The Polynomial Iteration Complexity for Variance Exploding Diffusion Models: Elucidating SDE and ODE Samplers

Ruofeng Yang **Bo Jiang** **Shuai Li***
John Hopcroft Center for Computer Science, Shanghai Jiao Tong University
{wanshuiyin, bjiang, shuaili8}@sjtu.edu.cn, * Corresponding Author

Abstract

Recently, variance exploding (VE) diffusion models have achieved state-of-the-art (SOTA) performance in two implementations: (1) the SDE-based implementation and (2) the probability flow ODE (PFODE) implementation. However, only a few works analyze the iteration complexity of VE-based models, and most focus on SDE-based implementation with strong assumptions. In this work, we prove the first polynomial iteration complexity under the realistic bounded support assumption for these two implementations. For the SDE-based implementation, we explain why the current SOTA VE-based model performs better than previous VE models. After that, we provide an improved result under the linear subspace data assumption and explain the great performance of VE models under the manifold data. For the PFODE-based implementation, the current results depend exponentially on problem parameters. Inspired by the previous predictor-corrector analysis framework, we propose the PFODE-Corrector algorithm and prove the polynomial complexity for the basic algorithm with uniform stepsize. After that, we show that VE-based models are more suitable for large stepsize and propose an exponential-decay stepsize version algorithm to improve the results.

1 Introduction

Recently, diffusion models have shown their power for generative modeling in many areas (Ho et al., 2020; Kim et al., 2022; Karras et al., 2022; Chen et al., 2024). The mathematical mechanism behind the diffusion model is two processes: the forward and reverse processes. The forward process $\{q_t\}_{t\in[0,T]}$ gradually injects noise into the data q_0 till it is close to pure Gaussian:

$$\mathrm{d}X_t = -f(t)X_t\ \mathrm{d}t + g(t)\ \mathrm{d}B_t, \quad X_0 \sim q_0 \in \mathbb{R}^d\,,$$

where $f(t)$ and $g(t)$ are non-negative non-decreasing sequence and $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion. One core of diffusion models is the forward process, and there are two typical forward processes (Song et al., 2020b): (1) Variance exploding (VE) SDE and (2) variance preserving (VP) SDE. When $f(t) = 1/2\beta_t$ and $g(t) = \sqrt{\beta_t}$ with a bounded β_t, the process is instantiated as VPSDE, whose stationary distribution is $\mathcal{N}(0, I)$. When the process only contains a diffusion term $g(t) = \sqrt{\mathrm{d}\sigma_t^2/\mathrm{d}t}$ and $f(t) \equiv 0$, the process is instantiated as VESDE. Two common VESDE are VESDE (SMLD) with $\sigma_t^2 = t$ and VESDE (SOTA) with $\sigma_t^2 = t^2$. These three processes have been widely used, and VESDE (SOTA) recently achieves great performance in many areas such as image generation (Teng et al., 2023; Kim et al., 2022), video generation (Blattmann et al., 2023). Furthermore, Karras et al. (2022) show that VESDE (SOTA) has a linear solution trajectory directly towards the data manifold, which is an important feature for the following application. More specifically, consistency models use VESDE (SOTA) as the forward process and obtain the first one-step generation model without adversarial training and distillation (Song et al., 2023; Kim et al., 2023). Stanczuk et al. (2024) use a VE-based diffusion model to detect the intrinsic dimension of data.

After choosing the forward process, the models reverse this process and obtain the reverse process:

$$\mathrm{d}Y_t = \left[f(T-t)Y_t + \frac{1+\eta^2}{2}g(T-t)^2\nabla \log q_{T-t}\left(Y_t\right)\right]\mathrm{d}t + \eta g(T-t)\ \mathrm{d}B_t\,, Y_0 \sim q_T\,,$$

where $(Y_t)_{t\in[0,T]} = (X_{T-t})_{t\in[0,T]}$. Since the above process has the same density function q_t with the cor-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

responding forward process (Cattiaux et al., 2021), diffusion models can generate samples by running the reverse process. Since the gradient of forward logarithmic density $\nabla \log q_t(\cdot)$ (a.k.a. score function) contains the data information, diffusion models use a neural network $s_t(\cdot)$ to approximate it (Vincent, 2011). With the approximated score, diffusion models discretize the reverse process and run the discrete process starting at a pure Gaussian noise to generate samples. The reverse process can be instantiated as reverse SDE ($\eta = 1$) or reverse probability flow ODE (PFODE, $\eta = 0$). The reverse SDE usually generates higher quality samples (Ning et al., 2023; Kim et al., 2022). The reverse PFODE always has a faster generation speed and is useful in other aspects such as calculating likelihoods (Song et al., 2020b) or one-step models (Song et al., 2023). Hence, the two reverse processes are both important, and we need to provide the theoretical grounding to elucidate the remarkable performance of VE-based models with these processes.

Though VE-based models have achieved great performance, the theoretical analysis for the iteration complexity is lacking. Most works focus on VPSDE and achieve polynomial iteration complexity in reverse SDE (Benton et al., 2023) and PFODE (Li et al., 2023; Chen et al., 2023c). For the VESDE, most works consider the reverse SDE and have strong assumptions about the data. Gao et al. (2023) assume the log-concave data and prove the polynomial complexity for a general VESDE class. With a slightly weaker log-Sobolev inequality (LSI) assumption, Lee et al. (2022) achieves the polynomial results for VESDE (SMLD). Even though the LSI assumption is slightly weaker than the log-concave distribution assumption, it still precludes the existence of multi-modal real-world data. For the reverse PFODE, Chen et al. (2023d) and Yang et al. (2024) achieve the results with exponential dependence on problem parameters. With a log-concave data distribution, Gao and Zhu (2024) achieve the polynomial complexity. Hence, the following question remains open:

Under the realistic bounded support assumption, is it possible to obtain a polynomial iteration complexity for VESDE (SOTA) with reverse SDE and PFODE?

