-1}$, so

$$\inf_{T \geq K^{-1}} \left(V^\star(\rho) - V^{\pi_{K,T}}(\rho) \right) \geq \frac{c_{M,R} \lambda_{M,R}}{K}.$$

Consequently, if $t_0(K) \geq K^{-1}$ for all sufficiently large K, then for all sufficiently large K,

$$\inf_{T \geq t_0(K)} \left(V^\star(\rho) - V^{\pi_{K,T}}(\rho) \right) \geq \frac{c_{M,R} \lambda_{M,R}}{K},$$

which proves the claim. □

B.8 Lower bound with state-dependent volatility

We now prepare the lower-bound proof for state-dependent volatility. Since the density argument used in the state-independent case no longer applies here directly, we replace it with a local Krylov estimate that gives a uniform probability of reaching small balls while remaining in a controlled region.

Lemma B.4 (Local Krylov lower estimate)**.** *Let $d \geq 2$ and fix $0 < \lambda \leq \Lambda < \infty$. Let*

$$X_u = \int_0^u \alpha_r \, dW_r + \int_0^u \beta_r \, dr, \qquad 0 \leq u \leq 1,$$

where $X_0 = 0$, the coefficients are progressively measurable, and

$$\lambda I_d \preceq \alpha_r \alpha_r^\top \preceq \Lambda I_d, \qquad 0 \le r \le 1.$$

Define the finite-horizon exit time $\tau_2^X := \inf\{u \in [0,1] : X_u \notin B_2\}$, with $\inf \varnothing := \infty$, and the no-exit event

$$\mathcal{E}_2^X := \{X_u \in B_2 \text{ for all } 0 \le u \le 1\} = \{\tau_2^X > 1\}.$$

Assume that $\|\beta_r\| \le 2$ for $0 \le r \le 1$ on $\{r < \tau_2^X\}$. Then there exist constants $N < \infty$ and $\nu > 0$, depending only on d, λ, Λ, such that for every $q \in B_1(0)$ and every $0 < \rho \le 1$,

$$\mathbb{P}\big(X_1 \in B_{2\rho}(q),\ \mathcal{E}_2^X\big) \ge N^{-1}\rho^\nu.$$

Proof. We reduce the claim to Theorem 2.1 of Krylov (2021). First define an auxiliary process on $[0,\infty)$ by stopping and modifying the coefficients after exit from B_2 and after time 1. Set

$$\widehat{\beta}_r := \begin{cases} \beta_r \, \mathbf{1}_{\{r < \tau_2^X\}}, & 0 \le r \le 1, \\ 0, & r > 1, \end{cases}$$

and

$$\widehat{\alpha}_r := \begin{cases} \alpha_r, & 0 \le r \le 1 \text{ and } r < \tau_2^X, \\ \sqrt{\lambda}\, I_d, & \text{otherwise.} \end{cases}$$

Define

$$\widehat{X}_u := \int_0^u \widehat{\alpha}_r \, dW_r + \int_0^u \widehat{\beta}_r \, dr, \qquad u \ge 0.$$

Then $\widehat{X}$ agrees with X up to the first exit of X from B_2 on the time interval $[0,1]$. If

$$\widehat{\tau}_2 := \inf\{u \ge 0 : \widehat{X}_u \notin B_2\},$$

then, for every Borel set $A \subseteq \mathbb{R}^d$, $\{\widehat{X}_1 \in A,\ \widehat{\tau}_2 > 1\} = \{X_1 \in A,\ \mathcal{E}_2^X\}$. Moreover, $\lambda I_d \preceq \widehat{\alpha}_r \widehat{\alpha}_r^\top \preceq \Lambda I_d$, $r \ge 0$, and $\|\widehat{\beta}_r\| \le 2\,\mathbf{1}_{\{0 \le r \le 1\}} \mathbf{1}_{\{\widehat{X}_r \in B_2\}}$.

Next reduce to symmetric volatility. Set $a_r := \widehat{\alpha}_r \widehat{\alpha}_r^\top$ and $S_r := a_r^{1/2}$. Then S_r satisfies $\delta_0 I_d \preceq S_r \preceq \delta_0^{-1} I_d$ with $\delta_0 := \min\{\sqrt{\lambda}, \Lambda^{-1/2}, 1/2\}$. Furthermore, the process

$$\widetilde{W}_u := \int_0^u S_r^{-1} \widehat{\alpha}_r \, dW_r$$

is a Brownian motion. By associativity of Ito integrals, $\widehat{X}$ is thus in Krylov's diffusion-type form with symmetric volatility with eigenvalues in $[\delta_0, \delta_0^{-1}]$.

It remains to verify Krylov's drift domination condition for the particular function

$$h(t,z) := 2\,\mathbf{1}_{\{0 \le t \le 1\}} \mathbf{1}_{\{z \in B_2\}}.$$

We choose the admissible exponents $p_0 = q_0 = d+1$, so that Krylov's mixed norm L_{p_0,q_0} coincides with the usual space-time L^{d+1} norm. Then $\|\widehat{\beta}_t\| \le h(t, \widehat{X}_t)$. For a parabolic cylinder $C_\ell(t_0,z_0) := [t_0, t_0+\ell^2) \times B_\ell(z_0)$, we have

$$\begin{aligned} \|h\|^{d+1}_{L^{d+1}(C_\ell(t_0,z_0))} &= \int_{C_\ell(t_0,z_0)} 2^{d+1} \mathbf{1}_{\{0 \le t \le 1\}} \mathbf{1}_{\{z \in B_2\}} \, dz\, dt \\ &= 2^{d+1} \big|[t_0, t_0+\ell^2) \cap [0,1]\big| \, |B_\ell(z_0) \cap B_2| \\ &\le 2^{d+1} v_d (\ell^2 \wedge 1)(\ell^d \wedge 2^d). \end{aligned}$$

where $v_d := \lambda_d(B_1(0)) = \pi^{d/2}/\Gamma(d/2+1)$ denotes the Lebesgue volume of the unit ball in $\mathbb{R}^d$. If $0 < \ell \le 1$, this is bounded by $C_d \ell^{d+2}$; if $\ell \ge 1$, it is bounded by $C_d \le C_d \ell$. Hence, for all $\ell > 0$,

$$\|h\|^{d+1}_{L^{d+1}(C_\ell(t_0,z_0))} \le C_d(\ell^{d+1} \wedge 1)\ell.$$

Thus Krylov's Assumption 2.1 holds with $\bar{b}_\ell := C_d(\ell^{d+1} \wedge 1)$. The function $\bar{b}_\ell$ is bounded, nondecreasing, and satisfies $\bar{b}_\ell \downarrow 0$ as $\ell \downarrow 0$. Therefore Krylov's Assumption 2.2 holds at some sufficiently

small scale $R_* > 0$ such that $\bar{N}\,\bar{b}_{R_*} < 1$. depending only on d and δ_0. In Krylov's notation, fix $\overline{R} = 2$.

Now apply Krylov's Theorem 2.1 with application radius $\mathsf{R} = 2$, $\kappa = \frac{1}{2}$, $\eta = \frac{1}{4}$, and $t = 1$. The time condition is $\eta\mathsf{R}^2 = 1 \le t = 1 \le \eta^{-1}\mathsf{R}^2 = 16$. Since $0, q \in B_1(0)$ with $\kappa\mathsf{R} = 1$, Krylov's estimate gives

$$\mathbb{P}\big(\widehat{X}_1 \in B_{\rho\mathsf{R}}(q),\ \widehat{\tau}_2 > 1\big) \ge N^{-1}\rho^{\nu}.$$

Because $\rho\mathsf{R} = 2\rho$ and $\{\widehat{X}_1 \in B_{2\rho}(q), \widehat{\tau}_2 > 1\} = \{X_1 \in B_{2\rho}(q), \mathcal{E}_2^X\}$, we obtain

$$\mathbb{P}\big(X_1 \in B_{2\rho}(q),\ \mathcal{E}_2^X\big) \ge N^{-1}\rho^{\nu}.$$

All constants depend only on d, λ, Λ. Hence we are done. □

B.9 Proof of Proposition 3.3, assumption (a): state-dependent volatility

We now complete the proof of Proposition 3.3 under the full state-dependent volatility assumption. The proof follows the same short-time vs. final-window decomposition as in the state-independent case, but the final concentration step is obtained from the local Krylov of Lemma B.4.

Proof. All constants below may depend on $m, d, \kappa, \Lambda, \mu_0, R$, but not on K, f, σ, t, nor on the locations or pairwise separations of the points in $\mathcal{Z}$.

Notational note. Throughout this proof, we will suppress the state argument in the drift $f(t, a, x)$ and $\sigma(t, a, x)$, so that they become $f(t, a)$ and $\sigma(t, a)$, respectively. This is because we do not use the state x in our analysis. Therefore, we clarify any indication of variables $x, z, \cdots \in \mathbb{R}^d$ in this section are actions, not states (d is the action dimension).

Step 1: the case $t \le K^{-1}$. For $\delta > 0$, define

$$U_\delta := \bigcup_{j=1}^{m} B(z_j, \delta).$$

Let $\phi_t(x)$ be the deterministic flow generated by f:

$$\frac{d}{dt}\phi_t(x) = f(t, \phi_t(x)), \qquad \phi_0(x) = x.$$

Since $f(t, \cdot)$ is K-Lipschitz, for any $x, z \in \mathbb{R}^d$,

$$\frac{d}{dt}|\phi_t(x) - \phi_t(z)|^2 \ge -2K|\phi_t(x) - \phi_t(z)|^2.$$

Therefore

$$|\phi_t(x) - \phi_t(z)| \ge e^{-Kt}|x - z|.$$

Thus ϕ_t is injective, and ϕ_t^{-1} is e^{Kt}-Lipschitz on $\phi_t(\mathbb{R}^d)$. Hence, for every Borel set $E \subset \mathbb{R}^d$ and every $t \le K^{-1}$,

$$\lambda_d(\phi_t^{-1}(E)) \le e^{dKt}\lambda_d(E) \le e^d\lambda_d(E).$$

Because $\mu_0 \ll \lambda_d$, there exists $\alpha > 0$ such that

$$\lambda_d(E) \le \alpha \quad \Longrightarrow \quad \mu_0(E) \le \frac{1}{4}.$$

Choose $\delta > 0$, depending only on m, d, μ_0, such that $e^d\lambda_d(U_{2\delta}) \le \alpha$. This is possible because

$$\lambda_d(U_{2\delta}) \le m v_d (2\delta)^d.$$

Then, for every $t \le K^{-1}$,

$$\mathbb{P}\{\phi_t(\bar{a}_0) \in U_{2\delta}\} = \mu_0(\phi_t^{-1}(U_{2\delta})) \le \frac{1}{4}.$$

Couple the diffusion and deterministic flow from the same initial point and set

$$D_t := \bar{a}_t - \phi_t(\bar{a}_0).$$

Then

$$dD_t = \{f(t, \bar{a}_t) - f(t, \phi_t(\bar{a}_0))\}\, dt + \sigma(t, \bar{a}_t)\, dB_t.$$

By Itô's formula, the Lipschitz bound on f, and the upper volatility bound,

$$\frac{d}{dt}\mathbb{E}\|D_t\|^2 \le 2K\mathbb{E}\|D_t\|^2 + d\Lambda.$$

Since $D_0 = 0$, Gronwall's inequality gives, for $t \le K^{-1}$,

$$\mathbb{E}\|D_t\|^2 \le \frac{d\Lambda}{2K}(e^{2Kt} - 1) \le \frac{d\Lambda}{2K}(e^2 - 1).$$

Thus, for all sufficiently large K,

$$\mathbb{P}\{\|D_t\| \ge \delta\} \le \frac{\mathbb{E}\|D_t\|^2}{\delta^2} \le \frac{1}{4}.$$

If $\bar{a}_t \in U_\delta$ and $\|D_t\| < \delta$, then $\phi_t(\bar{a}_0) \in U_{2\delta}$. Hence $\mathbb{P}\{\bar{a}_t \in U_\delta\} \le 1/4 + 1/4 = 1/2$. On U_δ^c, we have

$$\mathsf{dist}(x, \mathcal{Z})^2 \wedge R^2 \ge \delta^2 \wedge R^2.$$

Therefore, for every $t \le K^{-1}$ and all sufficiently large K,

$$\mathbb{E}\left[\mathsf{dist}(\bar{a}_t, \mathcal{Z})^2 \wedge R^2\right] \ge (\delta^2 \wedge R^2)\, \mathbb{P}\{\bar{a}_t \notin U_\delta\} \ge \frac{\delta^2 \wedge R^2}{2}.$$

In particular, for $K \ge 1$,

$$\mathbb{E}\left[\mathsf{dist}(\bar{a}_t, \mathcal{Z})^2 \wedge R^2\right] \ge \frac{\delta^2 \wedge R^2}{2K}.$$

Step 2: the case $t > K^{-1}$. Fix $t > K^{-1}$ and set $s := t - 1/K$. We prove a lower bound conditionally on $\mathcal{F}_s$. Set $y := \bar{a}_s$. Let ϕ_u solve

$$\frac{d}{du}\phi_u = f(u, \phi_u), \qquad \phi_s = y.$$

For $0 \le \ell \le 1$, define $Y_\ell := \sqrt{K}\left(\bar{a}_{s+\ell/K} - \phi_{s+\ell/K}\right)$ and $W_\ell := \sqrt{K}\left(B_{s+\ell/K} - B_s\right)$. Conditionally on $\mathcal{F}_s$, W is a standard Brownian motion. The process Y satisfies

$$dY_\ell = \beta_K(\ell, Y_\ell)\, d\ell + \widetilde{\sigma}_K(\ell, Y_\ell)\, dW_\ell, \qquad Y_0 = 0,$$

where

$$\beta_K(\ell, u) = \frac{f(s + \ell/K, \phi_{s+\ell/K} + u/\sqrt{K}) - f(s + \ell/K, \phi_{s+\ell/K})}{\sqrt{K}}$$

and

$$\widetilde{\sigma}_K(\ell, u) = \sigma(s + \ell/K, \phi_{s+\ell/K} + u/\sqrt{K}).$$

The K-Lipschitz property gives

$$\|\beta_K(\ell, u) - \beta_K(\ell, v)\| \le \|u - v\|.$$

Also $\beta_K(\ell, 0) = 0$. Hence, on B_2, $\|\beta_K(\ell, u)\| \le 2$. The rescaled covariance matrix satisfies, by Assumption 2.1,

$$\kappa I_d \preceq \widetilde{\sigma}_K(\ell, u)\widetilde{\sigma}_K(\ell, u)^\top \preceq \Lambda I_d.$$

We first treat $d \ge 2$. Let

$$\mathcal{E}_2^Y := \{Y_u \in B_2 \text{ for all } 0 \le u \le 1\}$$

Construct a concatenated process $\widehat{Y}$ as follows. Up to time τ_2, set $\widehat{Y}_\ell = Y_\ell$. After τ_2, continue with drift zero and covariance matrix κI_d. Then $\widehat{Y}$ agrees with Y on $\{\tau_2 > 1\}$, its covariance matrix remains between κI_d and ΛI_d, and its drift satisfies

$$\|\widehat{\beta}_\ell\| \le 2\, \mathbf{1}_{\{0 \le \ell \le 1\}} \mathbf{1}_{\{\widehat{Y}_\ell \in B_2\}}.$$

Thus Lemma B.4 applies conditionally on $\mathcal{F}_s$, with ellipticity constants κ, Λ.

Consequently, there exist constants $N < \infty$ and $\nu > 0$, depending only on d, κ, Λ, such that for every $q \in B_1$ and every $\rho \in (0, 1]$,

$$\mathbb{P}\left(Y_1 \in B_{2\rho}(q),\, \mathcal{E}_2^Y \,\middle|\, \mathcal{F}_s\right) \geq N^{-1}\rho^\nu.$$

Choose $m+1$ fixed points $q_0, \ldots, q_m \in B_1$ with positive minimum separation $\Delta := \min_{i \neq j} \|q_i - q_j\| > 0$. Set $r_0 := \frac{\Delta}{8}$. Since $q_i \in B_1$, we may choose the points so that $r_0 \leq 1/4$.

Let $y_1, \ldots, y_m \in \mathbb{R}^d$ be arbitrary. We claim that there exists $i \in \{0, \ldots, m\}$ such that

$$\operatorname{dist}(q_i, \{y_1, \ldots, y_m\}) > 2r_0.$$

If not, then each q_i is within distance $2r_0$ of at least one z_j. Since there are $m+1$ points q_i and only m points y_j, two distinct $q_i, q_{i'}$ must be within distance $2r_0$ of the same y_j. Then

$$\|q_i - q_{i'}\| \leq \|q_i - y_j\| + \|y_j - q_{i'}\| \leq 4r_0 = \frac{\Delta}{2},$$

which contradicts the definition of Δ. Thus such an i exists.

For this i,

$$B(q_i, r_0) \cap \bigcup_{j=1}^m B(y_j, r_0) = \varnothing.$$

We apply the lower estimate with $q = q_i, \rho = r_0/2$. Since $r_0 \leq 1/4$, $\rho \in (0, 1]$, and

$$B_{2\rho}(q_i) = B(q_i, r_0).$$

Therefore

$$\mathbb{P}\left(Y_1 \in B(q_i, r_0),\, \mathcal{E}_2^Y \,\middle|\, \mathcal{F}_s\right) \geq N^{-1}\left(\frac{r_0}{2}\right)^\nu.$$

Define

$$c_0 := N^{-1}\left(\frac{r_0}{2}\right)^\nu > 0.$$

Since $B(q_i, r_0)$ is disjoint from the target union,

$$\mathbb{P}\left(Y_1 \notin \bigcup_{j=1}^m B(y_j, r_0) \,\middle|\, \mathcal{F}_s\right) \geq c_0.$$

Now set $y_j := \sqrt{K}\left(z_j - \phi_{s+1/K}\right), \qquad j = 1, \ldots, m$. The event

$$\mathsf{dist}(\bar{a}_{s+1/K}, \mathcal{Z}) \leq \frac{r_0}{\sqrt{K}}$$

is exactly

$$Y_1 \in \bigcup_{j=1}^m B(y_j, r_0).$$

Therefore

$$\mathbb{P}\left(\mathsf{dist}(\bar{a}_{s+1/K}, \mathcal{Z}) > \frac{r_0}{\sqrt{K}} \,\middle|\, \mathcal{F}_s\right) \geq c_0.$$

For all K sufficiently large so that $r_0/\sqrt{K} \leq R$, we have on this event

$$\mathsf{dist}(\bar{a}_{s+1/K}, \mathcal{Z})^2 \wedge R^2 \geq \frac{r_0^2}{K}.$$

Consequently,

$$\begin{aligned}\mathbb{E}\left[\mathsf{dist}(\bar{a}_{s+1/K}, \mathcal{Z})^2 \wedge R^2 \,\middle|\, \mathcal{F}_s\right] &\geq \frac{r_0^2}{K}\mathbb{P}\left(\mathsf{dist}(\bar{a}_{s+1/K}, \mathcal{Z}) > \frac{r_0}{\sqrt{K}} \,\middle|\, \mathcal{F}_s\right) \\ &\geq \frac{c_0 r_0^2}{K}.\end{aligned}$$

Taking expectations gives, for every $t > K^{-1}$ and all sufficiently large K,

$$\mathbb{E}\left[\mathsf{dist}(\bar{a}_t, \mathcal{Z})^2 \wedge R^2\right] \geq \frac{c_0 r_0^2}{K}.$$

It remains to handle the one-dimensional action case $d = 1$. The local Krylov estimate above was stated for dimensions at least two, so we embed the scalar rescaled process into a two-dimensional one.

Work conditionally on $\mathcal{F}_s$. Enlarge the probability space, if necessary, by adding a one-dimensional Brownian motion G, independent of $\mathcal{F}_s$ and of the Brownian motion driving Y. Define the augmented process

$$\widehat{Y}_\ell := (Y_\ell, G_\ell) \in \mathbb{R}^2, \qquad 0 \leq \ell \leq 1.$$

If

$$dY_\ell = \beta_K(\ell, Y_\ell)d\ell + \widetilde{\sigma}_K(\ell, Y_\ell)dW_\ell,$$

then $\widehat{Y}$ satisfies

$$d\widehat{Y}_\ell = \widehat{\beta}_K(\ell, \widehat{Y}_\ell)d\ell + \widehat{\Sigma}_K(\ell, \widehat{Y}_\ell)d\widehat{W}_\ell,$$

where

$$\widehat{W}_\ell := (W_\ell, G_\ell), \qquad \widehat{\beta}_K(\ell, u, v) := (\beta_K(\ell, u), 0),$$

and

$$\widehat{\Sigma}_K(\ell, u, v) := \begin{pmatrix} \widetilde{\sigma}_K(\ell, u) & 0 \\ 0 & 1 \end{pmatrix}.$$

Therefore

$$\widehat{\Sigma}_K \widehat{\Sigma}_K^\top = \begin{pmatrix} \widetilde{\sigma}_K(\ell, u)^2 & 0 \\ 0 & 1 \end{pmatrix},$$

and Assumption 2.1 implies

$$(\kappa \wedge 1)I_2 \preceq \widehat{\Sigma}_K \widehat{\Sigma}_K^\top \preceq (\Lambda \vee 1)I_2.$$

Moreover, on the two-dimensional ball $B_2^{(2)} := \{(u, v) : u^2 + v^2 < 4\}$, we have $|u| \leq 2$. Since $\beta_K(\ell, \cdot)$ is 1-Lipschitz and $\beta_K(\ell, 0) = 0$,

$$\|\widehat{\beta}_K(\ell, u, v)\| = |\beta_K(\ell, u)| \leq |u| \leq 2 \qquad \text{on } B_2^{(2)}.$$

Thus the two-dimensional local Krylov lower estimate applies to $\widehat{Y}$, with constants depending only on κ and Λ.

We now perform the pigeonhole argument in the first coordinate. Choose $m + 1$ deterministic points $q_0, \ldots, q_m \in (-1, 1)$ with pairwise distance at least $\Delta_m > 0$, and set

$$r_0 := \Delta_m/8 \wedge 1/4.$$

Let $\zeta_1, \ldots, \zeta_m \in \mathbb{R}$ be arbitrary. Similar to before, since there are only m target points but $m + 1$ candidate points, there exists an index i such that

$$|q_i - \zeta_j| \geq 2r_0 \qquad \text{for all } j = 1, \ldots, m.$$

Equivalently,

$$(q_i - r_0, q_i + r_0) \cap \bigcup_{j=1}^{m} (\zeta_j - r_0, \zeta_j + r_0) = \varnothing.$$

Apply the two-dimensional Krylov estimate to the ball in $\mathbb{R}^2$ centered at $(q_i, 0)$ with radius r_0. Since the estimate gives a ball of radius 2ρ, take $\rho = r_0/2$. Then, conditionally on $\mathcal{F}_s$,

$$\mathbb{P}\Big(\widehat{Y}_1 \in B_{r_0}^{(2)}((q_i, 0)),\ \widehat{\mathcal{E}}_2 \,\Big|\, \mathcal{F}_s\Big) \geq N^{-1}(r_0/2)^\nu =: c_0,$$

where

$$\widehat{\mathcal{E}}_2 := \{\widehat{Y}_\ell \in B_2^{(2)} \text{ for all } 0 \leq \ell \leq 1\}.$$

Here the index i may be chosen $\mathcal{F}_s$-measurably; the bound is uniform over the finitely many possible choices of i.

On the event $\widehat{Y}_1 \in B_{r_0}^{(2)}((q_i, 0))$, the first coordinate satisfies $Y_1 \in (q_i - r_0, q_i + r_0)$. By the choice of q_i, this interval is disjoint from

$$\bigcup_{j=1}^{m} (\zeta_j - r_0, \zeta_j + r_0).$$

Hence $\operatorname{dist}(Y_1, \{\zeta_1, \dots, \zeta_m\}) \geq r_0$ on this event. Therefore

$$\mathbb{P}(\operatorname{dist}(Y_1, \{\zeta_1, \dots, \zeta_m\}) \geq r_0 \,|\, \mathcal{F}_s) \geq c_0.$$

Finally, take $\zeta_j := \sqrt{K}\,(z_j - \phi_T), j = 1, \dots, m$. Since

$$\bar{a}_T = \phi_T + \frac{Y_1}{\sqrt{K}},$$

we have the exact identity

$$\operatorname{dist}(\bar{a}_T, \mathcal{Z}) = \frac{1}{\sqrt{K}} \operatorname{dist}(Y_1, \{\zeta_1, \dots, \zeta_m\}).$$

Thus

$$\mathbb{P}\left(\operatorname{dist}(\bar{a}_T, \mathcal{Z}) \geq \frac{r_0}{\sqrt{K}} \,\middle|\, \mathcal{F}_s \right) \geq c_0.$$

For all sufficiently large K such that $r_0/\sqrt{K} \leq R$, it follows that

$$\mathbb{E}\big[\operatorname{dist}(\bar{a}_T, \mathcal{Z})^2 \wedge R^2 \,\big|\, \mathcal{F}_s\big] \geq \frac{r_0^2}{K} \mathbb{P}\left(\operatorname{dist}(\bar{a}_T, \mathcal{Z}) \geq \frac{r_0}{\sqrt{K}} \,\middle|\, \mathcal{F}_s \right) \geq \frac{c_0 r_0^2}{K}.$$

This proves the same clipped lower bound in dimension one. The constants depend only on m, κ, Λ, R and, when combined with the short-time case, on μ_0, but not on the locations or separations of the target points.

Combining the two regimes, for all sufficiently large K,

$$\mathbb{E}\left[\mathsf{dist}(\bar{a}_t, \mathcal{Z})^2 \wedge R^2\right] \geq \frac{1}{K} \min\left\{ \frac{\delta^2 \wedge R^2}{2}, c_0 r_0^2 \right\}.$$

Taking the infimum over all admissible f, σ, t proves the result. □

B.10 Lower bound failure without upper volatility control

Finally, we show that the upper bound $\sigma(t,x)\sigma(t,x)^\top \preceq \Lambda I_d$ in Assumption 2.1 is not merely technical for the lower-bound argument. Without this bound, state-dependent volatility can accelerate concentration near the target set and violate the K^{-1} lower bound.

Proposition B.5. *The upper bound $\sigma(t,x)\sigma(t,x)^\top \preceq \Lambda I_d$ in Proposition 3.3 cannot be dropped. In dimension $d = 1$, let*

$$A = \{-1, 1\}.$$

For every $K \geq 1$, there exists a conservative weak solution of

$$dX_t = \sigma_K(X_t) dB_t, \qquad X_0 \sim \mathcal{N}(0, 1),$$

such that the drift is identically zero, hence K-Lipschitz, and

$$\sigma_K(x)^2 \geq 1 \qquad \text{for all } x \in \mathbb{R},$$

but

$$\inf_{t \geq 0} \mathbb{E}[d(X_t, A)^2] \lesssim \frac{1}{K^2}.$$

Consequently, lower ellipticity alone does not imply the lower bound

$$\inf_{t} \mathbb{E}[d(X_t, A)^2] \gtrsim \frac{1}{K}.$$

The proof follows the classical speed-measure theory for one-dimensional diffusions. We use the standard Brownian time-change construction; see Karatzas & Shreve (1991, Ch. 5, §5.5A). We also use the fact that a regular one-dimensional diffusion in natural scale with finite speed measure has invariant probability measure given by the normalized speed measure and is ergodic; see Rogers & Williams (2000, Ch. V, §§47,53).

Proof. Let

$$r(x) := d(x, A), \qquad \varepsilon_K := K^{-1},$$

and define

$$q_K(x) := \frac{1}{1 + (r(x)/\varepsilon_K)^6}.$$

Set

$$\sigma_K(x)^2 := \frac{1}{q_K(x)} = 1 + \left(\frac{r(x)}{\varepsilon_K}\right)^6.$$

Then $\sigma_K(x)^2 \geq 1$ for all x, but $\sup_x \sigma_K(x) = \infty$.

We construct the diffusion by a time change of Brownian motion. This is a standard one-dimensional time-change construction; see Karatzas & Shreve (1991, Ch. 5, §5.5A) for the method of time change and see Rogers & Williams (2000, Ch. V, §47) for the speed-measure/time-substitution formulation. Specifically, let W be a one-dimensional Brownian motion with $W_0 \sim \mathcal{N}(0, 1)$, and define

$$S_u := \int_0^u q_K(W_s)\, ds.$$

Since q_K is strictly positive on neighborhoods of -1 and 1, and Brownian motion spends infinite occupation time in every nonempty open interval,

$$S_u \to \infty \qquad \text{a.s. as } u \to \infty.$$

Hence the inverse clock $\tau_t := \inf\{u : S_u > t\}$ is finite for every finite t. Define $X_t := W_{\tau_t}$. Then X_t is finite for every finite t, so the process is conservative. Moreover,

$$d\langle X\rangle_t = d\tau_t = \frac{1}{q_K(X_t)} dt = \sigma_K(X_t)^2 dt.$$

Therefore, by the martingale representation theorem, X is a weak solution of $dX_t = \sigma_K(X_t) dB_t$.

Now define $Z_K := \int_{\mathbb{R}} q_K(x)\, dx$ and $\pi_K(dx) := Z_K^{-1} q_K(x)\, dx$. Since $q_K(x) \asymp \varepsilon_K^6 |x|^{-6}$ as $|x| \to \infty$, $Z_K < \infty$. The time-changed process is a regular diffusion in natural scale with speed measure $m_K(dx) = 2q_K(x)\, dx$. Thus π_K is the normalized speed measure. Since $m_K(\mathbb{R}) < \infty$, the diffusion is positive recurrent and ergodic; see Rogers & Williams (2000, Ch. V, §§47,53). Let $P_t h(x) := \mathbb{E}_x[h(X_t)]$ denote the Markov semigroup of X. By ergodicity with invariant law π_K,

$$P_t h \to \int h\, d\pi_K \qquad \text{in } L^2(\pi_K), \qquad h \in L^2(\pi_K).$$

Let γ be the $\mathcal{N}(0, 1)$ law and let φ be its density. Since π_K has positive density everywhere,

$$h_K := \frac{d\gamma}{d\pi_K} = \frac{Z_K \varphi}{q_K}.$$

Because q_K^{-1} grows only polynomially while φ decays Gaussianly, we have $h_K \in L^2(\pi_K)$. Also

$$f(x) := d(x, A)^2 = r(x)^2$$

belongs to $L^2(\pi_K)$, since in the tails

$$r(x)^4 q_K(x) \asymp \varepsilon_K^6 |x|^{-2},$$

which is integrable. Therefore, by symmetry and $L^2(\pi_K)$ ergodicity,

$$\mathbb{E}_\gamma[f(X_t)] = \langle h_K, P_t f\rangle_{L^2(\pi_K)} = \langle P_t h_K, f\rangle_{L^2(\pi_K)} \to \int f\, d\pi_K.$$

Thus the invariant expectation is the actual limit of finite-time expectations from the required Gaussian initialization.

It remains to estimate $\int f\, d\pi_K$. Since $r(x) = d(x, A)$ has level sets of uniformly bounded multiplicity on $\mathbb{R}$,

$$Z_K = \int_{\mathbb{R}} \frac{dx}{1 + (r(x)/\varepsilon_K)^6} \asymp \varepsilon_K,$$

and

$$\int_{\mathbb{R}} r(x)^2 q_K(x)\, dx \lesssim \int_0^\infty \frac{s^2}{1 + (s/\varepsilon_K)^6}\, ds \asymp \varepsilon_K^3.$$

Therefore

$$\int d(x, A)^2\, \pi_K(dx) = \frac{\int r(x)^2 q_K(x)\, dx}{Z_K} \lesssim \varepsilon_K^2 = \frac{1}{K^2}.$$

Since $\mathbb{E}_\gamma[d(X_t, A)^2] \to \int d(x, A)^2\, \pi_K(dx)$, there exists a finite deterministic time t_K such that $\mathbb{E}_\gamma[d(X_{t_K}, A)^2] \lesssim \frac{1}{K^2}$. Hence

$$\inf_{t \geq 0} \mathbb{E}[d(X_t, A)^2] \lesssim \frac{1}{K^2},$$

which proves the claim. □

C Proofs for Section 4

C.1 Drift construction and detailed finite-sample bound

We give the architecture and sampling construction used in Section 4, followed by the full oracle inequality and the dependence of the rates on the state and action dimensions.

ReLU drift classes. We first introduce some constants that will be used to construct the policy class. Fix a sufficiently large constant $C_0 \geq 2$, independent of $(n, K, s, m, d, T, \delta)$. Choose constants $\ell_1, \ell_2 < \infty$, independent of (K, s), satisfying

$$\ell_1 \geq \frac{\sqrt{d}\, R_0}{1 - e^{-T}}, \qquad \ell_2 \geq \frac{C_0^m \sqrt{d}\, L_{g^\star}}{1 - e^{-T}}, \tag{C.1}$$

where R_0 is given in Assumption 4.1. For each $s \geq 1$, define

$$p_{s,1} := C_0^m d(m + d + 1)s \log(es), \tag{C.2}$$
$$p_{s,2} := C_0^m \log(es), \tag{C.3}$$
$$p_{s,3} := C_0^m d(m + d + 1)s, \tag{C.4}$$
$$p_{s,4} := C_0^m \{1 + \sqrt{d}(R_0 + L_{g^\star})\}(1 + s)^{C_0^m} (1 - e^{-T})^{-1}. \tag{C.5}$$

Let $\mathcal{N}_s$ be the set of vector-valued multi-layer ReLU networks $\Psi_\theta : [0, T] \times \mathbb{R}^d \times [0, 1]^m \to \mathbb{R}^d$ satisfying the following conditions: Ψ_θ has at most $p_{s,1}$ nonzero scalar parameters, depth at most $p_{s,2}$, width at most $p_{s,3}$, and parameter bound $|\theta_i| \leq p_{s,4}$. For every affine region S intersecting $[0, T] \times \mathbb{R}^d \times [0, 1]^m$,

$$\Psi_\theta(t, a, x) = M^1_{\theta,S} t + M^2_{\theta,S} a + M^3_{\theta,S} x + b_{\theta,S}, \tag{C.6}$$

for some $M^2_{\theta,S} \in \mathbb{R}^{d \times d}$ and $M^3_{\theta,S} \in \mathbb{R}^{d \times m}$ satisfying

$$\|M^2_{\theta,S}\|_{\mathrm{op}} \leq 1, \qquad \|M^3_{\theta,S}\|_{\mathrm{op}} \leq \ell_2,$$

and

$$\sup_{t \in [0,T],\, x \in [0,1]^m} \|\Psi_\theta(t, 0, x)\| \leq \ell_1.$$

Define the drift class

$$\mathcal{F}_{K,s} := \Big\{ f_\theta(t, a, x) = K \Psi_\theta(t, a, x) : \Psi_\theta \in \mathcal{N}_s \Big\}. \tag{C.7}$$

The architecture bounds, especially C_0^m, in the definition of $\mathcal{F}_{K,s}$ are chosen so that the class contains the mean-reverting drift of the form (3.2). Specifically, Lemma C.1 shows that for all $K \geq 1$, $\mathcal{F}_{K,s}$ contains

$$f_{K,s}(t,a,x) = K\left(\frac{h_s(x)}{1-e^{-KT}} - a\right), \tag{C.8}$$

where h_s is a suitable ReLU network of size s approximating the optimal selector $g^\star$. Thus the finite-sample analysis can use the approximation property from Section 3 for mean-reverting drifts.