The bounded support assumption is more realistic than previous assumptions since it admits the blow-up phenomenon of the score. More specifically, under this assumption, $\nabla \log q_{T-t}(Y_t)$ has order $1/\sigma^2_{T-t}$ and goes to $+\infty$ when t goes to T. To deal with this problem, Kim et al. (2021) use the early stopping technique, which stops the reverse process at $T - \delta$. We also adopt this technique in this work.

1.1 Our Contributions

As shown in Table 1, under the bounded support assumption, this work provides the first polynomial results for VESDE with reverse SDE and PFODE simultaneously and explains the great performance of VESDE (SOTA). During the analysis, we elucidate the design space of VE-based models, including the data structure and stepsize, and provide a basis for designing an algorithm with better performance.

The reverse SDE: the great performance of VESDE (SOTA). For the reverse SDE, we provide the first polynomial iteration complexity under the realistic EDM stepsize (4) for VESDE (SMLD) and VESDE (SOTA). During this process, we conduct a detailed analysis of various models and explain why VESDE (SOTA) and VPSDE perform well compared to VESDE (SMLD). More specifically, we need to control and balance the reverse beginning error $\mathrm{KL}(q_T \| \mathcal{N}(0, \sigma_T^2 I))$, discretization error and approximated score error. The great performance of VPSDE comes from the fast forward process convergence rate, which leads to a small reverse beginning error $\exp(-T)$. On the contrary, this error term is $1/\sigma_T^2$ for VESDE. This large T heavily influences the discretization term $d^2(T/\delta)^{\frac{1}{a}}/K$ and leads worse ϵ_{KL} dependence compared to VPSDE. However, VESDE (SOTA) has a better dependence on ϵ_{W_2}, which is determined by the designed noise schedule $g(t)$. More specifically, the requirement for the early stopping parameter δ is $W_2^2(q_\delta, q_0) = d\delta^2 \le \epsilon_{W_2}^2$, which leads to a δ with order ϵ_{W_2}. On the contrary, due to $\sigma_\delta^2 = \delta$, VPSDE needs a δ with order $\epsilon_{W_2}^2$. Hence, VESDE (SOTA) and VPSDE achieve the same ϵ dependence [1] and has different preference in empirical metric (detail in Section 4). For VESDE (SMLD), due to the slow forward convergence rate $1/T$ and variance exploding rate $\sigma_\delta^2 = \delta$, it has a worse iteration complexity.

The reverse SDE: the low-dimensional data. Since image datasets usually admit the manifold hypothesis (Pope et al., 2021) and VESDE has a great empirical performance under the low-dimensional data distribution (Song and Ermon, 2019; Karras et al., 2022; Tang and Yang, 2024), we analyze how VESDE takes advantage of manifold information after obtaining the results under the general bounded support assumption. More specifically, we assume data admit a low-dimensional d' linear subspace. Then, we prove that VESDE achieves $(d'^3 + d - d')/\epsilon_{\mathrm{KL}}^2$ results instead of $d^2/\epsilon_{\mathrm{KL}}^2$, which make the first step to explain why VESDE is more suitable for the manifold data [2].

[1] When discussing the complexity, we view ϵ_{KL} and ϵ_{W_2} to be equally important.

[2] Pope et al. (2021) shows that the d' of common image datasets are $20 \sim 40$ and $d = 3 * 256 * 256$. Hence, d'^3 is smaller than d, and the results have been directly improved.

	Distribution	Stepsize	Complexity	Reference
Reverse SDE	Log-concave	Uniform	$1/\epsilon_{W_2}^{5/2}$	Gao et al. (2023)
	Bounded support	Uniform	$1/\left(\epsilon_{\mathrm{TV}}^2\epsilon_{W_2}^8\right)$	Yang et al. (2024)
		EDM (4), $a\in[1,\infty)$	$1/\left(\epsilon_{\mathrm{KL}}^{2+1/a}\epsilon_{W_2}^{1/a}\right)$	Ours, Theorem 1
		Exponential	$1/\epsilon_{\mathrm{KL}}^2$	Ours, Corollary 1
		Exponential	$1/\epsilon_{W_2}^4$	Ours, Corollary 2
Reverse PFODE	Log-concave	Uniform	$1/\epsilon_{W_2}^4$	Gao and Zhu (2024)
	$\nabla\log q_t$ L-Lipschitz	Uniform	$e^{L/\epsilon_{\mathrm{KL}}}\mathrm{Poly}\left(1/\epsilon_{\mathrm{KL}}\right)$	Chen et al. (2023d)
	Bounded support	Uniform	$e^{1/\epsilon_{W_1}}\mathrm{Poly}\left(1/\epsilon_{W_1}\right)$	Yang et al. (2024)
		Uniform Predictor Uniform Corrector	Predictor: $1/\epsilon_{\mathrm{TV}}^2\epsilon_{W_2}^5$ Corrector: $1/\epsilon_{\mathrm{TV}}^2\epsilon_{W_2}^7$	Ours, Theorem 3
		Exponential Predictor	Predictor: $1/\epsilon_{\mathrm{TV}}^2\epsilon_{W_2}^2$ Corrector: $1/\epsilon_{\mathrm{TV}}^7\epsilon_{W_2}^6$	Ours, Theorem 4

Table 1: The results of VESDE (SOTA). The subscript of ϵ indicates the distance metric. For example, $1/(\epsilon_{\mathrm{KL}}^{2+1/a}\epsilon_{W_2}^{1/a})$ means the output is $\widetilde{O}(\epsilon_{\mathrm{KL}}^2)$ close to q_δ, which is $\epsilon_{W_2}^2$-close to q_0. The $\mathrm{KL}+W_2$ guarantee is stronger than $\mathrm{TV}+W_2$ due to Pinsker's inequality, and W_2 is stronger than the W_1 guarantee. For the reverse PFODE, this work uses a predictor-corrector type algorithm, and we provide the corresponding complexity. Finally, we note that Gao et al. (2023) and Gao and Zhu (2024) ignore an additional $1/\mathrm{Poly}(\epsilon_{W_2})$ term due to the log-concave data distribution (Remark 2).

The reverse PFODE: large stepsize is more suitable for VE-based models. For the reverse PFODE setting, inspired by the predictor-corrector framework in VPSDE (Chen et al., 2023c), we propose the PFODE-Corrector algorithm (Algorithm 1), which switches between a PFODE predictor and underdamped Langevin diffusion (ULD) corrector. As the first step, we consider the PFODE-Corrector algorithm with a uniform stepsize and achieve the first polynomial complexity for VE-based models with a PFODE predictor. However, the result of the uniform version algorithm has a bad dependence on early stopping parameter δ, which comes from the rough uniform predictor stepsize and is unfriendly to microscopic sample quality (Kim et al., 2021). Hence, we propose an exponential-decay PFODE-Corrector algorithm with exponential-decay predictor stepsize and time-dependent corrector, which improves the predictor iteration complexity and is slightly worse on the corrector iteration complexity. This existence of the exponential decay predictor depends heavily on the reverse process without the drift term (Section 5.3), which may have independent interest.

2 Related Work

Theoretical analysis for reverse SDE. For VP-based models, many works achieve polynomial iteration complexity (Lee et al., 2022; Chen et al., 2022b, 2023a; Benton et al., 2023). More specifically, Chen et al. (2022b) achieve the first polynomial result under the bounded support assumption. Chen et al. (2023a) and Benton et al. (2023) further relax the assumption to the second moment bounded assumption and achieve $1/\epsilon_{\mathrm{KL}}^2$ by using the exponential decay stepsize. Very recently, a series of works analyzes the iteration complexity and accelerating algorithm of VP-based models from a pure discrete time perspective and achieves great results (Li et al., 2023; Li and Yan, 2024b; Liang et al., 2024). However, their specially designed noise schedule relies heavily on the VP forward process and can not extend to the VE-based models.

Different from the VPSDE, there are a few works (De Bortoli et al., 2021; Lee et al., 2022; Gao et al., 2023) analyze VESDE. Specifically, De Bortoli et al. (2021) achieve the first exponential iteration complexity for VESDE (SMLD). After that, some works assume log-concave or LSI holds on data distributions to improve the results. Lee et al. (2022) propose the first polynomial iteration complexity for VESDE (SMLD) under LSI assumption. Gao et al. (2023) provide the polynomial results for a series of VESDE under log-concave data distribution. As mentioned above, these assumptions are strong and preclude the presence of highly multi-modal data distributions. For bounded support distributions, Yang et al. (2024) recently provide polynomial complexity for VE-based models. However, their forward process has a drift term, which differs from the VESDE in applications.

Theoretical analysis for reverse PFODE. When considering reverse PFODE, most of the works focus on the VPSDE setting. Chen et al. (2023d) and Yang et al. (2024) propose the convergence guarantee of PFODE with exponential dependence on the problem parameter. Li et al. (2023) and Li et al. (2024) assume an accurate enough Jacobian matrix and achieve a polynomial result for VPSDE. Gao and Zhu (2024)

assume a log-concave distribution and achieve a polynomial complexity for VP and VESDE. Without any additional assumption, Chen et al. (2023c) introduce a Langevin process as the corrector to achieve polynomial results. In this work, we adapt the predictor-corrector framework of Chen et al. (2023c) to analyze VESDE. However, unlike previous uniform stepsize, we show VESDE can choose different stepsize, depending on which problem parameter we focus on (Section 5.3).

3 Variance Exploding Diffusion Models

A diffusion model usually consists of a forward process and reverse process (Song et al., 2020b). The forward process injects Gaussian noise into data step by step, and The reverse process sequentially removes noise from data to generate samples. Recently, variance exploding diffusion models have achieved SOTA performance under this paradigm (Teng et al., 2023; Karras et al., 2022). In Section 3.1, we first introduce the typical VESDE models and its corresponding reverse process. Then, Section 3.2 introduces the underdamped Langevin diffusion, which is used as a corrector in the reverse PFODE setting.

3.1 The Variance Exploding Diffusion Model

Variance Exploding Forward Process. Let q_0 denote the data distribution, $\{\sigma_t^2\}_{t\in[0,T]}$ a non-decreasing sequence and $g(t) = \sqrt{\mathrm{d}\sigma_t^2/\mathrm{d}t}$. The variance exploding forward process is defined by:

$$\mathrm{d}X_t = g(t)\ \mathrm{d}B_t, \quad X_0 \sim q_0 \in \mathbb{R}^d\,, \tag{1}$$

where $(B_t)_{t\geq 0}$ is a d-dimensional Brownian motion. There are two common choices of variance σ_t^2: (1) VESDE (SMLD) with $\sigma_t^2 = t$ (Song et al., 2020a); (2) VESDE (SOTA) with $\sigma_t^2 = t^2$. This work analyzes these models simultaneously and shows why VESDE (SOTA) achieve a better performance (Karras et al., 2022).