Lemma C.1 (ReLU approximation and mean-reverting drift)**.** *Under Assumption 4.1, there exists a constant $C_0 \geq 0$ such that for every $s \geq 1$ there is a vector-valued ReLU network $h_s : [0,1]^m \to \mathbb{R}^d$ satisfying*

$$\sup_{x\in[0,1]^m} \|h_s(x) - g^\star(x)\|^2 \leq C_0^m m d L_{g^\star}^2 s^{-2/m}, \tag{C.9}$$

$$\sup_{x\in[0,1]^m} \|h_s(x)\| \leq \sqrt{d}\, R_0, \tag{C.10}$$

$$\mathrm{Lip}(h_s) \leq C_0^m \sqrt{d}\, L_{g^\star}. \tag{C.11}$$

Moreover, h_s has at most (i) $C_0^m ds \log(es)$ nonzero scalar parameters, (ii) depth at most $C_0^m \log(es)$, (iii) width at most $C_0^m ds$, and (iv)weights bounded by $C_0^m(1 + R_0 + L_{g^\star})(1+s)^{C_0^m}$.

It holds that $f_{K,s} \in \mathcal{F}_{K,s}$, where $f_{K,s}$ is defined in (C.8)*.*

The proof is given in Appendix C.2.

Induced diffusion policies and statistical samples. For a drift $f \in \mathcal{F}_{K,s}$, let $\pi_f(\cdot \mid x)$ denote the terminal law at time T of the following SDE: for each fixed x, write $\bar{a}^x$ for the solution of

$$d\bar{a}_t^x = f(t, \bar{a}_t^x, x)\, dt + \sigma(t, \bar{a}_t^x, x)\, dB_t, \qquad \bar{a}_0^x \sim \mu_0.$$

The dependence on f is only through the displayed drift. Write $V(f) := V^{\pi_f}(\rho)$. For the empirical value function, fix i.i.d. random variables

$$\xi_i := \big(X_{i,0}, \{\zeta_{i,j}, B_{i,j}, U_{i,j}\}_{j\geq 0}\big), \qquad i = 1, \ldots, n,$$

where $X_{i,0} \sim \rho$, $\zeta_{i,j} \sim \mu_0$, the $B_{i,j}$'s are independent Brownian motions on $[0,T]$, and $U_{i,j} \sim \mathrm{Unif}[0,1]$, all mutually independent. Given f, set $X_{i,0}^f := X_{i,0}$. Recursively, given $X_{i,j}^f$, let $\bar{a}_{i,j}$ solve

$$d\bar{a}_{i,j,t} = f(t, \bar{a}_{i,j,t}, X_{i,j}^f)\, dt + \sigma(t, \bar{a}_{i,j,t}, X_{i,j}^f)\, dB_{i,j,t}, \qquad \bar{a}_{i,j,0} = \zeta_{i,j},$$

and define

$$A_{i,j}^f := \bar{a}_{i,j,T}, \qquad X_{i,j+1}^f := \Phi_P\big(X_{i,j}^f, A_{i,j}^f, U_{i,j}\big).$$

Then, for each fixed f, $\{(X_{i,j}^f, A_{i,j}^f)\}_{j\geq 0}$ follows the MDP trajectory law induced by π_f. The empirical value $\widehat{V}_n(f)$ is defined by (4.3). For each fixed f, $\widehat{V}_n(f)$ is an unbiased estimator of $V(f)$. Using the same fixed exogenous variables $\xi_1, \ldots, \xi_n$ for all $f \in \mathcal{F}_{K,s}$ makes $f \mapsto \widehat{V}_n(f)$ a well-defined random function, which is used in the proof to control the uniform deviation over $f \in \mathcal{F}_{K,s}$.

Define $r_{\max} := \|r\|_\infty$, set $H_n = 1$ if $\gamma = 0$, and $H_n := \max\left\{1, \left\lceil \frac{\log(8\sqrt{n})}{\log(1/\gamma)} \right\rceil\right\}$ if $0 < \gamma < 1$. Define the logarithmic factor

$$\begin{aligned}\Gamma_n(K,s,m,d,T) := \log\Big(1 + 2n^2\Big\{1 + T + \sqrt{m} + \frac{r_{\max}}{1-\gamma}\sqrt{H_n} \\ \times \big(m_2(\mu_0) + K^2\ell_1^2 T^2 + d\Lambda T\big)^{1/2} e^{C_3 KT}\Big\} \\ \times (p_{s,1}p_{s,4})^{p_{s,2}+2} C_3 H_n \Big\{C_3\big[1 + (1+\ell_2)\sqrt{KT}\, e^{C_3 KT}\big]\Big\}^{H_n}\Big),\end{aligned} \tag{C.12}$$

where $C_3 \geq 1$ is a sufficiently large constant depending only on $(L_\sigma, L_r^x, L_r^a, L_P^x, L_P^a, \gamma)$. Then we have the following main result.

Theorem C.2 (Finite-sample oracle inequality for Lipschitz drifts)**.** *Assume Assumptions 2.1, 4.1, and 4.2 hold. Suppose also that the quadratic Bellman upper bound* (4.4) *holds. Then there exists a constant $C_1 < \infty$, independent of $(n, K, s, m, d, T, \delta)$, such that, for all $K \geq 1$, $s \geq 1$, and $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$V^\star(\rho) - V(\widehat{f}_{n,K,s}) \leq C_1 \Bigg[e^{-2KT} m_2(\mu_0) + \frac{d\Lambda}{K} + C_0^m m d L_{g^\star}^2 s^{-2/m} + \frac{r_{\max}}{1-\gamma} \sqrt{\frac{p_{s,1}\Gamma_n(K, s, m, d, T) + \log(2/\delta)}{n}} + \frac{r_{\max}}{1-\gamma} \frac{p_{s,1}\Gamma_n(K, s, m, d, T) + \log(2/\delta)}{n} \Bigg]. \tag{C.13}$$

The constant C_1 depends only on $(L_Q, L_\sigma, L_r^x, L_r^a, L_P^x, L_P^a, \Lambda, \gamma)$ and numerical constants.

The proof is given in Appendix C.3. For fixed dimensions and problem parameters, $p_{s,1} = \widetilde{O}(s)$ and $\Gamma_n(K, s, m, d, T) = \widetilde{O}(K)$, while $e^{-2KT} m_2(\mu_0) = O(K^{-1})$ for $K \geq 1$. Thus (C.13) implies Theorem 4.1, including the linear statistical term in (4.5).

Remark C.1 (Balancing the leading terms)**.** *Assume $m_2(\mu_0) \leq C_2 d$ and treat all problem parameters other than (m, d, K, s, n) as fixed. Since $p_{s,1} = \widetilde{O}(C_0^m d(m + d + 1)s)$ and $\Gamma_n(K, s, m, d, T) = \widetilde{O}(KT)$, the leading terms in* (C.13)*, after suppressing logarithmic and fixed problem-dependent factors, are*

$$\frac{d}{K}, \qquad C_0^m m d\, s^{-2/m}, \qquad \sqrt{\frac{C_0^m d(m + d + 1)sK}{n}}.$$

Balancing the first two terms gives $s \asymp (C_0^m mK)^{m/2}$. Substituting this choice into the statistical term and balancing with d/K yields $K \asymp \left(\frac{dn}{C_0^m(m+d+1)(C_0^m m)^{m/2}}\right)^{2/(m+6)}$ and

$$\text{(C.13)} = \widetilde{\mathcal{O}}\left(d^{(m+4)/(m+6)} \left(\frac{C_0^m(m + d + 1)(C_0^m m)^{m/2}}{n} \right)^{2/(m+6)} \right). \tag{C.14}$$

In particular, the exponent of n is $-2/(m + 6)$.

Remark C.2 (Dissipative drift classes and better rate)**.** *Imposing a one-sided dissipativity condition on the drift class $\mathcal{F}_{K,s}$ can improve the error bound* (C.14)*. Specifically, restrict the class to drifts satisfying* (4.6)*. This condition is satisfied by the mean-reverting drift $f_{K,s}$ defined in* (C.8)*, for which $\lambda = 1$.*

In this case, the approximation error terms in (C.13) *are unchanged, but the statistical error term reduces to $\sqrt{\frac{C_0^m d(m+d+1)s}{n}}$, compared with $\sqrt{\frac{C_0^m d(m+d+1)sK}{n}}$ in the generic K-Lipschitz case. Thus the leading term in* (C.13) *becomes*

$$\frac{d}{K} + C_0^m m d s^{-2/m} + \sqrt{\frac{C_0^m d(m + d + 1)s}{n}}.$$

Since the last two terms do not depend on K, balancing them gives $s \asymp \left(\frac{C_0^m m^2 dn}{m+d+1}\right)^{m/(m+4)}$. For this choice, it suffices to take

$$K \gtrsim \left(\frac{dn}{C_0^m(m + d + 1)(C_0^m m)^{m/2}} \right)^{2/(m+4)},$$

so that d/K is no larger than the other two terms. Thus, up to logarithmic factors and fixed problem constants, the exponent of n improves from $-2/(m + 6)$ to $-2/(m + 4)$. Here, the displayed order of K is a minimum sufficient scale rather than an optimizer: any larger K achieves the same leading rate. This sharper rate comes from the stronger condition (4.6)*; without it, Theorem 4.1 applies to the larger generic ReLU drift class, but retains the extra K-dependence in the statistical term.*

C.2 Auxiliary facts about ReLU network class

Here we present some elementary properties of $\mathcal{F}_{K,s}$, which will be used in the proof of Theorem 4.1.

Lemma C.3 (Properties of the ReLU drift class)**.** *Every $f \in \mathcal{F}_{K,s}$ satisfies, for all $t \in [0,T]$, $a, a' \in \mathbb{R}^d$ and $x, x' \in [0,1]^m$,*

$$\|f(t,0,x)\| \leq \ell_1 K, \tag{C.15}$$
$$\|f(t,a,x) - f(t,a',x)\| \leq K\|a - a'\|, \tag{C.16}$$
$$\|f(t,a,x) - f(t,a,x')\| \leq \ell_2 K\|x - x'\|. \tag{C.17}$$

Consequently, $\|f(t,a,x)\| \leq K(\ell_1 + \|a\|)$. Under Assumptions 2.1 and 4.2, the state-conditioned action SDE has a unique strong solution for every $f \in \mathcal{F}_{K,s}$ and $x \in [0,1]^m$.

Proof. Note that (C.15) is part of the definition. For the action Lipschitz bound, fix t, x, a, a', set $v = a - a'$, and write $a_r = a' + rv$. Since Ψ_θ is continuous and piecewise affine, $r \mapsto \Psi_\theta(t, a_r, x)$ is absolutely continuous and

$$\Psi_\theta(t,a,x) - \Psi_\theta(t,a',x) = \int_0^1 J_a\Psi_\theta(t,a_r,x)v\,dr,$$

where $J_a\Psi_\theta$ is the action-slope matrix on the relevant affine region for a.e. r. The bound $\|M^2_{\theta,S}\|_{\mathrm{op}} \leq 1$ gives (C.16). The proof of (C.17) is identical along the line segment between x and x', using $\|M^3_{\theta,S}\|_{\mathrm{op}} \leq \ell_2$. The linear-growth bound follows from (C.15) and (C.16). Finally, the linear-growth and global Lipschitz bounds on f, together with the boundedness and Lipschitz assumptions on σ, give the stated strong well-posedness. □

Proof of Lemma C.1. The existence and approximation properties of h_s follow from the standard finite-element/ReLU approximation result for Lipschitz maps on the cube, obtained by nodal interpolation on a Kuhn triangulation and exact ReLU realization of continuous piecewise-affine finite-element functions; see Yarotsky (2017); Petersen & Voigtlaender (2018). Applying the scalar construction coordinate by coordinate and summing the squared coordinate errors gives (C.9). The bounds (C.10)–(C.11) follow from the boundedness and Lipschitz properties of the clipped finite-element interpolant.

It remains to verify that the associated mean-reverting reference drift belongs to $\mathcal{F}_{K,s}$. The map $x \mapsto h_s(x)$ is a ReLU network. The affine map $a \mapsto -a$ is exactly representable by ReLUs since

$$-a_j = \mathrm{ReLU}(-a_j) - \mathrm{ReLU}(a_j), \qquad j = 1, \ldots, d.$$

Thus (C.8) is a ReLU network in (t,a,x), with no dependence on t. The additional action-identity branch and final affine readout increase the size, depth, width, and weight bounds by numerical factors. The factor $(1-e^{-T})^{-1}$ in (C.5) covers $(1-e^{-KT})^{-1}$ because $K \geq 1$, and the constant C_0 in the architecture bounds is chosen large enough to absorb these fixed realization overheads.

The action slope is $-I_d$, so its operator norm is 1. By (C.11),

$$\mathrm{Lip}\left(\frac{h_s}{1-e^{-KT}}\right) \leq \frac{C_0^m\sqrt{d}\,L_{g^\star}}{1-e^{-T}} \leq \ell_2.$$

Also, by (C.10),

$$\sup_{t,x}\left\|\frac{h_s(x)}{1-e^{-KT}}\right\| \leq \frac{\sqrt{d}\,R_0}{1-e^{-T}} \leq \ell_1.$$

Hence $f_{K,s} \in \mathcal{F}_{K,s}$. □

C.3 Proof of Theorem C.2

We organize the proof into five parts. First, we reduce empirical maximization to a comparison inequality and bound the comparison term using the mean-reverting reference drift. The next three parts establish the localization, covering, and finite-horizon stability estimates. The final part converts these estimates into a uniform-convergence bound and substitutes it into the comparison inequality.

Step 1: Empirical reduction and reference-drift approximation. For $f \in \mathcal{F}_{K,s}$, write

$$G_f(\xi) := \sum_{j=0}^{\infty} \gamma^j r(X_j^f, A_j^f), \qquad PG_f := \mathbb{E}G_f(\xi) = V(f),$$

$$P_n G_f := \frac{1}{n}\sum_{i=1}^{n} G_f(\xi_i) = \widehat{V}_n(f).$$

For every f, $|G_f| \le r_{\max}/(1-\gamma)$. Let $f_0 \in \mathcal{F}_{K,s}$ be arbitrary. Since $\widehat{f}_{n,K,s}$ is an exact empirical maximizer,

$$P_n G_{\widehat{f}_{n,K,s}} \ge P_n G_{f_0}.$$

Therefore,

$$V^\star(\rho) - V(\widehat{f}_{n,K,s}) \le V^\star(\rho) - V(f_0) + 2 \sup_{f\in\mathcal{F}_{K,s}} |(P_n - P)G_f|\,. \tag{C.18}$$

If an η-maximizer is used instead, the right-hand side gains η.

We next record the performance identity used to control the comparison term. For every Markov policy π and every $x \in \mathcal{X}$,

$$V^\star(x) - V^\pi(x) = \mathbb{E}_x^\pi \sum_{j=0}^{\infty} \gamma^j \{V^\star(X_j) - Q^\star(X_j, A_j)\}. \tag{C.19}$$

Indeed,

$$\mathbb{E}_x^\pi\{V^\star(X_0) - Q^\star(X_0,A_0)\} = V^\star(x) - \mathbb{E}_x^\pi r(X_0,A_0) - \gamma \mathbb{E}_x^\pi V^\star(X_1),$$

and

$$V^\pi(x) = \mathbb{E}_x^\pi r(X_0,A_0) + \gamma \mathbb{E}_x^\pi V^\pi(X_1).$$

Adding the two displays yields the one-step identity

$$V^\star(x) - V^\pi(x) = \mathbb{E}_x^\pi\{V^\star(X_0) - Q^\star(X_0,A_0)\} + \gamma \mathbb{E}_x^\pi\{V^\star(X_1) - V^\pi(X_1)\}.$$

Iteration for H steps gives

$$V^\star(x) - V^\pi(x) = \mathbb{E}_x^\pi \sum_{j=0}^{H-1} \gamma^j \{V^\star(X_j) - Q^\star(X_j,A_j)\} + \gamma^H \mathbb{E}_x^\pi\{V^\star(X_H) - V^\pi(X_H)\}.$$

Since $|V^\star| \vee |V^\pi| \le r_{\max}/(1-\gamma)$, the remainder tends to zero as $H \to \infty$, proving (C.19).