Two Typical Reverse Processes When reversing the forward SDE, we obtain the reverse process $(Y_t)_{t\in[0,T]} = (X_{T-t})_{t\in[0,T]}$:

$$\mathrm{d}Y_t = \frac{1+\eta^2}{2} g(T-t)^2 \nabla \log q_{T-t}\left(Y_t\right)\mathrm{d}t + \eta g(T-t)\ \mathrm{d}B_t\,, Y_0 \sim q_T \tag{2}$$

where $\nabla \log q_t(\cdot)$ is the ground truth score function and $\eta \in [0,1]$. Recently, the reverse SDE ($\eta = 1$) and the reverse probability flow ODE (PFODE, $\eta = 0$) both achieve great empirical performance. Hence, we analyze these two typical reverse processes in this work.

Before the generation process, we need to approximate the unknown ground truth score function $\nabla \log q_t(\cdot)$ and the reverse beginning distribution q_T since these terms contain data information. The ground true score function is approximated by a score neural network $\{s_t(\cdot)\}_{t\in[0,T]}$, which is learned by the score matching technique (Song and Ermon, 2019). Since the forward process converts distribution to pure Gaussian noise, we choose $q_\infty = \mathcal{N}(0, \sigma_T^2 I)$ to approximate q_T. With these terms, the continuous-time reverse process $(\widehat{Y}_t)_{t\in[0,T]}$ is:

$$\mathrm{d}\widehat{Y}_t = \frac{1+\eta^2}{2} g(T-t)^2 s_{T-t}(\widehat{Y}_t)\mathrm{d}t + \eta g(T-t)\ \mathrm{d}B_t, \quad \widehat{Y}_0 \sim q_\infty\,, \tag{3}$$

To obtain an implementable algorithm, the diffusion model discretes the above process. Let $t_0 \leq t_1 \leq \cdots \leq t_K = T$ be the discretization points in the forward time and $h_k := t_k - t_{k-1}$ be the stepsize. As discussed in Section 1, the score function blow-ups at the end of the reverse process under the real-world image datasets (Kim et al., 2021) and our bounded-support assumption also admits this phenomenon. Kim et al. (2021) use the early stopping technique by setting $t_0 = \delta$ to avoid the blow-up phenomenon of the score, and we adopt this technique in our work. When considering the reverse process, we define $t'_k = T - t_{K-k}$ and $h'_k = h_{K-k}$. In this work, we consider two stepsize schemes: (1) the EDM stepsize (Karras et al., 2022); (2) the exponential-decay stepsize. The EDM stepsize is

$$t_k = (\delta + kh)^a \text{ and } h = (T^{1/a} - \delta)/K\,, \tag{4}$$

where $a \in [1, +\infty)$. When $a = 1$, EDM stepsize represents the uniform stepsize used by most of the theoretical works (Chen et al., 2022b; Gao and Zhu, 2024). When $a = 7$, this stepsize shows great performance (Karras et al., 2022). The exponential decay stepsize is $h_k = rt_k$, where r is a small coefficient corresponding to accuracy parameters ϵ. Recently, this stepsize has been used in theoretical works (Chen et al., 2023a; Benton et al., 2023) to improve results.

After choosing the stepsize scheme, we define the discretization process for reverse SDE and PFODE. More specifically, at the k-th interval $t \in [t'_k, t'_{k+1}]$, the diffusion model freezes the score at time t'_k and runs the following process:

$$\mathrm{d}\bar{Y}_t = \frac{1+\eta^2}{2} g(T-t)^2 s_{T-t'_k}(\bar{Y}_{t'_k})\mathrm{d}t + \eta g(T-t)\mathrm{d}B_t\,. \tag{5}$$

3.2 The Underdamped Langevin Diffusion

The role of reverse SDE and PFODE is to predict and remove noise from data. As shown in (Song et al., 2020b), diffusion models can employ a suitable

Langevin process to correct the marginal distribution provide by the predictor. We also introduce the underdamped Langevin Diffusion (ULD) corrector to achieve a polynomial result for VESDE with a PFODE predictor, which is better than the previous exponential-dependence result (Yang et al., 2024) (see Table 1).

Since the underdamped Langevin corrector works with a fixed t, we use U_t as a shorthand for the potential $-\ln q_t$. To distinguish from the predictor stage, we denote m by the time of the corrector. Let $\rho > 0$ be the friction parameter. The ULD is a stochastic process $(z_m, v_m)_{m\geq 0}$ over $\mathbb{R}^d \times \mathbb{R}^d$ given by $\mathrm{d}z_m = v_m \, \mathrm{d}m$ and

$$\mathrm{d}v_m = -\left(\nabla U_t\left(z_m\right) + \rho v_m\right)\mathrm{d}m + \sqrt{2\rho}\mathrm{d}B_m \, .$$

Let $\boldsymbol{q} := q \otimes \gamma^d$, where $\gamma^d = \mathcal{N}(0, I)$. The stationary distribution of this process is $\boldsymbol{q}_t = q_t \otimes \gamma^d$. Let h_{corr} be the corrector stepsize. We also use the EI discretization to discretize the above process, and the results $(\bar{z}_m, \bar{v}_m)_{m\geq 0}$ is given by $\mathrm{d}\bar{z}_m = \bar{v}_m \, \mathrm{d}m$ and

$$\mathrm{d}\bar{v}_m = \left(s_t(\bar{z}_{\lfloor \frac{m}{h_{\mathrm{corr}}} \rfloor h_{\mathrm{corr}}}) - \rho \bar{v}_m\right)\mathrm{d}m + \sqrt{2\rho}\mathrm{d}B_m. \quad (6)$$

Notations. For $x \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d\times d}$, we denote by $\|x\|$ and $\|A\|$ the L_2 norm for vector and matrix. For a random variable X, the sub-exponential and sub-gaussian norms are defined by

$$\|X\|_{\psi_k} := \inf\left\{t > 0 : \mathbb{E}\exp\left(|X|^k/t\right) \leq 2\right\}, k = 1, 2 \, .$$

For random vector, we denote by $\|x\|_{\psi_k} := \|\|x\|\|_{\psi_k}$.

4 Guarantee for reverse SDE

This section provides the first polynomial iteration complexity for VESDE with reverse SDE under the bounded support assumption and EDM scheme (Theorem 1). Then, we show why VESDE (SOTA) and VPSDE perform better than VESDE (SMLD). After that, we prove that VESDE can use low-dimensional information to improve the result, which explains the great performance of VESDE under the manifold data.

Before providing the results, we introduce some assumptions about the data distribution and the score function. For the data distribution, we assume it is bounded support with diameter R.

Assumption 1. *q_0 is supported on a compact set $\mathcal{M}$ and $0 \in \mathcal{M}$ and $R = \sup\{\|x - y\| : x, y \in \mathcal{M}\} \geq 1$.*

This assumption is widely used by current works (De Bortoli, 2022; Chen et al., 2022b) and is the weakest one for VE-based models.

Remark 1. *When considering the VP-based model, recent works (Chen et al., 2023a; Benton et al., 2023) assume the second moment $\mathbb{E}[\|q_0\|_2^2]$ is bounded, which is slightly weaker. We note that when considering VESDE with reverse SDE,* **Assumption 1** *can also be relaxed to this assumption. However, when considering the reverse PFODE, our exponential-decay version algorithm depends on a time-dependent Lipschitz constant L_{T-t}, which holds under the bounded support assumption.*

For the approximated score, we assume it is close enough to the ground-truth score function..

Assumption 2. *There exists a constant $\epsilon_{\mathrm{score}}$ such that for any $t \in [\delta, T]$,*

$$\mathbb{E}_{X\sim q_t}\left[\|s_t(X) - \nabla \log q_t(X)\|^2\right] \leq \epsilon_{\mathrm{score}}^2/\sigma_t^2 \, .$$

This assumption matches the order $1/\sigma_t^2$ of the ground-truth score function (Remark 1, (Chen et al., 2022a)) and is more realistic compared with the previous uniform L_2-accurate approximated score assumption (Chen et al., 2022b; Benton et al., 2023).

Theorem 1. *Assume Assumption 1 and 2 hold. Let $p_{T-\delta}$ be the output of Eq. (5). If using the EDM stepsize (Eq. (4)), then we have that*

$$\mathrm{KL}\left(p_{T-\delta}, q_\delta\right) \leq \bar{D}^2/\sigma_T^2 + d^2(T/\delta)^{\frac{1}{a}}/K + \epsilon_{\mathrm{score}}^2 \log(T/\delta) \, ,$$

for VESDE (SMLD) and VESDE (SOTA), where c is the eigenvalue of $\mathrm{Cov}\left[q_0\right]$ with the largest absolute value and $\bar{D} = d|c| + \mathbb{E}[q_0] + R$.

Furthermore, by choosing $\sigma_T^2 \geq \bar{D}^2/\epsilon_{\mathrm{KL}}^2$, $\sigma_\delta^2 = \epsilon_{W_2}^2/d$, the output is $\widetilde{O}\left(\epsilon_{\mathrm{KL}}^2 + \epsilon_{\mathrm{score}}^2\right)$-close to q_δ, which is $\epsilon_{W_2}^2$-close to q_0, with iteration complexity

$$K = \Theta\left(\frac{d^{2+\frac{1}{2a}}\bar{D}^{1/a}}{\epsilon_{\mathrm{KL}}^{2+1/a}\epsilon_{W_2}^{1/a}}\right) .$$

for VESDE (SOTA). For VESDE (SMLD), we require $K = \Theta\left(d^{2+\frac{1}{a}}\bar{D}^{\frac{2}{a}}/\left(\epsilon_{\mathrm{KL}}^{2+2/a}\epsilon_{W_2}^{2/a}\right)\right)$ to achieve the same guarantee.

We also provide the results for VPSDE under the EDM stepsize $\tilde{\Theta}\left(d^2(\sqrt{d}(R \vee \sqrt{d}))^{1/a}/(\epsilon_{\mathrm{KL}}^2\epsilon_{W_2}^{2/a})\right)$ (detail in Appendix B). We note that the accuracy parameter ϵ_{KL} and ϵ_{W_2} have the same order in VPSDE and VESDE (SOTA), and these results explain the great performance of VESDE (SOTA). The better dependence on ϵ_{KL} for VPSDE is due to the faster forward convergence rate, which is important to the Fréchet Inception Distant (FID) metric. The better dependence on ϵ_{W_2} for VESDE (SOTA) is due to $\sigma_\delta^2 = \delta^2$ instead of δ, which allows larger early stopping parameter and is important to the microscopic sample quality (the Negative Log-Likelihood (NLL) metric). The slightly worse result for VESDE (SMLD) is due to the slower forward convergence rate compared to VPSDE and a smaller early stopping parameter than VESDE (SOTA).