We now apply the above identity to the mean-reverting reference drift. Let h_s and $f_{K,s}$ be as in (C.8) and Lemma C.1, and set $f_0 = f_{K,s}$. By Lemma C.1, $f_0 \in \mathcal{F}_{K,s}$. Fix $x \in [0,1]^m$. Let $\bar{a}_t^x$ denote the solution of the action diffusion driven by f_0 at state x. Define

$$\mathfrak{m}_t(x) := \frac{1 - e^{-Kt}}{1 - e^{-KT}} h_s(x), \qquad 0 \le t \le T,$$

and set $Y_t := \bar{a}_t^x - \mathfrak{m}_t(x)$. Since $\mathfrak{m}_T(x) = h_s(x)$, direct differentiation gives

$$dY_t = -KY_t\,dt + \sigma(t, \bar{a}_t^x, x) dB_t.$$

Itô's formula and Assumption 2.1 imply

$$\frac{d}{dt}\mathbb{E}\|Y_t\|^2 \le -2K\mathbb{E}\|Y_t\|^2 + d\Lambda, \qquad \mathbb{E}\|Y_0\|^2 = m_2(\mu_0).$$

Gronwall's inequality yields

$$\sup_{x\in[0,1]^m} \mathbb{E}\|\bar{a}_T^x - h_s(x)\|^2 \le e^{-2KT} m_2(\mu_0) + \frac{d\Lambda}{2K}(1 - e^{-2KT}). \tag{C.20}$$

Combining (C.20) with Lemma C.1,

$$\sup_{x\in[0,1]^m} \mathbb{E}\|\bar{a}_T^x - g^\star(x)\|^2 \le C\left(e^{-2KT} m_2(\mu_0) + \frac{d\Lambda}{K} + C_0^m m d L_{g^\star}^2 s^{-2/m}\right). \tag{C.21}$$

By (C.19) and (4.4),

$$V^\star(\rho) - V(f_0) \le C\left(e^{-2KT} m_2(\mu_0) + \frac{d\Lambda}{K} + C_0^m m d L_{g^\star}^2 s^{-2/m}\right), \tag{C.22}$$

where C depends only on the fixed problem constants in Theorem C.2.

Step 2: Path localization under second moments. For the localization estimates, write

$$D_{K,T} := \left(m_2(\mu_0) + K^2\ell_1^2 T^2 + d\Lambda T\right)^{1/2}.$$

We first prove a path localization estimate.

Lemma C.4 (Path localization)**.** *For every $f \in \mathcal{F}_{K,s}$ and every $x \in [0,1]^m$,*

$$\mathbb{E}\left[\sup_{0\le t\le T} \|\bar{a}_t^{f,x}\|^2\right] \le Ce^{C_3 KT} D_{K,T}^2. \tag{C.23}$$

Consequently, for every $R \ge 1$,

$$\mathbb{P}\left(\sup_{0\le t\le T} \|\bar{a}_t^{f,x}\| > R\right) \le \frac{Ce^{C_3 KT} D_{K,T}^2}{R^2}. \tag{C.24}$$

Proof. Fix $f \in \mathcal{F}_{K,s}$ and $x \in [0,1]^m$. By Lemma C.3,

$$\|f(t,a,x)\| \le K(\ell_1 + \|a\|).$$

Let

$$S_t := \sup_{0\le u\le t} \|\bar{a}_u^{f,x}\|, \qquad M_t := \int_0^t \sigma(u, \bar{a}_u^{f,x}, x) dB_u.$$

The integral form of the SDE gives

$$S_T \le \|\bar{a}_0\| + K\ell_1 T + K\int_0^T S_u\, du + \sup_{0\le u\le T} \|M_u\|.$$

By Gronwall's inequality,

$$S_T \le e^{KT}\left(\|\bar{a}_0\| + K\ell_1 T + \sup_{0\le u\le T} \|M_u\|\right).$$

Therefore,

$$\mathbb{E}S_T^2 \le Ce^{C_3 KT}\left(m_2(\mu_0) + K^2\ell_1^2 T^2 + \mathbb{E}\sup_{0\le u\le T} \|M_u\|^2\right).$$

Doob's L^2 maximal inequality and Itô's isometry give

$$\mathbb{E}\sup_{0\le u\le T} \|M_u\|^2 \le 4\mathbb{E}\|M_T\|^2 \le 4d\Lambda T.$$

Thus (C.23) follows from the definition of $D_{K,T}$, and (C.24) follows from Markov's inequality. □

Step 3: Localized covering of the drift class. For $R \ge 1$, define

$$Z_R := [0,T] \times B_R^d \times [0,1]^m, \qquad d_{K,R}(f,g) := \sup_{z\in Z_R} \frac{\|f(z) - g(z)\|}{K},$$

where B_R^d is the Euclidean ball in $\mathbb{R}^d$. Also define

$$\rho_R := 1 + T + R + \sqrt{m}. \tag{C.25}$$

Lemma C.5 (Localized covering number)**.** *For every $R \ge 1$ and every $0 < \varepsilon \le 1$,*

$$\log \mathcal{N}(\varepsilon, \mathcal{F}_{K,s}, d_{K,R}) \le p_{s,1} \log\left(1 + \frac{2\rho_R (p_{s,1} p_{s,4})^{p_{s,2}+2}}{\varepsilon}\right). \tag{C.26}$$

Moreover, after replacing ε by 2ε, the cover may be chosen to consist of elements of $\mathcal{F}_{K,s}$.

Proof. Consider the ambient parameter box corresponding to networks with at most $p_{s,1}$ nonzero scalar parameters, depth at most $p_{s,2}$, width at most $p_{s,3}$, and parameter magnitudes bounded by $p_{s,4}$. The standard layer-by-layer telescoping estimate for ReLU networks on bounded input sets gives, for any two parameter vectors θ, θ' in this ambient box,

$$\sup_{z \in Z_R} \|\Psi_\theta(z) - \Psi_{\theta'}(z)\| \le \rho_R (p_{s,1} p_{s,4})^{p_{s,2}+1} \|\theta - \theta'\|_\infty. \tag{C.27}$$

This deterministic perturbation estimate is a standard ingredient in covering-number bounds for ReLU networks; see Bartlett et al. (2019); Petersen & Voigtlaender (2018). Thus an ℓ_∞ grid on $[-p_{s,4}, p_{s,4}]^{p_{s,1}}$ with mesh size

$$\eta := \varepsilon \{\rho_R (p_{s,1} p_{s,4})^{p_{s,2}+1}\}^{-1}$$

induces an external ε-cover of the ambient network class on Z_R. Its cardinality is at most

$$\left(1 + \frac{2p_{s,4}}{\eta}\right)^{p_{s,1}} \le \left(1 + \frac{2\rho_R (p_{s,1} p_{s,4})^{p_{s,2}+2}}{\varepsilon}\right)^{p_{s,1}}.$$

Since $\mathcal{F}_{K,s}$ is a subclass of this ambient class, the same upper bound holds for the external covering number of $\mathcal{F}_{K,s}$ under $d_{K,R}$. Equivalently, it bounds the ε-packing number of $\mathcal{F}_{K,s}$. A maximal 2ε-separated subset of $\mathcal{F}_{K,s}$ is a 2ε-cover by elements of $\mathcal{F}_{K,s}$ and has cardinality bounded by the preceding packing bound. This proves (C.26). □

Step 4: Finite-horizon stability. For the stability estimates, write

$$L_{K,T} := 1 + (1 + \ell_2)\sqrt{KT}\, e^{C_3 KT}, \qquad M_{K,H,T} := C_3 H \{C_3 L_{K,T}\}^H.$$

For $H \ge 1$, define

$$G_H(f) := \sum_{j=0}^{H-1} \gamma^j r(X_j^f, A_j^f).$$

Let $\Pi_R : \mathbb{R}^d \to B_R^d$ denote Euclidean projection. For $f \in \mathcal{F}_{K,s}$ define

$$f^R(t, a, x) := f(t, \Pi_R(a), x), \qquad \sigma^R(t, a, x) := \sigma(t, \Pi_R(a), x).$$

Let $\widetilde{G}_H^R(f)$ denote the H-step return generated by replacing, at every within-step action diffusion, (f, σ) with (f^R, σ^R).

Lemma C.6 (Projected return stability)**.** *For all $f, g \in \mathcal{F}_{K,s}$,*

$$\|\widetilde{G}_H^R(f) - \widetilde{G}_H^R(g)\|_{L^2(P)} \le M_{K,H,T}\, d_{K,R}(f, g). \tag{C.28}$$

Proof. Fix $f, g \in \mathcal{F}_{K,s}$ and two states $x, x' \in [0,1]^m$. On the same probability space, let $\zeta \sim \mu_0$ and let B be a Brownian motion. Define $\widetilde{a}^{f,x}$ and $\widetilde{a}^{g,x'}$ as the projected within-step action diffusions

$$\begin{aligned} d\widetilde{a}_t^{f,x} &= f^R(t, \widetilde{a}_t^{f,x}, x)\, dt + \sigma^R(t, \widetilde{a}_t^{f,x}, x)\, dB_t, \qquad \widetilde{a}_0^{f,x} = \zeta, \\ d\widetilde{a}_t^{g,x'} &= g^R(t, \widetilde{a}_t^{g,x'}, x')\, dt + \sigma^R(t, \widetilde{a}_t^{g,x'}, x')\, dB_t, \qquad \widetilde{a}_0^{g,x'} = \zeta. \end{aligned}$$

Set $Z_t := \widetilde{a}_t^{f,x} - \widetilde{a}_t^{g,x'}$. Then $Z_0 = 0$.

Using Lemma C.3, the Lipschitz property of Π_R, the definition of $d_{K,R}$, and Assumption 4.2, Itô's formula gives

$$\frac{d}{dt}\mathbb{E}\|Z_t\|^2 \le C_3 K \mathbb{E}\|Z_t\|^2 + C_3 K (1 + \ell_2)^2 \|x - x'\|^2 + C_3 K d_{K,R}(f, g)^2.$$

Since $Z_0 = 0$, Gronwall's inequality gives

$$\left(\mathbb{E}\|Z_T\|^2\right)^{1/2} \le C_3 (1 + \ell_2)\sqrt{KT}\, e^{C_3 KT} \left\{\|x - x'\| + d_{K,R}(f, g)\right\}.$$

Let

$$D_j := \left(\mathbb{E}\|\widetilde{X}_j^f - \widetilde{X}_j^g\|^2\right)^{1/2}, \qquad E_j := \left(\mathbb{E}\|\widetilde{A}_j^f - \widetilde{A}_j^g\|^2\right)^{1/2}.$$

The preceding estimate implies

$$E_j \le C_3 L_{K,T}\{D_j + d_{K,R}(f,g)\}.$$

By the synchronous L^2 representation of the transition kernel in Assumption 4.2,

$$D_{j+1} \le C_3 L_{K,T}\{D_j + d_{K,R}(f,g)\}.$$

Since $D_0 = 0$, induction gives, for all $0 \le j \le H$,

$$D_j + E_j \le C_3\big(C_3 L_{K,T}\big)^H d_{K,R}(f,g).$$

Using the Lipschitz continuity of r and summing over $j = 0, \dots, H-1$,

$$\|\widetilde{G}_H^R(f) - \widetilde{G}_H^R(g)\|_{L^2(P)} \le C_3 H\big(C_3 L_{K,T}\big)^H d_{K,R}(f,g),$$

which is (C.28). □

Lemma C.7 (Original return stability)**.** *For all $H \ge 1$, all $R \ge 1$, and all $f, g \in \mathcal{F}_{K,s}$,*

$$\|G_H(f) - G_H(g)\|_{L^2(P)} \le M_{K,H,T} d_{K,R}(f,g) + C\frac{r_{\max}}{1-\gamma}\frac{\sqrt{H}\, D_{K,T} e^{C_3 KT}}{R}. \tag{C.29}$$

Proof. Fix $f \in \mathcal{F}_{K,s}$. For one decision step, Lemma C.4 gives, uniformly over $x \in [0,1]^m$,

$$\mathbb{P}\left(\sup_{0\le t\le T} \|\bar{a}_t^{f,x}\| > R\right) \le \frac{C e^{C_3 KT} D_{K,T}^2}{R^2}.$$

For an H-step trajectory, let $E_R^f(H)$ be the event that at least one of the H within-step action paths exits B_R^d. By the union bound,

$$\mathbb{P}(E_R^f(H)) \le \frac{CH e^{C_3 KT} D_{K,T}^2}{R^2}.$$

On the complement of $E_R^f(H)$, the original trajectory and projected trajectory agree through time H. Since both $G_H(f)$ and $\widetilde{G}_H^R(f)$ are bounded in absolute value by $r_{\max}/(1-\gamma)$,

$$\|G_H(f) - \widetilde{G}_H^R(f)\|_{L^2(P)} \le C\frac{r_{\max}}{1-\gamma}\frac{\sqrt{H}\, D_{K,T} e^{C_3 KT}}{R}.$$

The same bound holds for g. The triangle inequality and Lemma C.6 prove (C.29). □

Step 5: Entropy of truncated returns and uniform convergence. Let $\mathcal{G}_H := \{G_H(f) : f \in \mathcal{F}_{K,s}\}$. Fix $0 < \varepsilon \le r_{\max}/(1-\gamma)$ and define

$$R_{\varepsilon,H} := \max\left\{1, \frac{C(r_{\max}/(1-\gamma))\sqrt{H}\, D_{K,T} e^{C_3 KT}}{\varepsilon}\right\}. \tag{C.30}$$

By Lemma C.7, if

$$d_{K,R_{\varepsilon,H}}(f,g) \le \frac{\varepsilon}{2M_{K,H,T}},$$

then $\|G_H(f) - G_H(g)\|_{L^2(P)} \le \varepsilon$. Therefore, Lemma C.5 implies

$$\log \mathcal{N}(\varepsilon, \mathcal{G}_H, L^2(P)) \le p_{s,1} \log\left(1 + \frac{C\{1 + T + \sqrt{m} + R_{\varepsilon,H}\}(p_{s,1}p_{s,4})^{p_{s,2}+2} M_{K,H,T}}{\varepsilon}\right). \tag{C.31}$$

We now take $H = H_n$. If $\gamma = 0$, then $G_f = G_{H_n}(f)$. If $0 < \gamma < 1$, then, by $H_n = \max\left\{1, \left\lceil \frac{\log(8\sqrt{n})}{\log(1/\gamma)} \right\rceil\right\}$,

$$\gamma^{H_n} \le \frac{1}{8\sqrt{n}}.$$

Thus, in all cases,

$$\sup_{f\in\mathcal{F}_{K,s}} |G_f - G_{H_n}(f)| \le \frac{r_{\max}}{1-\gamma}\frac{1}{8\sqrt{n}}. \tag{C.32}$$

For $(r_{\max}/(1-\gamma))/n \le \varepsilon \le r_{\max}/(1-\gamma)$, (C.31), (C.30), and (C.12) imply

$$\log \mathcal{N}(\varepsilon, \mathcal{G}_{H_n}, L^2(P)) \le p_{s,1}\Gamma_n(K,s,m,d,T)\log\left(\frac{e r_{\max}}{(1-\gamma)\varepsilon}\right). \tag{C.33}$$

The factor n^2 in (C.12) dominates the powers of ε^{-1} generated by R_{ε,H_n} and by the covering radius.

We use the following bounded entropy maximal inequality. If $\mathcal{G}$ is separable, $|g| \le r_0$ for all $g \in \mathcal{G}$, and

$$\log \mathcal{N}(\varepsilon, \mathcal{G}, L^2(P)) \le E\log(e\,c/\varepsilon), \qquad r_0/n \le \varepsilon \le r_0,$$

then, with probability at least $1-\delta$,

$$\sup_{g\in\mathcal{G}} |(P_n - P)g| \le C r_0 \left\{ \sqrt{\frac{E + \log(2/\delta)}{n}} + \frac{E + \log(2/\delta)}{n} \right\}. \tag{C.34}$$

This follows from symmetrization and Dudley's entropy integral bound (Wainwright, 2019, Theorem 5.22), together with the standard bounded-difference concentration inequality for suprema of bounded empirical processes.

Applying (C.34) to $\mathcal{G}_{H_n}$ with $r_0 = r_{\max}/(1-\gamma)$ and $E = p_{s,1}\Gamma_n(K,s,m,d,T)$, we obtain, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}_{K,s}} |(P_n - P)G_{H_n}(f)| \le C\frac{r_{\max}}{1-\gamma}\left\{ \sqrt{\frac{p_{s,1}\Gamma_n(K,s,m,d,T) + \log(2/\delta)}{n}} + \frac{p_{s,1}\Gamma_n(K,s,m,d,T) + \log(2/\delta)}{n} \right\}. \tag{C.35}$$

By (C.32),

$$\sup_{f\in\mathcal{F}_{K,s}} |(P_n - P)(G_f - G_{H_n}(f))| \le \frac{r_{\max}}{1-\gamma}\frac{1}{4\sqrt{n}},$$

which is absorbed into the right side of (C.35). Hence the same bound holds with G_f in place of $G_{H_n}(f)$, after increasing the numerical constant.

Finally, combining (C.18), (C.22), and the uniform-convergence bound from Step 5 gives

$$\begin{aligned} V^\star(\rho) - V(\widehat{f}_{n,K,s}) \le C_1 \Bigg[& e^{-2KT} m_2(\mu_0) + \frac{d\Lambda}{K} + C_0^m m d L_{g^\star}^2 s^{-2/m} \\ & + \frac{r_{\max}}{1-\gamma}\sqrt{\frac{p_{s,1}\Gamma_n(K,s,m,d,T) + \log(2/\delta)}{n}} \\ & + \frac{r_{\max}}{1-\gamma}\frac{p_{s,1}\Gamma_n(K,s,m,d,T) + \log(2/\delta)}{n} \Bigg], \end{aligned}$$

which is (C.13). This proves Theorem C.2.

D Proof for Section 5.1

This appendix proves the policy-gradient formula used in Section 5. The argument first represents the diffusion policy through a reference driftless diffusion and a Girsanov change of measure, then differentiates expectations with respect to the drift parameter, and finally inserts this derivative into the standard performance-difference identity for discounted MDPs.

D.1 Assumptions and derivative identity

We state the assumptions used in Proposition 5.1 and the derivative identity underlying its proof.

Assumption D.1. *Assume that f is continuously differentiable in θ and*

$$\sup_{s,\theta,x} \|f(s,0,\theta,x)\| + \|\nabla_\theta f(s,0,\theta,x)\| < \infty.$$

Moreover, f and $\nabla_\theta f$ are α-Hölder continuous in a; that is, there exists a constant $C > 0$ and $\alpha \in (0,1]$ such that for all $s \in [0,T], x \in \mathcal{X}$, $\theta \in \mathbb{R}^k$ and $a, a' \in \mathbb{R}^d$,

$$\|f(s,a,\theta,x) - f(s,a',\theta,x)\| + \|\nabla_\theta f(s,a,\theta,x) - \nabla_\theta f(s,a',\theta,x)\| \le C\left(\|a-a'\|^\alpha \vee \|a-a'\|\right).$$

Furthermore, we impose an additional condition on the initial distribution μ_0:

$$\mathbb{E}_{\mu_0}\left[\exp\left(ca^2\right)\right] < \infty, \qquad \text{where } a \sim \mu_0, \tag{D.1}$$

for some sufficiently small constant $c > 0$.

By Assumption D.1, and holding $a' = 0$ fixed, there exists constant $C > 0$ such that

$$\|f(s,a,\theta,x)\| + \|\nabla_\theta f(s,a,\theta,x)\| \le C(\|a\| + 1).$$

This growth bound is used to verify the martingale and differentiability conditions required by the Girsanov-based argument.

To achieve our primary objective of identifying the gradient of the value function $\nabla_\theta V^{\pi_\theta}(\rho)$, we begin with characterizing the derivative of $\mathbb{E}^{\mathbb{P}_\theta}\left[\varphi(\bar{a}_T)\right]$ with a bounded measurable function φ.

Lemma D.1. *Suppose Assumptions D.1, 2.1 and* (4.2) *hold. Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a bounded measurable function, and define $U_\theta^\varphi(x) = \mathbb{E}^{\mathbb{P}_\theta}\left[\varphi(\bar{a}_T^x)\right]$. Then,*

$$\nabla_\theta U_\theta^\varphi(x) = \mathbb{E}^{\mathbb{P}_\theta}\left[\left(\int_0^T \nabla_\theta f_\theta(t,\bar{a}_t,x)\left(\sigma^\top(t,\bar{a}_t,x)\right)^{-1} \mathrm{d}B_t\right)\varphi(\bar{a}_T)\right].$$

Applying Lemma D.1 with the bounded test function $\varphi(a) = Q^{\pi_\theta}(x,a)$ inside the discounted performance-difference identity yields Proposition 5.1. Here the normalized discounted state distribution is

$$d_\rho^\pi(\mathrm{d}x) := \int_{\mathcal{X}} d^\pi(\mathrm{d}x|x')\rho(\mathrm{d}x'), \qquad d^\pi(\mathrm{d}x|x') := (1-\gamma)\sum_{n=0}^\infty \gamma^n P_\pi^n(\mathrm{d}x|x'),$$

where P_π^n is the n-fold transition kernel under π, with $P_\pi^0(\mathrm{d}x|x') = \delta_{x'}(\mathrm{d}x)$.

This identity is in the same spirit as the classical policy-gradient theorem in RL (Sutton et al., 1999) and related policy-gradient formulas for discrete-time diffusion policies with Gaussian transition kernels (Psenka et al., 2024). In the discrete-time setting, the internal denoising trajectory has an explicit product density, so the policy score decomposes into a sum of transition-kernel scores. In contrast, our policy is defined as the terminal law of a continuous-time SDE with a general drift and state-dependent volatility. The terminal action density is generally unavailable, and the corresponding pathwise likelihood ratio must be represented through a Girsanov change of measure. This replaces the discrete sum of internal scores by the stochastic integral $\int_0^T \nabla_\theta f_\theta(t,\bar{a}_t^x,x)\left(\sigma^\top(t,\bar{a}_t^x,x)\right)^{-1} dB_t$, which accounts for the additional technical work in Appendix D. At the same time, the continuous-time expression gives a discretization-independent training identity, applies directly to general SDE-based diffusion policies, and separates the role of the drift sensitivity $\nabla_\theta f_\theta$ from the sampling discretization used in implementation. This gives a compact training formula for SDE-based diffusion policies and supports the implementation of the diffusion-policy classes studied above.

D.2 Constructing policy from a reference probability space using Girsanov's theorem

Consider a reference probability space $(\Omega', \mathbb{F}' = (\mathcal{F}_t')_{t\in[0,T]}, \mathbb{Q})$ where $\{B_s^{\mathbb{Q}}\}_{s\in[0,T]}$ is a d-dimensional $\mathbb{Q}$-Brownian motion and

$$\mathrm{d}\bar{a}_s = \sigma(s,\bar{a}_s,x)\cdot \mathrm{d}B_s^{\mathbb{Q}}, \tag{D.2}$$

where the existence of the strong solution is guaranteed by the Lipschitz property of σ. For fixed $\theta \in \mathbb{R}^k$ and $x \in \mathcal{X}$, we defined the process $\{M_t(\theta, x)\}_{t\in[0,T]}$ by

$$M_t(\theta,x) := \exp\Big(\int_0^t f(s,\bar{a}_s,\theta,x)^\top \big(\sigma^\top(s,\bar{a}_s,x)\big)^{-1} \mathrm{d}B_s^{\mathbb{Q}} - \frac{1}{2}\int_0^t f(s,\bar{a}_s,\theta,x)^\top \big((\sigma\sigma^\top)(s,\bar{a}_s,x)\big)^{-1} f(s,\bar{a}_s,\theta,x)\mathrm{d}s\Big). \tag{D.3}$$

The following lemma verifies that the conditions of Karatzas & Shreve (1991, Corollary 5.14 in Chapter 3) hold with an even weaker assumption than Assumption D.1.

Lemma D.2. *Assume that for all $x \in \mathcal{X}$, $\theta \in \mathbb{R}^k$, there exists a constant C such that*

$$\big\|f(s,a,\theta,x)\big\| \le C(|a|+1).$$

Then for any $x \in \mathcal{X}$, $\theta \in \mathbb{R}^k$, there exists a sequence of real numbers $0 = t_0 < t_1 < ... < t_n = T$ such that

$$\mathbb{E}^{\mathbb{Q}}\Big[\exp\Big(\int_{t_i}^{t_{i+1}} f(s,\bar{a}_s,\theta,x)^\top \big((\sigma\sigma^\top)(s,\bar{a}_s,x)\big)^{-1} f(s,\bar{a}_s,\theta,x)\mathrm{d}s\Big)\Big] < \infty,$$

for $i = 0, 1, .., n-1$, $x \in \mathcal{X}$, $\theta \in \mathbb{R}^d$. Consequently, $\{M_t(\theta,x)\}_{t\in[0,T]}$ is a $\mathbb{Q}$-martingale.

Proof. By the linear growth condition of f, we get

$$\begin{aligned}\int_{t_i}^{t_{i+1}} f(s,\bar{a}_s,\theta,x)^\top \big((\sigma\sigma^\top)(s,\bar{a}_s,x)\big)^{-1} f(s,\bar{a}_s,\theta,x)\mathrm{d}s &\le C\int_{t_i}^{t_{i+1}} \big|f(s,\bar{a}_s,\theta,x)\big|^2\mathrm{d}s \\ &\le C\int_{t_i}^{t_{i+1}} \big(|\bar{a}_s|^2+1\big)\mathrm{d}s \le C(t_{i+1}-t_i)\Big(\sup_{s\in[t_i,t_{i+1}]}|\bar{a}_s|^2+1\Big),\end{aligned}$$

where the second inequality holds since $(|\bar{a}_s|+1)^2 \le 2(|\bar{a}_s|^2+1)$ and constant C may vary line by line. Then, we have

$$\begin{aligned}&\exp\Big(\int_{t_i}^{t_{i+1}} f(s,\bar{a}_s,\theta,x)^\top \big((\sigma\sigma^\top)(s,\bar{a}_s,x)\big)^{-1} f(s,\bar{a}_s,\theta,x)\mathrm{d}s\Big) \\ &\le \exp\Big(C(t_{i+1}-t_i)\Big(\sup_{s\in[t_i,t_{i+1}]}|\bar{a}_s|^2+1\Big)\Big) \\ &= \sup_{s\in[t_i,t_{i+1}]}\exp\Big(C(t_{i+1}-t_i)\big(|\bar{a}_s|^2+1\big)\Big).\end{aligned}$$

Note that $\mathrm{d}\bar{a}_t = \sigma(t,\bar{a}_t,x)\cdot \mathrm{d}B_t^{\mathbb{Q}}$ and $\sigma(s,a,x) \prec CI_d$, using Hajek (1985, Theorem 1.3), (D.1) in Assumption D.1 and the fact that $\phi(a) = \exp\big(C(t_{i+1}-t_i)(|a|^2+1)\big)$ is a non-decreasing convex function, we know

$$\mathbb{E}^{\mathbb{Q}}\Big[\exp\Big(C(t_{i+1}-t_i)\big(|\bar{a}_s|^2+1\big)\Big)\Big] \le \mathbb{E}\big[\exp\big(C(t_{i+1}-t_i)(|Z|^2+1)\big)\big],$$

where $Z \sim N(0, CsI_d)$. Set $t_{i+1}-t_i$ small enough (for instance $t_{i+1}-t_i = \frac{1}{4C^2T}$), we have

$$\mathbb{E}\big[\exp\big(C(t_{i+1}-t_i)(|Z|^2+1)\big)\big] < \infty.$$

Define $Y_s := \exp\big(C(t_{i+1}-t_i)(|\bar{a}_s|^2+1)\big)$, then, again by $\mathrm{d}\bar{a}_t = \sigma(t,\bar{a}_t,x)\cdot \mathrm{d}B_t^{\mathbb{Q}}$, and using the fact that $Y_s = \phi(a_s)$ with $\phi(a) = \exp\big(C(t_{i+1}-t_i)(|a|^2+1)\big)$ is a nondecreasing convex function, we get that $\{\sqrt{Y_s}\}_{s\in[0,T]}$ is a submartingale (Karatzas & Shreve, 1991, Problem 3.7 in Chapter 1), then using Doob's martingale inequality (Karatzas & Shreve, 1991, Theorem 3.8 in Chapter 1), we achieve

$$\mathbb{E}\Big[\sup_{s\in[t_i,t_{i+1}]}\exp\Big(C(t_{i+1}-t_i)\big(|\bar{a}_s|^2+1\big)\Big)\Big] \le \mathbb{E}\Big[\sup_{s\in[0,T]} Y_s\Big] \le 4\mathbb{E}\big[Y_T\big] < \infty,$$

provided that $t_{i+1}-t_i$ small enough, and the final statements holds due to Karatzas & Shreve (1991, Corollary 5.14 in Chapter 3). □

It is not hard to see that our Assumption D.1 is stronger than the assumptions of Lemma D.2. Therefore, the Novikov's condition holds, and we can apply Girsanov theorem to define a family of new probability measures $\{\mathbb{P}^{\theta,x}\}_{\theta\in\mathbb{R}^k,x\in\mathcal{X}}$ by

$$\frac{\mathrm{d}\mathbb{P}^{\theta,x}}{\mathrm{d}\mathbb{Q}} = M_t(\theta,x), \tag{D.4}$$

and get that

$$B_t^{\theta,x} := B_t^{\mathbb{Q}} - \int_0^t \sigma(s,\bar{a}_s,x)^{-1} f(s,\bar{a}_s,\theta,x)\mathrm{d}s, \text{ is a Brownian motion under } \mathbb{P}^{\theta,x}.$$

Therefore,

$$\mathrm{d}\bar{a}_s = \sigma(s,\bar{a}_s,x)\cdot \mathrm{d}B_s^{\mathbb{Q}} = f(s,\bar{a}_s,\theta,x)\mathrm{d}s + \sigma(s,\bar{a}_s,x)\cdot \mathrm{d}B_s^{\theta,x}.$$

Finally, for any fixed θ and x, the SDE above satisfies the standard SDE with Lipschitz drift and diffusion coefficients, the weak uniqueness holds (Karatzas & Shreve, 1991, Proposition 2.13 in Chapter 5), and therefore, $\bar{a}_T$ under $\mathbb{P}^{\theta,x}$ should share same distribution of $\bar{a}_T^{\theta,x}$ in (2.3). We summarize our discussion in this subsection as the following proposition.

Proposition D.3. *Suppose Assumptions D.1, 2.1 and* (4.2) *hold, consider the SDE* (2.3) *posed on the probability space* $(\Omega,\mathbb{F}=(\mathcal{F}_t)_{t\in[0,T]},\mathbb{P})$ *supporting the* $\mathbb{P}$*-Brownian motion* $\{B_s\}_{s\in[0,T]}$ *and the SDE* (D.2) *posed on the reference probability space* $(\Omega',\mathbb{F}'=(\mathcal{F}'_t)_{t\in[0,T]},\mathbb{Q})$ *supporting the* $\mathbb{Q}$*-Brownian motion* $\{B_s^{\mathbb{Q}}\}_{s\in[0,T]}$*. Then, a family of probability measure* $\{\mathbb{P}^{\theta,x}\}_{\theta\in\mathbb{R}^k,x\in\mathcal{X}}$ *are well-defined by* (D.4)*, and*

$$\mathbb{P}^{\theta,x}\circ(\bar{a}_T)^{-1} = \mathbb{P}\circ(\bar{a}_T^{\theta,x})^{-1}.$$

.

D.3 Derivative for $\mathbb{E}^{\mathbb{P}^{\theta,x}}\big[\varphi(\bar{a}_T)\big]$ when φ is a continuous bounded function independent of θ.

Let φ be a continuous bounded function independent of θ and by (D.4), we have

$$\mathbb{E}^{\mathbb{P}^{\theta,x}}\big[\varphi(\bar{a}_T)\big] = \mathbb{E}^{\mathbb{Q}}\big[M_T(\theta,x)\varphi(\bar{a}_T)\big],$$

where $M_t(\theta,x)$ is defined in (D.3) and further,

$$\mathrm{d}M_t(\theta,x) = M_t(\theta,x)f(t,\bar{a}_t,\theta,x)^\top\sigma(t,\bar{a}_t,x)^{-1}\mathrm{d}B_t^{\mathbb{Q}},$$
$$\mathrm{d}\bar{a}_t = \sigma(t,\bar{a}_t,x)\cdot\mathrm{d}B_t^{\mathbb{Q}}, \text{ for all } t\in[0,T].$$

We first present two technical lemmas which are crucial for the follow-up analysis.

Lemma D.4. *Suppose that for every fixed* $M>0$ *and* $\beta\in\mathbb{R}^k$*,*

$$\lim_{\varepsilon\to 0}\sup_{\theta\in\mathbb{R}^d,x\in\mathcal{X}}\mathbb{E}^{\mathbb{Q}}\left[\sup_{0\le s\le T,0\le t\le\varepsilon}\|\nabla_\theta f(s,\bar{a}_s,\theta+t\beta,x)-\nabla_\theta f(s,\bar{a}_s,\theta,x)\|_{\mathrm{op}}^M\right]=0 \tag{D.5}$$

Then for any $p\ge 1$*,*

$$\lim_{\epsilon\to 0}\sup_{\theta\in\mathbb{R}^d,x\in\mathcal{X}}\mathbb{E}^{\mathbb{Q}}\left[\sup_{0\le t\le T}\left|\frac{1}{\varepsilon}\left(\frac{M_T(\theta+\epsilon\beta,x)}{M_T(\theta,x)}-1\right)-Y_T(\theta,\beta,x)\right|^p\right]=0. \tag{D.6}$$

Proof. We define the following quantities:

$$\begin{aligned} r_1(\varepsilon,s,\bar{a}_s,\beta;\theta,x) &= f(s,\bar{a}_s,\theta+\varepsilon\beta,x)-f(s,\bar{a}_s,\theta,x)\\ &\quad -\varepsilon\langle\nabla_\theta f(s,\bar{a}_s,\theta,x),\beta\rangle; \end{aligned} \tag{D.7}$$

$$\begin{aligned} r_2(\varepsilon,s,\bar{a}_s,\beta;\theta,x) &= \varepsilon^2\langle\nabla_\theta f,\beta\rangle^\top\big((\sigma\sigma^\top)(s,\bar{a}_s,x)\big)^{-1}\langle\nabla_\theta f,\beta\rangle\\ &\quad +2\varepsilon\langle\nabla_\theta f,\beta\rangle^\top\big((\sigma\sigma^\top)(s,\bar{a}_s,x)\big)^{-1}r_1(\varepsilon)\\ &\quad +r_1(\varepsilon)^\top\big((\sigma\sigma^\top)(s,\bar{a}_s,x)\big)^{-1}r_1(\varepsilon)\\ &\quad +2r_1(\varepsilon)^\top\big((\sigma\sigma^\top)(s,\bar{a}_s,x)\big)^{-1}f(s,\bar{a}_s,\theta,x), \end{aligned} \tag{D.8}$$

where f is shorthand for $f(s, \bar{a}_s, \theta, x)$ and r_1 is shorthand for the quantity defined in the preceding equation. The reason for defining r_1 is clear: it is the error of the linear approximation of f at θ. Similarly, r_2 is the remainder term for approximating the quadratic term $f(\theta+\varepsilon\beta)^\top((\sigma\sigma^\top)(s,\bar{a}_s,x))^{-1}f(\theta+\varepsilon\beta)$ by $f(\theta)^\top\left((\sigma\sigma^\top)(s,\bar{a}_s,x)\right)^{-1}f(\theta)$. We have

$$\begin{aligned}&\frac{M_t(\theta+\varepsilon\beta,x)}{M_t(\theta,x)}\\&=\exp\left(\begin{array}{l}\varepsilon Y_t(\theta,\beta,x)+\displaystyle\int_0^t r_1(\varepsilon,s,\bar{a}_s,\beta;\theta,x)\mathrm{d}B_s^Q\\-\dfrac{1}{2}\displaystyle\int_0^t r_2(\varepsilon,s,\bar{a}_s,\beta;\theta,x)\mathrm{d}s\end{array}\right).\end{aligned}$$