Theorem 1 explains the great empirical results of Karras et al. (2022) with $a = 7$ compared with the uniform stepsize ($a = 1$). Corollary 1 considers a more theory-friendly exponential stepsize and achieves an improved result. However, as shown in Figure 13 of Karras et al. (2022), the FID will increase when $a > 7$. Hence, whether the exponential decay stepsize is effective in application still needs to be explored.

Corollary 1. *Following the setting of Theorem 1 with exponential decay stepsize $h_k = rt_k$, where $r = \epsilon_{\mathrm{KL}}^2/\log(T/\delta)$, then $\mathrm{KL}\left(p_{T-\delta}, q_\delta\right)$ is bounded by*

$$\frac{(d|c| + \mathbb{E}[q_0] + R)^2}{\sigma_T^2} + \frac{d^2 \log^2(T/\delta)}{K} + \epsilon_{\mathrm{score}}^2 \log(T/\delta)\,,$$

for VESDE (SMLD) and VESDE (SOTA). By choosing $\sigma_T^2 \geq \bar{D}^2/\epsilon_{\mathrm{KL}}^2$, $\sigma_\delta^2 = \epsilon_{W_2}^2/d$, the output is $\widetilde{O}\left(\epsilon_{\mathrm{KL}}^2 + \epsilon_{\mathrm{score}}^2\right)$-close to q_δ, which is $\epsilon_{W_2}^2$-close to q_0, with iteration complexity $K = \widetilde{\Theta}\left(d^2 \log^2(T/\delta)/\epsilon_{\mathrm{KL}}^2\right)$.

To compare more comprehensively with Gao et al. (2023), we also provide a pure W_2 guarantee. The proof process is the same compared to Chen et al. (2022b), and we provide this result for completeness.

Corollary 2. *Defined by $p_{T-\delta,R_0}$ the output $p_{T-\delta}$ of Corollary 1 projected onto $\mathrm{B}(0, R_0)$ for $R_0 = \widetilde{\Theta}(R)$. Then, $W_2\left(p_{T-\delta,R_0}, q_0\right) \leq \epsilon_{W_2}$ with iteration complexity $K = \widetilde{\Theta}\left(d^2 R^4 \log^2(T/\delta)/\epsilon_{W_2}^4\right)$ and $\epsilon_{\mathrm{score}} \leq \widetilde{O}(\epsilon_{W_2})$.*

Remark 2. *Gao et al. (2023) achieve $1/\epsilon_{W_2}^{2.5}$ results under the log-concave data assumption, which is better than Corollary 2. However, this assumption is stronger than* **Assumption 1** *and far away from the multi-modal real-world data. Furthermore, the score does not blow up, which conflicts with the experimental phenomena. Hence, Gao et al. (2023) and Gao and Zhu (2024) ignore the influence of δ and an additional $1/Poly(\epsilon_{W_2})$ dependence, which is a key part of iteration complexity.*

Remark 3. *Very recently, Wang et al. (2024) also made a great step to explain the role of EDM in the training and sampling phase of diffusion models. For the sampling phase, their work provides the first polynomial iteration complexity $1/(\epsilon_{\mathrm{KL}}^{2+1/a}\delta^{1/a})$ for VESDE with reverse SDE under the EDM stepsize to guarantee $\mathrm{KL}\left(p_{T-\delta}, q_\delta\right) \leq \epsilon_{\mathrm{KL}}^2$ (Corollary 1 of their work), where the first inequality of Theorem 1 also provides similar results. However, our work further considers the order of the early stopping parameter δ to guarantee $W_2^2(q_0, q_\delta) \leq \epsilon_{W_2}^2$, which is the core to show why VESDE (SMLD) performs worse than VESDE (SOTA) and VPSDE (Discussed in Section 1 and before Corollary 1). Furthermore, we also analyze the iteration complexity of VESDE under the low-dimensional manifold setting for the first time and prove that the VE-based models can make full use of low-dimensional information to improve the iteration complexity.*

4.1 VESDE Uses the Manifold Information

Recently, many works have shown that diffusion models adapt to the intrinsic manifold data structure (Tang and Yang, 2024; Chen et al., 2023b), and this phenomenon is more obvious in VESDE. More specifically, Song and Ermon (2019) show that VESDE (SMLD) improves the score function learning process when data is supported on the manifold. Furthermore, Karras et al. (2022) unify VP and VESDE and show that the optimal solution trajectory corresponds to VESDE (SOTA), which is directly towards the data manifold. To understand how VESDE uses manifold information, we consider low-dimensional linear structured data, which has been widely used in current works (Chen et al., 2023b; Yuan et al., 2023; Guo et al., 2024).

Assumption 3. *X_0 admit a linear structure $X_0 = AZ^{\mathrm{LD}}$ where $A \in \mathbb{R}^{d\times d'}$ with orthonormal columns and $Z^{\mathrm{LD}} \sim q_z^{\mathrm{LD}}$. q_z^{LD} is supported on a compact set $\mathcal{M}'$ and $0 \in \mathcal{M}'$.*

We also define by R' the diameter of $\mathcal{M}'$ and $\bar{D}' = d|c| + \mathbb{E}\left[q_0\right] + R'$. Then, we obtain the following improved result, which uses the manifold information.