Now we use the following inequality: for every $z\in\mathbb{R}$:

$$|e^z-1-z|\le z^2e^{|z|},$$

which gives us that

$$\begin{aligned}&\left|\frac{M_t(\theta+\varepsilon\beta,x)}{M_t(\theta,x)}-1-\left(\begin{array}{l}\varepsilon Y_t(\theta,\beta,x)+\displaystyle\int_0^t r_1(\varepsilon,s,\bar{a}_s,\beta;\theta,x)dB_s^Q\\-\dfrac{1}{2}\displaystyle\int_0^t r_2(\varepsilon,s,\bar{a}_s,\beta;\theta,x)\mathrm{d}s\end{array}\right)\right|\\\le&\left(\begin{array}{l}\varepsilon Y_t(\theta,\beta,x)+\displaystyle\int_0^t r_1(\varepsilon,s,\bar{a}_s,\beta;\theta,x)\mathrm{d}B_s^Q\\-\dfrac{1}{2}\displaystyle\int_0^t r_2(\varepsilon,s,\bar{a}_s,\beta;\theta,x)\mathrm{d}s\end{array}\right)^2.\\\le&\exp\left(\left|\begin{array}{l}\varepsilon Y_t(\theta,\beta,x)+\displaystyle\int_0^t r_1(\varepsilon,s,\bar{a}_s,\beta;\theta,x)\mathrm{d}B_s^Q\\-\dfrac{1}{2}\displaystyle\int_0^t r_2(\varepsilon,s,\bar{a}_s,\beta;\theta,x)\mathrm{d}s\end{array}\right|\right).\end{aligned}$$

We bound the $M=M(p)$-th moment of the rightmost term, taken supremum over $0\le t\le T$, when ε is sufficiently small (M is constant with respect to ε). Firstly, by using AM-GM, we will bound the $3M$-th moment of each exponential term. To this end,

$$\begin{aligned}&\mathbb{E}\left[\sup_{0\le t\le T}\exp\left(3M\varepsilon|Y_t(\theta,\beta,x)|\right)\right]\\\le&\mathbb{E}\left[\sup_{0\le t\le T}\exp\left(3M\varepsilon\left|\int_0^t\langle\nabla_\theta f,\beta\rangle^\top(\sigma^\top(s,\bar{a}_s,x))^{-1}\mathrm{d}B_s^Q\right|\right)\right]\\+&\mathbb{E}\left[\sup_{0\le t\le T}\exp\left(3M\varepsilon\left|\int_0^t\langle\nabla_\theta f,\beta\rangle^\top\left((\sigma\sigma^\top)(s,\bar{a}_s,x)\right)^{-1}f\mathrm{d}s\right|\right)\right].\end{aligned}$$

For the first term, we further use the following crude bound:

$$\begin{aligned}&\mathbb{E}\left[\sup_{0\le t\le T}\exp\left(3M\varepsilon\left|\int_0^t\langle\nabla_\theta f,\beta\rangle^\top\left(\sigma^\top(s,\bar{a}_s,x)\right)^{-1}\mathrm{d}B_s^Q\right|\right)\right]\\\le&\,\mathbb{E}\left[\sup_{0\le t\le T}\exp\left(3M\varepsilon\int_0^t\langle\nabla_\theta f,\beta\rangle^\top\left(\sigma^\top(s,\bar{a}_s,x)\right)^{-1}\mathrm{d}B_s^Q\right)\right]\\&+\mathbb{E}\left[\sup_{0\le t\le T}\exp\left(-3M\varepsilon\int_0^t\langle\nabla_\theta f,\beta\rangle^\top\left(\sigma^\top(s,\bar{a}_s,x)\right)^{-1}\mathrm{d}B_s^Q\right)\right].\end{aligned}$$

To bound the right-hand side, we let

$$Y_t^{(1)}(\theta,\beta,x)=\int_0^t\langle\nabla_\theta f,\beta\rangle^\top\left(\sigma^\top(s,\bar{a}_s,x)\right)^{-1}\mathrm{d}B_s^Q,$$

so that $\{Y_t^{(1)}\}_{t\geq 0}$ is a martingale. First, by L^2 martingale inequality we get that

$$\mathbb{E}\left[\sup_{0\leq t\leq T}\exp\left(3M\varepsilon\cdot Y_t^{(1)}(\theta,\beta,x)\right)\right]\leq 4\mathbb{E}\left[\exp\left(3M\varepsilon\cdot Y_T^{(1)}(\theta,\beta,x)\right)\right].$$

By Burkholder-Davis-Gundy inequality, for every $p\geq 1$ we have

$$\mathbb{E}\left[|Y_T^{(1)}(\theta,\beta,x)|^p\right]\leq (C_0p)^p\mathbb{E}[\langle Y^{(1)}(\theta,\beta,x)\rangle_T^{p/2}],$$

for an absolute constant C_0. Using linear growth on $\nabla_\theta f$, we get that

$$\langle Y^{(1)}(\theta,\beta,x)\rangle_T\leq C_2\|\beta\|^2T\sup_{0\leq s\leq T}\{C(\|\bar{a}_s\|^2+1)\}.$$

Taking expectations on both sides, using L^p martingale inequality on $\bar{a}_s$ and operator norm boundedness of σ, we get that for $p\geq 2$,

$$\mathbb{E}\left[|Y_T^{(1)}(\theta,\beta,x)|^p\right]\leq (C_0p)^p\cdot C_1^p\|\beta\|^pd^pT^p.$$

Hence, by taking ε such that $M\varepsilon$ is small enough, and noticing that our proof does not change when we have the $-$ sign, we get from Taylor series that

$$\sup_{\theta,x}\mathbb{E}\left[\sup_{0\leq t\leq T}\exp\left(3M\varepsilon\left|\int_0^t\langle\nabla_\theta f,\beta\rangle^\top\left(\sigma^\top(s,\bar{a}_s,x)\right)^{-1}\mathrm{d}B_s^Q\right|\right)\right]<\infty.$$

Next, we deal with the same term with r_1 in place of $Y_t^{(1)}(\theta,\beta,x)$. Define

$$Y_t^{(2)}(\varepsilon,\theta,\beta,x)=\int_0^t r_1(\varepsilon,s,\bar{a}_s,\beta;\theta,x)\mathrm{d}B_s^Q,$$

and $(Y_t^{(2)}(\theta,\beta,x))$ is a martingale. Proceeding similarly as before, we use the bound

$$\mathbb{E}\left[\sup_{0\leq t\leq T}\exp\left(3M\cdot|Y_t^{(2)}(\varepsilon,\theta,\beta,x)|\right)\right]\leq 4\mathbb{E}\left[\exp\left(3M\cdot|Y_T^{(2)}(\varepsilon,\theta,\beta,x)|\right)\right].$$

By Taylor series and monotone convergence,

$$\mathbb{E}\left[\sup_{0\leq t\leq T}\exp\left(3M\cdot|Y_t^{(2)}(\varepsilon,\theta,\beta,x)|\right)\right]\leq 4\sum_{k=0}^\infty\frac{(3M)^k\mathbb{E}[|Y_T^{(2)}(\varepsilon,\theta,\beta,x)|^k]}{k!}.$$

For $k\geq 2$, we have by Burkholder-Davis-Gundy inequality that

$$\mathbb{E}[|Y_T^{(2)}(\varepsilon,\theta,\beta,x)|^k]\leq (C_0k)^k\mathbb{E}[\langle Y^{(2)}(\varepsilon,\theta,\beta,x)\rangle_T^{k/2}],$$

where

$$\langle Y^{(2)}(\varepsilon,\theta,\beta,x)\rangle_T=\int_0^T\|r_1(\varepsilon,s,\bar{a}_s,\beta,\theta,x)\|^2\mathrm{d}s.$$

We bound $\|r_1\|$ as follows

$$\begin{aligned}\|r_1(\varepsilon,s,\bar{a}_s,\beta,\theta,x)\|=&\left\|\int_0^\varepsilon\langle\nabla_\theta f(s,\bar{a}_s,\theta+t\beta,x)-\nabla_\theta f(s,\bar{a}_s,\theta,x),\beta\rangle\mathrm{d}t\right\|\\ \leq&\int_0^\varepsilon\|\nabla_\theta f(s,\bar{a}_s,\theta+t\beta,x)-\nabla_\theta f(s,\bar{a}_s,\theta,x)\|_{\mathrm{op}}\cdot\|\beta\|\mathrm{d}t\\ \leq&\varepsilon\cdot\sup_{0\leq t\leq\varepsilon}\|\nabla_\theta f(s,\bar{a}_s,\theta+t\beta,x)-\nabla_\theta f(s,\bar{a}_s,\theta,x)\|_{\mathrm{op}}\cdot\|\beta\|\\ \leq&\varepsilon\|\beta\|\cdot C(\|\bar{a}_s\|+1).\end{aligned}$$

Raising to the k-th power, we get that

$$\langle Y^{(2)}(\varepsilon,\theta,\beta,x)\rangle_T^{k/2}\leq\varepsilon^{k/2}\|\beta\|^{k/2}C^kT^{k/2}\sup_{0\leq s\leq T}(\|\bar{a}_s\|^k+1).$$

From L^k martingale and Burkholder-Davis-Gundy inequalities, we get that

$$\mathbb{E}\left[\sup_{0\le s\le T}\|\bar{a}_s\|^k\right] \le 4\mathbb{E}\left[\|\bar{a}_T\|^k\right] \le 4\cdot\mathbb{E}\left[\langle\bar{a}\rangle_T^{k/2}\right] \le 4d^k C_2^{k/2}T^{k/2}.$$

Consequently,

$$\mathbb{E}\left[|Y_T^{(2)}(\varepsilon,\theta,\beta,x)|^k\right] \le (C_0k)^k\cdot\varepsilon^{k/2}\|\beta\|^{k/2}C^k d^k C_2^{k/2}T^k,$$

and so

$$\mathbb{E}\left[\sup_{0\le t\le T}\exp\left(3M\cdot|Y_t^{(2)}(\varepsilon,\theta,\beta,x)|\right)\right] \le 4\sum_{k=0}^{\infty}\frac{(3M)^k(C_0k)^k\cdot\varepsilon^{k/2}\|\beta\|^{k/2}C^k d^k C_2^{k/2}T^k}{k!},$$

and by making ε small enough, the series on the RHS converges. Lastly, we bound the term concerning r_2. We have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{0\le t\le T}\exp\left(3M\left|\int_0^t r_2(\varepsilon,s,\bar{a}_s,\theta,\beta,x)ds\right|\right)\right]\\
&\le \mathbb{E}\left[\sup_{0\le t\le T}\exp\left(3M\int_0^t|r_2(\varepsilon,s,\bar{a}_s,\theta,\beta,x)|ds\right)\right]\\
&= \mathbb{E}\left[\exp\left(3M\int_0^T|r_2(\varepsilon,s,\bar{a}_s,\theta,\beta,x)|ds\right)\right]\\
&\le \mathbb{E}\left[\sup_{0\le s\le T}\exp\left(3MT\cdot|r_2(\varepsilon,s,\bar{a}_s,\theta,\beta,x)|\right)\right].
\end{aligned}$$

By triangle inequality,

$$|r_2(\varepsilon,s,\bar{a}_s,\theta,\beta,x)| \le C(\varepsilon+\varepsilon^2)\|\beta\|^2(\|\bar{a}_s\|^2+1) \le C\varepsilon\|\beta\|^2(\|\bar{a}_s\|^2+1),$$

and hence

$$\begin{aligned}
&\mathbb{E}\left[\sup_{0\le t\le T}\exp\left(3M\left|\int_0^t r_2(\varepsilon,s,\bar{a}_s,\theta,\beta,x)ds\right|\right)\right]\\
&\le \mathbb{E}\left[\sup_{0\le s\le T}\exp\left(3MT\cdot C\varepsilon\|\beta\|^2(\|\bar{a}_s\|^2+1)\right)\right]\\
&\le 4\mathbb{E}\left[\exp\left(3MT\cdot C\varepsilon\|\beta\|^2(\|\bar{a}_T\|^2+1)\right)\right]\\
&= 4\exp(3MT\cdot C\varepsilon\|\beta\|^2)\mathbb{E}\left[\exp\left(3MT\cdot C\varepsilon\|\beta\|^2\cdot\|\bar{a}_T\|^2\right)\right]\\
&\le 4\exp(3MT\cdot C\varepsilon\|\beta\|^2)\mathbb{E}\left[\exp\left(3MT\cdot C\varepsilon\|\beta\|^2\cdot\|\mathcal{N}(0,C_2TI_d)\|^2\right)\right].
\end{aligned}$$

By making ε small enough as a function of (M,T,β), we get that the RHS is finite. We summarize that for each fixed M and $\varepsilon=\varepsilon(M)$ small enough,

$$\sup_{\theta,x}\mathbb{E}\left[\exp\left(M\varepsilon\left|\begin{array}{l}Y_t(\theta,\beta,x)\\ +\int_0^t r_1(\varepsilon,s,\bar{a}_s,\beta;\theta,x)\mathrm{d}B_s^Q\\ -\frac{1}{2}\int_0^t r_2(\varepsilon,s,\bar{a}_s,\beta;\theta,x)\mathrm{d}s\end{array}\right|\right)\right] < \infty.$$

Next, we show that for every fixed $M=M(p)$,

$$\lim_{\varepsilon\to0}\sup_{\theta,x}\mathbb{E}\left[\left|\frac{1}{\varepsilon}\begin{pmatrix}\varepsilon Y_t(\theta,\beta,x)\\ +\int_0^t r_1(\varepsilon,s,\bar{a}_s,\beta;\theta,x)\mathrm{d}B_s^Q\\ +\int_0^t r_2(\varepsilon,s,\bar{a}_s,\beta;\theta,x)\mathrm{d}s\end{pmatrix}^2\right|^M\right] = 0.$$

However, this is clear from the fact that upper bounds of $\|r_1\|, \|r_2\|$ are linear in ε. Now it suffices to show that

$$\lim_{\varepsilon\to 0}\sup_{\theta,x}\mathbb{E}\left[\sup_{0\le t\le T}\left|\int_0^t \frac{1}{\varepsilon}r_1(\varepsilon,s,\bar{a}_s,\theta,\beta,x)dB_s^Q+\int_0^t\frac{1}{\varepsilon}r_2(\varepsilon,s,\bar{a}_s,\beta,\theta,x)\right|^M\right]=0.$$

By L_p-martingale inequality, we get that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{0\le t\le T}\left|\int_0^t\frac{1}{\varepsilon}r_1(\varepsilon,s,\bar{a}_s,\theta,\beta,x)\mathrm{d}B_s^Q\right|^M\right]\\
&\le 4\mathbb{E}\left[\left|\int_0^T\frac{1}{\varepsilon}r_1(\varepsilon,s,\bar{a}_s,\theta,\beta,x)\mathrm{d}B_s^Q\right|^M\right]\\
&\le 4C(M)\cdot\mathbb{E}\left[\left(\int_0^T\frac{1}{\varepsilon^2}\|r_1(\varepsilon,s,\bar{a}_s,\theta,\beta,x)\|^2\mathrm{d}s\right)^{M/2}\right].
\end{aligned}$$

Using the upper bound on $\|r_1\|$ earlier, we get

$$\begin{aligned}
&\mathbb{E}\left[\sup_{0\le t\le T}\left|\int_0^t\frac{1}{\varepsilon}r_1(\varepsilon,s,\bar{a}_s,\theta,\beta,x)\mathrm{d}B_s^Q\right|^M\right]\\
&\le 4C(M)\cdot T^{M/2}\|\beta\|^M\cdot\mathbb{E}\left[\sup_{0\le s\le T,0\le t\le\varepsilon}\|\nabla_\theta f(s,\bar{a}_s,\theta+t\beta,x)-\nabla_\theta f(s,\bar{a}_s,\theta,x)\|_{\mathrm{op}}^M\right].
\end{aligned}$$

Next, for r_2, note that triangle inequality gives us

$$\frac{1}{\varepsilon}|r_2(\varepsilon,s,\bar{a}_s,\theta,\beta,x)|\le C\varepsilon\|\beta\|^2(\|\bar{a}_s\|^2+1)+\frac{1}{\varepsilon}\|r_1(\varepsilon,s,\bar{a}_s,\theta,\beta,x)\|\,C(\|\bar{a}_s\|+1).$$