Theorem 2. *Assume* **Assumption 2** *and 3 holds. If choosing $\sigma_T^2 \geq \bar{D}'^2/\epsilon_{\mathrm{KL}}^2$, $\sigma_\delta^2 = \epsilon_{W_2}^2/d$ and using exponential decay stepsize with $r = \epsilon_{\mathrm{KL}}^2/\log(T/\delta)$, then for VESDE (SMLD) and VESDE (SOTA), the output is $\widetilde{O}\left(\epsilon_{\mathrm{KL}}^2 + \epsilon_{\mathrm{score}}^2\right)$-close to q_δ, which is $\epsilon_{W_2}^2$-close to q_0, with iteration complexity*

$$K = \widetilde{\Theta}\left(\left(d'^3 + d - d'\right)\log^2(T/\delta)/\epsilon_{\mathrm{KL}}^2\right)\,.$$

The first d'^3 term is determined by the on-support space, and the orthogonal space determines the second $d - d'$ term. As shown in Pope et al. (2021), the latent dimension d' of common image datasets is smaller than 45. For example, the latent dimensions of the CelebA and ImageNet datasets are 24 and 43, respectively. We also know that the image dimension is $d = 3 * 256 * 256$ for these two datasets, larger than d'^3. Hence, the iteration complexity of Theorem 2 is better than Corollary 1. We note that Li and Yan (2024a) show that VPSDE also adapts to unknown low-dimensional manifold under a special noise schedule and achieves $\tilde{O}(d'^4/\epsilon_{\mathrm{TV}}^2)$ result. Very recently, Azangulov et al. (2024) further improve the above result and achieve $\tilde{O}(d'^3/\epsilon_{\mathrm{KL}}^2)$ for VPSDE setting, which has the same order with Theorem 2.

5 Guarantee for PFODE Predictor

This section provides the first polynomial complexity for VESDE with a PFODE predictor. Section 5.1 introduces two versions of the PFODE-Corrector algorithm: the uniform and exponential-decay stepsize version. The uniform version is similar to Chen et al. (2023c), which is mainly used in VPSDE. Based on

it, we propose the exponential-decay version algorithm specifically for VESDE, which has better dependence on δ and is friendly to microscopic sample quality(Kim et al., 2021). Section 5.2 provides the guarantee for the above algorithms and discusses the technical novelty.

5.1 PFODE-Corrector Algorithm for VESDE

At the beginning, we introduce some useful notations. By simple algebra, we know that under Assumption 1:

$$\|\nabla^2 \log q_t(X)\| \leq (1+R^2)/\sigma_t^4 \,, \forall X \in \mathbb{R}^d \,.$$

We define by $L_t = (1+R^2)/\sigma_{T-t}^4$ the Lipshcitz constant of $\nabla \log q_{T-t}$ and $L_{\max} = (1+R^2)/\sigma_\delta^4$ the maximum Lipshcitz constant. After that, we describe the PFODE-Corrector algorithm, which switches between the predictor (Eq. 5, $\eta = 0$) and the corrector (Eq. 6). Since the corrector stage does not increase reverse time, we define by K_0 the number of predictor stages.

For the predictor stage, we proposed two stepsize (a) run PFODE predictor with uniform stepsize h_{pred} for time $T_{\text{pred}} = 1/L_{\max}$ in K_0 stages or; (b) run one exponential decay stepsize $h'_k = r t_{T-k}$ once at stage k, where $r \leq 1$. After determining the choice of predictor, we define T. For the uniform stepsize, $T = K_0 T_{\text{pred}} + \delta$. For the exponential-decay stepsize, $K_0 = 1/r \log(T/\delta)$.

For the corrector stage, the algorithm run ULMC (Eq. (6)) with uniform stepsize h_{corr} for time T_{corr}. The choice of T_{corr} corresponds to the choice of the predictor stepsize, as we highlight in Algorithm 1. As shown in Section 5.2, different predictor stepsize will introduce different dependence on accuracy parameter ϵ_{TV} and ϵ_{W_2}. We can choose the appropriate algorithm to meet different requirements.

5.2 The Theoretical Guarantee

This section provides the iteration complexity of the above algorithms. Similar to Chen et al. (2023c), we make an additional assumption on the approximated score due to the corrector stage. Since the Lipschitz constant of $\nabla \log q_t$ has order $\left(1+R^2\right)/\sigma_t^4$, we also assume the Lipschitz constant of s_t is dependent on t.

Assumption 4. *For all discretization points, s_t is L_{T-t} Lipschitz, where $L_t = \left(1+R^2\right)/\sigma_{T-t}^4$.*

After that, we prove the results for Algorithm 1. We first provide the results for the uniform version algorithm, which achieves competitive results compared to the DPUM algorithm for the VPSDE.

Theorem 3. *[Uniform Stepsize] Assume* **Assumption 1**, *2 and 4 hold. Let $p_{T-\delta}^{\text{ULMC}}$ be the output of the uniform version algorithm. Then, if $\sigma_t^2 = t^2$ (VESDE (SOTA)), $\text{TV}(p_{T-\delta}^{\text{ULMC}}, q_\delta)$ is bounded by*

$$\frac{\bar{D}}{T} + \frac{\sqrt{d}R^3 h_{\text{corr}}}{\delta^5} + \frac{R^5(R+\sqrt{d})h_{\text{pred}}}{\delta^5} + \frac{R\epsilon_{\text{score}} T}{\delta^2} \,.$$

Furthermore, for VESDE (SOTA), by ignoring the approximated score error and choosing $T \geq \bar{D}/\epsilon_{\text{TV}}, \delta = \epsilon_{W_2}/\sqrt{d}, h_{\text{corr}} \leq \delta^5 \epsilon_{\text{TV}}/(\sqrt{d}R^3), h_{\text{pred}} \leq \delta^5 \epsilon_{\text{TV}}/(R^5(R+\sqrt{d}))$, the output is ϵ_{TV}-close to q_δ, which is ϵ_{W_2}-close to q_0, with total iteration complexity

$$\max\left\{\frac{T}{h_{\text{pred}}}, \frac{T\sqrt{L_{\max}}}{h_{\text{corr}}}\right\} \leq \tilde{O}\left(\frac{\bar{D}R^5 d^{3.5}(R+\sqrt{d})}{\epsilon_{\text{TV}}^2 \epsilon_{W_2}^7}\right) .$$

For VESDE (SMLD) setting ($\sigma_t^2 = t$), the total iteration complexity is $\tilde{O}\left(\frac{\bar{D}^2 R^5 d^3 (R+\sqrt{d})}{\epsilon_{\text{TV}}^3 \epsilon_{W_2}^6}\right)$.