We get

$$\begin{aligned}
&\mathbb{E}\left[\sup_{0\le t\le T}\left|\int_0^t\frac{1}{\varepsilon}r_2(\varepsilon,s,\bar{a}_s,\theta,\beta,x)\mathrm{d}s\right|^M\right]\\
&\le\mathbb{E}\left[\left(\int_0^T\frac{1}{\varepsilon}|r_2(\varepsilon,s,\bar{a}_s,\theta,\beta,x)|\mathrm{d}s\right)^M\right]\\
&\le C(M)\cdot\left\{\mathbb{E}\left[(CT\cdot\varepsilon\|\beta\|^2)^M\sup_{0\le s\le T}(\|\bar{a}_s\|^{2M}+1)\right]\right.\\
&\qquad\left.+\mathbb{E}\left[T^M\sup_{0\le s\le T}\left(\frac{1}{\varepsilon}\|r_1(\varepsilon,s,\bar{a}_s,\theta,\beta,x)\|\right)^M\cdot C(\|\bar{a}_s\|^M+1)\right]\right\}\\
&\le O_{T,\beta}(\varepsilon^M)+C(M)T^M\mathbb{E}\left[\sup_{0\le s\le T}\left(\frac{1}{\varepsilon}\|r_1(\varepsilon,s,\bar{a}_s,\theta,\beta,x)\|\right)^{2M}\right]^{1/2}\\
&\qquad\mathbb{E}\left[\sup_{0\le s\le T}C(\|\bar{a}_s\|^{2M}+1)\right]^{1/2}\\
&\le O_{T,\beta}(\varepsilon^M)+C(M)T^M\|\beta\|^M\mathbb{E}\left[\sup_{0\le s\le T,0\le t\le\varepsilon}\left\|\begin{matrix}\nabla_\theta f(s,\bar{a}_s,\theta+t\beta,x)\\-\nabla_\theta f(s,\bar{a}_s,\theta,x)\end{matrix}\right\|_{\mathrm{op}}^{2M}\right]^{1/2}.
\end{aligned}$$

By Condition (D.5), we are done. □

Lemma D.5. *Suppose Assumptions D.1, 2.1 and* (4.2) *hold. Then there exists $c=c(T)>0$ small enough so that with $p=1+c$,*

$$\sup_{\theta\in\mathbb{R}^k,x\in\mathcal{X}}\mathbb{E}^{\mathbb{Q}}\left[\sup_{0\le t\le T}M_t(\theta,x)^p\right]<\infty.$$

Proof. By the proof of Theorem 1 from Grigelionis & Mackevičius (2003), we get that

$$\begin{aligned}
\mathbb{E}^{\mathbb{Q}}\left[M_t(\theta,x)^p\right] &\le \mathbb{E}\left[\exp\left(\frac{p(\sqrt{p}+\sqrt{p-1})\sqrt{p-1}}{2}\langle M(\theta,x)\rangle_t\right)\right] \\
&= \mathbb{E}\left[\exp\left(\frac{p(\sqrt{p}+\sqrt{p-1})\sqrt{p-1}}{2}\cdot\int_0^t f(s,\bar{a}_s,\theta,x)^\top\sigma_s^{-2}f(s,\bar{a}_s,\theta,x)\mathrm{d}s\right)\right] \\
&\le \mathbb{E}\left[\exp\left(\frac{p(\sqrt{p}+\sqrt{p-1})\sqrt{p-1}}{2}\cdot\sup_{0\le s\le t}C(\|\bar{a}_s\|^2+1)\cdot t\right)\right] \\
&\le \mathbb{E}\left[\sup_{0\le s\le t}\exp\left(\frac{p(\sqrt{p}+\sqrt{p-1})\sqrt{p-1}}{2}\cdot Ct\cdot(\|\bar{a}_s\|^2+1)\right)\right] \\
&\overset{(a)}{\le} 4\mathbb{E}\left[\exp\left(\frac{p(\sqrt{p}+\sqrt{p-1})\sqrt{p-1}}{2}\cdot Ct\cdot(\|\bar{a}_t\|^2+1)\right)\right],
\end{aligned}$$

where in (a) we use L_2-martingale inequality (here it does not matter if the inequality is vacuous - i.e. the last expression is ∞). Next, noting that $\mathrm{d}\bar{a}_t = \sigma(t,\bar{a}_t,x)\mathrm{d}t$, either by the diffusion mean-comparison theorem or Burkholder-Davis-Gundy inequality in the proof of Lemma D.4, we can choose $p>1$ so that the last expression above is finite. Hence there exists $c=c(T)$ small enough, such that for $p=1+c$,

$$\mathbb{E}^{Q}[M_t(\theta,x)^p]\le C(T),$$

for a constant $C=C(T)$. By L_p-martingale inequality, we get that

$$\mathbb{E}^{\mathbb{Q}}\left[\sup_{0\le t\le T}M_t(\theta,x)^p\right]\le\left(\frac{p}{p-1}\right)^p C(T).$$

Since our argument does not matter on θ, x,

$$\sup_{\theta\in\mathbb{R}^k,x\in\mathcal{X}}\mathbb{E}^{\mathbb{Q}}\left[\sup_{0\le t\le T}M_t(\theta,x)^p\right]<\infty.$$

□

To compute the gradient of $\mathbb{E}^{\mathbb{P}^{\theta,x}}\left[\varphi(\bar{a}_T)\right]$ in θ, we would like to compute that, for any $\beta\in\mathbb{R}^k$,

$$\lim_{\epsilon\to0}\frac{1}{\epsilon}\Big(\mathbb{E}^{\mathbb{P}^{\theta+\epsilon\beta,x}}\left[\varphi(\bar{a}_T)\right]-\mathbb{E}^{\mathbb{P}^{\theta,x}}\left[\varphi(\bar{a}_T)\right]\Big)=\lim_{\epsilon\to0}\frac{1}{\epsilon}\mathbb{E}^{\mathbb{Q}}\left[\big(M_T(\theta+\epsilon\beta,x)-M_T(\theta,x)\big)\varphi(\bar{a}_T)\right].$$

Define $Y_t(\theta,\beta,x)$ by

$$\begin{aligned}
Y_t(\theta,\beta,x) :=& \int_0^t\big\langle\nabla_\theta f(s,\bar{a}_s,\theta,x),\beta\big\rangle^\top\big(\sigma^\top(s,\bar{a}_s,x)\big)^{-1}\mathrm{d}B_s^{\mathbb{Q}} \\
&-\int_0^t\big\langle\nabla_\theta f(s,\bar{a}_s,\theta,x),\beta\big\rangle^\top\big((\sigma\sigma^\top)(s,\bar{a}_s,x)\big)^{-1}f(s,\bar{a}_s,\theta,x)\mathrm{d}s,
\end{aligned}$$

for $t\ge0$ and $\langle\nabla_\theta f,\beta\rangle=\big(\langle\nabla_\theta f_1,\beta\rangle,..,\langle\nabla_\theta f_d,\beta\rangle\big)^\top\in\mathbb{R}^d$. Recall $\bar{a}_t=\int_0^t\sigma(s,\bar{a}_s,x)\cdot\mathrm{d}B_s^{\mathbb{Q}}$, $f,\nabla_\theta f$ have linear growth on a uniformly in s,θ,x, it is not hard to see that

$$\sup_{\theta\in\mathbb{R}^d,x\in\mathcal{X}}\mathbb{E}^{\mathbb{Q}}\Big[\sup_{0\le t\le T}\big|Y_t(\theta,\beta,x)\big|^p\Big]<\infty, \tag{D.9}$$

for any $\beta\in\mathbb{R}^k$ and any finite $p\ge1$. Set $V_t(\theta,\beta,x):=Y_t(\theta,\beta,x)M_t(\theta,x)$, and use Itô's formula, we have

$$\begin{aligned}
\mathrm{d}V_t(\theta,\beta,x) &= M_t(\theta,x)\mathrm{d}Y_t(\theta,\beta,x) \\
&\quad+Y_t(\theta,\beta,x)\mathrm{d}M_t(\theta,x) \\
&\quad+\mathrm{d}\big\langle M(\theta,x),Y(\theta,\beta,x)\big\rangle_t \\
&= \begin{pmatrix}M_t(\theta,x)\big\langle\nabla_\theta f(t,\bar{a}_t,\theta,x),\beta\big\rangle^\top\big(\sigma^\top(t,\bar{a}_t,x)\big)^{-1}\\+V_t(\theta,\beta,x)f(t,\bar{a}_t,\theta,x)^\top\big(\sigma^\top(t,\bar{a}_t,x)\big)^{-1}\end{pmatrix}\mathrm{d}B_t^{\mathbb{Q}}.
\end{aligned}$$

Next lemma proves a similar result as in Carmona (2016, Lemma 4.7).

Lemma D.6. *For any $\beta \in \mathbb{R}^k$, there exists $c = c(T) > 0$ independent of β such that for $p = 1 + c$, we have*

$$\lim_{\epsilon \to 0} \sup_{\theta \in \mathbb{R}^d, x \in \mathcal{X}} \mathbb{E}^{\mathbb{Q}} \left[\sup_{0 \le t \le T} \left| \frac{M_t(\theta + \epsilon\beta, x) - M_t(\theta, x)}{\epsilon} - V_t(\theta, \beta, x) \right|^p \right] = 0$$

Proof. We choose $c_1 = c_1(T)$ from Lemma D.5 and set $c = c_1/2$. With $p = 1 + c$, we have

$$\begin{aligned}
&\mathbb{E}^{\mathbb{Q}} \left[\sup_{0 \le t \le T} \left| \frac{M_t(\theta + \epsilon\beta, x) - M_t(\theta, x)}{\epsilon} - V_t(\theta, \beta, x) \right|^p \right] \\
\le &\mathbb{E}^{\mathbb{Q}} \left[\sup_{0 \le t \le T} \left| \frac{1}{\varepsilon} \left(\frac{M_t(\theta + \varepsilon\beta, x)}{M_t(\theta, x)} - 1 \right) - Y_t(\theta, \beta, x) \right|^p \cdot \sup_{0 \le t \le T} M_t(\theta, x)^p \right] \\
\le &\mathbb{E}^{\mathbb{Q}} \left[\sup_{0 \le t \le T} M_t(\theta, x)^{1+c_1} \right]^{\frac{p}{1+c_1}} \cdot \mathbb{E}^{\mathbb{Q}} \left[\sup_{0 \le t \le T} \left| \frac{1}{\varepsilon} \left(\frac{M_t(\theta + \varepsilon\beta, x)}{M_t(\theta, x)} - 1 \right) - Y_t(\theta, \beta, x) \right|^{\frac{1+c_1}{c}} \right]^{\frac{c}{1+c_1}}
\end{aligned}$$

where the last line uses Holder's inequality with conjugates $(1 + c_1)/p$ and $(1 + c_1)/c$. Now, we use Lemma D.4 with $(1 + c_1)/c < \infty$ to obtain that

$$\lim_{\varepsilon \to 0} \sup_{\theta \in \mathbb{R}^k, x \in \mathcal{X}} \mathbb{E}^{\mathbb{Q}} \left[\sup_{0 \le t \le T} \left| \frac{1}{\varepsilon} \left(\frac{M_t(\theta + \varepsilon\beta, x)}{M_t(\theta, x)} - 1 \right) - Y_t(\theta, \beta, x) \right|^{\frac{1+c_1}{c}} \right]^{\frac{c}{1+c_1}} = 0,$$

and Lemma D.5 to obtain that

$$\sup_{\theta \in \mathbb{R}^k, x \in \mathcal{X}} \mathbb{E}^{\mathbb{Q}} \left[\sup_{0 \le t \le T} M_t(\theta, x)^{1+c_1} \right]^{\frac{p}{1+c_1}} < \infty,$$

which concludes the statement. □

Together with the boundedness of the φ, we have

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \mathbb{E}^{\mathbb{Q}} \big[\big(M_T(\theta + \epsilon\beta, x) - M_T(\theta, x) \big) \varphi(\bar{a}_T) \big] = \mathbb{E}^{\mathbb{Q}} \big[V_T(\theta, \beta, x) \varphi(\bar{a}_T) \big],$$

Define $Y_t(\theta, \beta, x) = \frac{V_t(\theta, \beta, x)}{M_t(\theta, x)}$ and use Ito's formula, we have

$$\begin{aligned}
\mathrm{d}Y_t(\theta, \beta, x) &= \frac{\mathrm{d}V_t(\theta, \beta, x)}{M_t(\theta, x)} - \frac{V_t(\theta, \beta, x)}{(M_t(\theta, x))^2} \mathrm{d}M_t(\theta, x) - \frac{\mathrm{d}\big\langle V(\theta, \beta, x), M(\theta, x) \big\rangle_t}{M_t(\theta, x)^2} \\
&\quad + \frac{V_t(\theta, \beta, x)}{(M_t(\theta, x))^3} \mathrm{d}\big\langle M(\theta, x) \big\rangle_t(x) \\
&= \big\langle \nabla_\theta f_\theta(t, \bar{a}_t, x), \beta \big\rangle (\sigma^\top)^{-1} \mathrm{d}B_t^{\mathbb{Q}} - \big\langle \nabla_\theta f_\theta(t, \bar{a}_t, x), \beta \big\rangle (\sigma\sigma^\top)^{-1} f_\theta(t, \bar{a}_t, x) \mathrm{d}t.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\lim_{\epsilon \to 0} \frac{1}{\epsilon} \mathbb{E}^{\mathbb{Q}} \left[\big(M_T(\theta + \epsilon\beta, x) - M_T(\theta, x) \big) \varphi(\bar{a}_T) \right] \\
&= \mathbb{E}^{\mathbb{Q}} \big[M_T(\theta, x) Y_T(\theta, \beta, x) \varphi(\bar{a}_T) \big] \\
&= \mathbb{E}^{\mathbb{P}_\theta} \big[Y_T(\theta, \beta, x) \varphi(\bar{a}_T) \big] \\
&= \mathbb{E}^{\mathbb{P}_\theta} \left[\int_0^T \begin{pmatrix} \big\langle \nabla_\theta f_\theta(t, \bar{a}_t, x), \beta \big\rangle (\sigma^\top)^{-1} \mathrm{d}B_t^{\mathbb{Q}} \\ - \big\langle \nabla_\theta f_\theta(t, \bar{a}_t, x), \beta \big\rangle (\sigma\sigma^\top)^{-1} f_\theta(t, \bar{a}_t, x) \mathrm{d}t \end{pmatrix} \varphi(\bar{a}_T) \right] \\
&= \mathbb{E}^{\mathbb{P}_\theta} \left[\left(\int_0^T \big\langle \nabla_\theta f_\theta(t, \bar{a}_t, x), \beta \big\rangle (\sigma^\top)^{-1} \mathrm{d}B_t \right) \varphi(\bar{a}_T) \right].
\end{aligned}$$

To conclude, let $U^\varphi(\theta) := \mathbb{E}^{\mathbb{P}_\theta} \big[\varphi(\bar{a}_T) \big]$, using the previous calculation, we have

$$\nabla_\theta U^\varphi(\theta) = \mathbb{E}^{\mathbb{P}_\theta} \left[\left(\int_0^T \nabla_\theta f_\theta(t, \bar{a}_t, x) \big(\sigma^\top(t, \bar{a}_t, x) \big)^{-1} \mathrm{d}B_t \right) \varphi(\bar{a}_T) \right].$$

D.4 Derivative for value function $V^{\pi_\theta}(\rho)$.

Next, we proceed to the policy gradient. Recall that

$$V^{\pi_\theta}(x) = \mathbb{E}^{\pi_\theta}\Big[\sum_{h=0}^{\infty}\gamma^h r(x_h,a_h)\Big|x_0=x\Big] \text{ and } V^{\pi_\theta}(\rho) = \int_S V^{\pi_\theta}(x)\rho(\mathrm{d}x).$$

Using Kerimkulov et al. (2025, Lemma 2.3) , we have

$$\begin{aligned}&V^{\pi_{\theta+\epsilon\beta}}(\rho) - V^{\pi_\theta}(\rho)\\ &= \frac{1}{1-\gamma}\int_{\mathcal{X}}\Big(\int_{\mathcal{A}} Q^{\pi_\theta}(x,a)(\pi_{\theta+\epsilon\beta}-\pi_\theta)\big(\mathrm{d}a\big|x\big)\Big)d_\rho^{\pi_{\theta+\epsilon\beta}}(\mathrm{d}x).\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\frac{1}{\epsilon}\Big(V^{\pi_{\theta+\epsilon\beta}}(\rho) - V^{\pi_\theta}(\rho)\Big)\\ &= \underbrace{\frac{1}{1-\gamma}\int_{\mathcal{X}}\frac{1}{\epsilon}\Big(\int_{\mathcal{A}} Q^{\pi_\theta}(x,a)(\pi_{\theta+\epsilon\beta}-\pi_\theta)\big(\mathrm{d}a\big|x\big)\Big)\Big(d_\rho^{\pi_{\theta+\epsilon\beta}}(\mathrm{d}x) - d_\rho^{\pi_\theta}(\mathrm{d}x)\Big)}_{I_{1,\epsilon}}\\ &\quad + \underbrace{\frac{1}{1-\gamma}\int_{\mathcal{X}}\frac{1}{\epsilon}\Big(\int_{\mathcal{A}} Q^{\pi_\theta}(x,a)(\pi_{\theta+\epsilon\beta}-\pi_\theta)\big(\mathrm{d}a\big|x\big)\Big)d_\rho^{\pi_\theta}(\mathrm{d}x)}_{I_{2,\epsilon}}.\end{aligned}$$

By Lemma D.1 whose proof only requires boundedness of the test function, we have

$$\begin{aligned}\lim_{\epsilon\to 0}\sup_{x\in\mathcal{X}}\Bigg|&\frac{1}{\epsilon}\int_{\mathcal{A}} Q^{\pi_\theta}(x,a)(\pi_{\theta+\epsilon\beta}-\pi_\theta)\big(\mathrm{d}a\big|x\big)\\ &- \mathbb{E}^{\mathbb{P}_\theta}\Big[\Big(\int_0^T\big\langle\nabla_\theta f_\theta(t,\bar a_t,x),\beta\big\rangle\big(\sigma^\top(t,\bar a_t,x)\big)^{-1}\mathrm{d}B_t\Big)Q^{\pi_\theta}(x,\bar a_T)\Big]\Bigg| = 0.\end{aligned}$$

Then, by Dominated Convergence theorem, we have

$$I_{2,\epsilon} \to \frac{1}{1-\gamma}\int_{\mathcal{X}}\mathbb{E}^{\mathbb{P}_\theta}\Big[\Big(\int_0^T\nabla_\theta f_\theta(t,\bar a_t^x,x)\big(\sigma^\top(t,\bar a_t^x,x)\big)^{-1}\mathrm{d}B_t\Big)Q^{\pi_\theta}(x,\bar a_T^x)\Big]d_\rho^{\pi_\theta}(\mathrm{d}x),$$

as $\epsilon\to 0$, and

$$\begin{aligned}&\Bigg|I_{1,\epsilon} - \frac{1}{1-\gamma}\int_{\mathcal{X}}\mathbb{E}^{\mathbb{P}_\theta}\Big[\Big(\int_0^T\nabla_\theta f_\theta(t,\bar a_t^x,x)\big(\sigma^\top(t,\bar a_t^x,x)\big)^{-1}\mathrm{d}B_t\Big)Q^{\pi_\theta}(x,\bar a_T^x)\Big]\\ &\qquad\Big(d_\rho^{\pi_{\theta+\epsilon\beta}}(\mathrm{d}x) - d_\rho^{\pi_\theta}(\mathrm{d}x)\Big)\Bigg|\\ &\le \sup_{x\in\mathcal{X}}\Bigg|\frac{1}{\epsilon}\int_{\mathcal{A}} Q^{\pi_\theta}(x,a)(\pi_{\theta+\epsilon\beta}-\pi_\theta)\big(\mathrm{d}a\big|x\big)\\ &\qquad - \mathbb{E}^{\mathbb{P}_\theta}\Big[\Big(\int_0^T\big\langle\nabla_\theta f_\theta(t,\bar a_t,x),\beta\big\rangle\big(\sigma^\top(t,\bar a_t,x)\big)^{-1}\mathrm{d}B_t\Big)Q^{\pi_\theta}(x,\bar a_T)\Big]\Bigg|\\ &\qquad\cdot\frac{1}{1-\gamma}\int_{\mathcal{X}}\Big(d_\rho^{\pi_{\theta+\epsilon\beta}}(\mathrm{d}x) + d_\rho^{\pi_\theta}(\mathrm{d}x)\Big) \to 0, \text{ when } \epsilon\to 0.\end{aligned}$$

Moreover, by Assumption (D.1), we know that

$$\mathbb{E}^{\mathbb{P}_\theta}\Big[\Big(\int_0^T\nabla_\theta f_\theta(t,\bar a_t^x,x)\big(\sigma^\top(t,\bar a_t^x,x)\big)^{-1}\mathrm{d}B_t\Big)Q^{\pi_\theta}(x,\bar a_T^x)\Big]$$

is uniformly bounded for all $x\in\mathcal{X}$. It suffices to show that

$$\int_{\mathcal{X}}\psi(x)\big(d_\rho^{\pi_{\theta+\epsilon\beta}}(\mathrm{d}x) - d_\rho^{\pi_\theta}(\mathrm{d}x)\big) \to 0.$$

for any bounded measurable function $\psi(x)$. Recall the definition of $d_\rho^\pi(\mathrm{d}x)$, we have

$$d_\rho^\pi(\mathrm{d}x) = \int_{\mathcal{X}} d^\pi(\mathrm{d}x|x')\rho(\mathrm{d}x') = (1-\gamma)\sum_{n=0}^{\infty}\gamma^n P_\pi^n(\mathrm{d}x|x')\rho(\mathrm{d}x')$$

and

$$P_\pi^1(\mathrm{d}x|x') = \int_A P(\mathrm{d}x|x',a)\pi(\mathrm{d}a|x'), \quad P_\pi^n(\mathrm{d}x|x') = \int_{\mathcal{X}} P_\pi^{n-1}(\mathrm{d}x|x'')P_\pi^1(\mathrm{d}x''|x').$$

By Lemma D.6, there exist constant $C, \epsilon_0 > 0$ depending only on $\|\psi\|_\infty$ such that for $\epsilon < \epsilon_0$

$$\begin{aligned}
&\sup_{x'\in\mathcal{X}}\left|\int_{\mathcal{X}}\psi(x)\big(P^1_{\pi_{\theta+\epsilon\beta}}(\mathrm{d}x|x') - P^1_{\pi_\theta}(\mathrm{d}x|x')\big)\right| \\
&= \sup_{x'\in\mathcal{X}}\left|\int_A\left|\int_X \psi(x)P(\mathrm{d}x|x',a)\right|\big(\pi_{\theta+\epsilon\beta}(\mathrm{d}a|x') - \pi_\theta(\mathrm{d}a|x')\big)\right| \le C\epsilon,
\end{aligned}$$

Then, if we have shown $\sup_{x'\in\mathcal{X}}\left|\int_{\mathcal{X}}\psi(x)\big(P^{n-1}_{\pi_{\theta+\epsilon\beta}}(\mathrm{d}x|x') - P^{n-1}_{\pi_\theta}(\mathrm{d}x|x')\big)\right| \le C(n-1)\epsilon$, by induction, we get

$$\begin{aligned}
&\sup_{x'\in\mathcal{X}}\left|\int_{\mathcal{X}}\psi(x)\big(P^n_{\pi_{\theta+\epsilon\beta}}(\mathrm{d}x|x') - P^n_{\pi_\theta}(\mathrm{d}x|x')\big)\right| \\
&\le \sup_{x'\in\mathcal{X}}\left|\int_A\left|\int_X \psi(x)P^{n-1}_{\pi_{\theta+\epsilon\beta}}(\mathrm{d}x|x'')\right|\big(P^1_{\pi_{\theta+\epsilon\beta}}(\mathrm{d}x''|x') - P^1_{\pi_\theta}(\mathrm{d}x''|x')\big)\right. \\
&\quad + \sup_{x'\in\mathcal{X}}\left|\int_A\left|\int_X \psi(x)\big(P^{n-1}_{\pi_{\theta+\epsilon\beta}}(\mathrm{d}x|x'') - P^{n-1}_{\pi_\theta}(\mathrm{d}x|x'')\big)\right|P^1_{\pi_\theta}(\mathrm{d}x''|x')\big)\right. \\
&\le C(n-1)\epsilon + C\epsilon = Cn\epsilon,
\end{aligned}$$

Finally, combining the estimate

$$\sup_{x'\in\mathcal{X}}\left|\int_{\mathcal{X}}\psi(x)\big(P^n_{\pi_{\theta+\epsilon\beta}} - P^n_{\pi_\theta}\big)(dx|x')\right| \le Cn\epsilon,$$

with the definition of the discounted occupancy measure,

$$d_\rho^\pi(dx) = (1-\gamma)\sum_{n=0}^{\infty}\gamma^n P_\pi^n(dx|x')\rho(dx'),$$

we obtain

$$\begin{aligned}
&\left|\int_{\mathcal{X}}\psi(x)\big(d_\rho^{\pi_{\theta+\epsilon\beta}} - d_\rho^{\pi_\theta}\big)(dx)\right| \\
&\le (1-\gamma)\sum_{n=0}^{\infty}\gamma^n \sup_{x'\in\mathcal{X}}\left|\int_{\mathcal{X}}\psi(x)\big(P^n_{\pi_{\theta+\epsilon\beta}} - P^n_{\pi_\theta}\big)(dx|x')\right| \\
&\le C\epsilon(1-\gamma)\sum_{n=0}^{\infty} n\gamma^n = \frac{C\gamma}{1-\gamma}\epsilon \longrightarrow 0,
\end{aligned}$$

as $\epsilon \to 0$. Therefore,

$$\int_{\mathcal{X}}\psi(x)\big(d_\rho^{\pi_{\theta+\epsilon\beta}} - d_\rho^{\pi_\theta}\big)(dx) \longrightarrow 0,$$

which completes the proof.

E Additional related literature

Diffusion planning, trajectory diffusion, and diffusion imitation learning. Diffusion models have also been used as trajectory-level planners and sequence/action generators for offline decision making and robot imitation learning (Janner et al., 2022; Ajay et al., 2023; Li et al., 2023; Chi et al., 2025; Ke et al., 2024; Scheikl et al., 2024; Huang et al., 2024; Balim et al., 2025). These works provide important empirical evidence that diffusion parameterizations can capture multimodal, high-dimensional, and temporally correlated behavior distributions. Our setting differs in that we study Markov diffusion policies through their one-step state-conditioned action laws and prove explicit expressivity and statistical trade-offs.

Neural-network approximation, statistical learning, and PDE/SDE tools. Finally, the finite-sample part of our work builds on neural-network approximation and statistical learning theory for ReLU networks (Yarotsky, 2017; Petersen & Voigtlaender, 2018; Schmidt-Hieber, 2020; Bartlett et al., 2019; Ou & Bölcskei, 2024), while the continuous-time analysis uses tools from stochastic calculus, stochastic control, and parabolic PDE theory, including Girsanov transformations, Fokker–Planck smoothing, and density/lower-estimate arguments for uniformly elliptic diffusions (Stroock & Varadhan, 1979; Karatzas & Shreve, 1991; Øksendal, 2003; Krylov, 2021; Bogachev et al., 2015; Aronson, 1967; Kazamaki, 2006; Carmona, 2016; Hajek, 1985).

F Details of the generic-drift experiment

We collect the MDP specifications and optimization settings for Section 5.

The reward is the negative stage cost,

$$r(x,a) = -\Big(c_a(a - g_{\mathrm{ref}}(x))^2 + c_x\big((x^{(1)})^2 + 0.8(x^{(2)})^2\big) + c_s\sin^2(\kappa_s x^{(1)}) + c_u a^2\Big).$$

The state transition is nonlinear and action-dependent:

$$\begin{aligned} X_{h+1}^{(1)} &= \alpha_{11}X_h^{(1)} + \alpha_{12}X_h^{(2)} + \beta_1\tanh(A_h) + \delta_1\sin(\omega_1 X_h^{(2)}) + \varepsilon_h^{(1)}, \\ X_{h+1}^{(2)} &= \alpha_{22}X_h^{(2)} - \alpha_{21}\sin(X_h^{(1)}) + \beta_2\tanh(\omega_a A_h) + \delta_2\sin(\omega_2 X_h^{(1)}X_h^{(2)}) + \varepsilon_h^{(2)}. \end{aligned}$$

In the experiment, we set $R_a = 2.4, X_0 \sim \mathrm{Unif}([-2.2, 2.2]^2)$, and define

$$\begin{aligned} g_{\mathrm{ref}}(x) = R_a\tanh\Big(1.15\big[1.55\sin(2.25x^{(1)} + 0.85x^{(2)}) + 0.90\cos(1.65x^{(1)} - 2.15x^{(2)}) \\ + 0.55\sin(3.2x^{(1)}x^{(2)}) - 0.35x^{(1)} + 0.25x^{(2)}\big]\Big). \end{aligned}$$

The cost parameters are specified as $c_a = 2.0, c_x = 0.055, c_s = 0.030, c_u = 0.010$, and $\kappa_s = 1.4$. The transition parameters are $\alpha_{11} = 0.78,\ \alpha_{12} = 0.27, \alpha_{22} = 0.70, \alpha_{21} = 0.22, \beta_1 = 0.26, \beta_2 = 0.20, \delta_1 = 0.08, \delta_2 = 0.06, \omega_1 = 1.6, \omega_2 = 2$, and $\omega_a = 1.2$. The noise satisfies $(\varepsilon_h^{(1)}, \varepsilon_h^{(2)}) \sim N(0, 0.03^2 I_2)$, and the next state is clipped to $[-3, 3]^2$. We use horizon $H = 10$ and discount factor $\gamma = 0.95$.

We use Adam with actor learning rate 2.5×10^{-3}, critic learning rate 5×10^{-3}, gradient clipping at 8, weight clipping at 4, and 50 policy-gradient updates. We evaluate $K \in \{2, 4, 8, 12, 16, 24, 32, 48, 64, 96, 128, 256, 512\}$, use $n \in \{64, 256, 2048\}$ trajectories per policy-gradient update, average over 50 independent seeds, and evaluate each trained policy using 1024 held-out episodes.

G Additional experiments with dissipative drifts

In this section, we consider a simpler setting of one-step contextual bandit and parameterie the drift in the dissipative regime described in Remark 4.1. Here the policy uses a mean-reverting diffusion around a learned center, so the drift is one-sided dissipative in the action variable. We first choose $s(n) \asymp \big(n/\log n\big)^{m/(m+4)}$ and hold this center-class size fixed while varying K. Consequently, there is no U-shaped trade-off in K: increasing K reduces the diffusion error until the policy regret reaches the sample-dependent center estimation error.

Setup. We consider a simpler setting of one-step contextual bandit with $X \sim \mathrm{Unif}([0,1]^m)$, action space $\mathcal{A} = \mathbb{R}^d$, and reward $r(x,a) = -\|a - g^\star(x)\|^2$, where $g^\star$ is a bounded Lipschitz selector generated from random Fourier features. The optimal value is zero, so the value gap is exactly the mean-squared distance from the sampled action to $g^\star(X)$.

Implementation details. We use $m = 4$, $d = 2$, $T = 1$, and simulate the action diffusion with 24 time steps. The initial action has scale 1, and the volatility takes values in $[0.10, 1.00]$. We construct the oracle selector $g^\star$ using 2048 random Fourier features and choose its output scale so that the target actions have magnitude at most $R_0 = 1.5$.

For each sample size n, we parameterize the center by

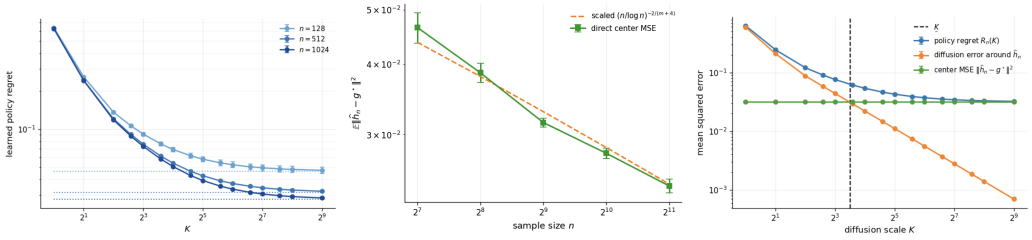
$$h_W(x) = \Phi_{s(n)}(x)^\top W, \qquad \Phi_s(x) = (1, \phi_1(x), \ldots, \phi_s(x)),$$

where $s(n) = \lfloor 16 \left(\frac{n}{\log n}\right)^{m/(m+4)} \rfloor$. Given noisy samples $Y_i = g^\star(X_i) + \xi_i$, with $\xi_i \sim N(0, 0.10^2 I_d)$, we fit W by ridge regression with regularization $\lambda = 10^{-6} \max\left\{1, \frac{s(n)+1}{n}\right\}$. Note that for each sample size n, we choose $s = s(n)$, fit the center $\widehat{h}_n$, and keep it fixed while varying K. Thus, changing K affects the diffusion approximation error but not the center estimation error.

We evaluate $K \in \{1, 2, 4, 6, 8, 12, 16, 24, 32, 48, 64, 96, 128, 192, 256, 512\}$ and $n \in \{128, 256, 512, 1024, 2048\}$, using five independent training seeds. Evaluation uses 1024 contexts and four terminal-action samples per context.

We report **(i)** the policy regret $R_n(K) := \mathbb{E}\left[\left\|\bar{a}_T^{\widehat{h}_n, X} - g^\star(X)\right\|^2\right]$, **(ii)** the diffusion error around the fitted center $\mathbb{E}\left[\left\|\bar{a}_T^{\widehat{h}_n, X} - \widehat{h}_n(X)\right\|^2\right]$, and **(iii)** the direct center MSE $\mathbb{E}\left[\left\|\widehat{h}_n(X) - g^\star(X)\right\|^2\right]$. We mark $\underline{K}$ as the smallest value of K for which the diffusion error is comparable to the center estimation error. Beyond this point, increasing K gives little further improvement in policy regret.

Results. Figure 2(a) shows that, for each fixed sample size, the learned policy gap decreases monotonically with K and approaches the corresponding direct center-MSE plateau. The plateaus decrease as n increases because a larger sample size supports a more accurate fitted center. In particular, unlike the generic neural-drift experiment, there is no U-shaped dependence on K in this dissipative experiment. Figure 2(b) shows that the direct center MSE decreases consistently with the theoretical guide $(n/\log n)^{-2/(m+4)}$. This is the approximation–estimation trade-off that determines $s(n)$, and it is independent of K. Figure 2(c) shows the behavior at $n = 512$. The diffusion error around the fitted center decreases with K, while the direct center MSE remains constant because the same center $\widehat{h}_n$ is used for every K. Consequently, the policy regret decreases toward the center-estimation floor. The marked $\underline{K}$ is the smallest diffusion scale for which the diffusion term is comparable to this floor; taking a larger K does not improve the leading statistical rate.



(a) Policy regret for different n. (b) Direct center MSE versus n. (c) Error components at $n = 512$.

Figure 2: Dissipative regime with $s = s(n)$ fixed across K. Panel (a) shows that the learned gaps decrease with K toward the sample-dependent direct center-MSE plateaus, indicated by the dotted lines. Panel (b) compares the direct center MSE with the scaled rate $(n/\log n)^{-2/(m+4)}$. Panel (c) compares the policy regret, diffusion error around the fitted center, and direct center MSE at $n = 512$; the dashed line marks $\underline{K}$, the smallest diffusion scale at which the diffusion error becomes comparable to the center MSE. (Markers show averages over five independent repetitions, and error bars indicate standard errors.)