Theorem 3 provides the first polynomial iteration complexity for VESDE with a PFODE predictor. This result is competitive with the DPUM algorithm (an algorithm for VPSDE with the ULD corrector, Chen et al. (2023c)). More specifically, when considering **Assumption 1**, the result of DPUM is $\tilde{O}(R^4 d^{0.5}(\sqrt{d}(R \vee \sqrt{d}))^2/(\epsilon_{\text{TV}}\epsilon_{W_2}^8))$, which has the same order in ϵ with Theorem 3 (detail in Appendix B). After obtaining these results, we note that the dependence on ϵ_{W_2} is not good enough for VESDE (SOTA) under the uniform stepsize. Hence, we design an exponential-decay version algorithm to provide an δ-friendly result.

Theorem 4. *[Exponential-decay stepsize] Following the setting of Theorem 3, choosing the exponential-decay version algorithm with $r = \frac{\epsilon_{\text{TV}}^2 \sigma_\delta^2}{d^2 R^2 \log^3(T/\delta)}$, $\bar{D}/\sigma_T \leq \epsilon_{\text{TV}}$, $\sigma_\delta^2 = \epsilon_{W_2}^2/d$ and considering VESDE (SOTA), the iteration complexity for the predictor is*

$$O\Big(\frac{d^3 R^2 \log^3(T/\delta)}{\epsilon_{\text{TV}}^2 \epsilon_{W_2}^2}\Big) ,$$

and the iteration complexity for the corrector is $O\Big(\frac{d^{7.5} R^5 \bar{D}^2 \log^6(T/\delta)}{\epsilon_{\text{TV}}^7 \epsilon_{W_2}^6}\Big)$. The corrector iteration complexity dominates the total iteration complexity. We achieve the same result for VESDE (SMLD).

Remark 4. *When focusing on ϵ_{W_2}, the result of Theorem 4 is better than the uniform one. Furthermore, the predictor iteration complexity of Theorem 4 is better than the one $\tilde{O}(1/(\epsilon_{\text{TV}}^2 \epsilon_{W_2}^5))$ of Theorem 3[3]. The source of the slightly worse total iteration complexity is the corrector part, which has a large corrector time T_{corr} and small uniform stepsize h_{corr}. More specifically, $T_{\text{corr}} = 1/\sqrt{L_t}$ goes to T/R and introduce additional ϵ_{TV} dependence when $t \to T$. It is an interesting future work to obtain a better result using a more refined analysis for ULD (Foster et al., 2021).*

[3]This predictor iteration complexity is T/h_{pred} of VESDE (SOTA) in Theorem 3 and we ignore $d, R, \bar{D}$ here.

Algorithm 1 PFODE with Corrector

Input: Total time T, approximated score s, uniform predictor stepsize h_{pred}, exponential-predictor predictor ratio r, uniform corrector stepsize h_{corr}.
Initialization: Draw $\bar{Y}_0$ from $\mathcal{N}(0, \sigma_T^2 I_D)$.
if The uniform PFODE-Corrector Algorithm **then**
  for $k = 0, 1, \ldots, K_0 - 1$ **do**
    Predictor. Starting from $\bar{Y}_{k/L_{\max}}$, run the PFODE (5, $\eta = 0$) from time $\frac{k}{L_{\max}}$ to $\frac{k+1}{L_{\max}}$ with step size h_{pred} to obtain $\bar{Y}'_{(k+1)/L_{\max}}$.
    Corrector. Starting from $\bar{Y}'_{(k+1)/L_{\max}}$, run ULMC (6) for total time $1/\sqrt{L_{\max}}$ with step size h_{corr} and score $s_{(k+1)/L_{\max}}$ to obtain $\bar{Y}_{(k+1)/L_{\max}}$.
  end for
end if
if The exponential-decay PFODE-Corrector Algorithm **then**
  Set $h_k = t_k - t_{k-1} = r t_k$ and $h'_k = h_{K-k}$ for $\forall k \in [0, K_0 - 1]$.
  for $k = 0, 1, \ldots, K_0 - 1$ **do**
    Predictor. Starting from $\bar{Y}_{t'_k}$, run the PFODE with stepsize h'_k to obtain $\bar{Y}'_{t'_{k+1}}$.
    Corrector. Starting from $\bar{Y}'_{t'_{k+1}}$, run ULMC for total time $1/\sqrt{L_{t'_k}}$ with stepsize h_{corr} to obtain $\bar{Y}_{t'_{k+1}}$.
  end for
end if
return $\bar{Y}_{T-\delta}$

Remark 5. *We note that the dependence of T and $1/\delta$ in the approximated score term is polynomial. Hence, if ϵ_{score} is much smaller than ϵ_{TV} and ϵ_{W_2}, we can consider the approximated score (Detail in Appendix E.3).*

5.3 Technical Challenge

In this section, we first briefly introduce the basic proof idea. Then, we focus on the two technical challenges.

Proof Sketch. The proof idea of the PFODE-Corrector type algorithm is similar to Chen et al. (2023c). For the predictor stage, the predictor pushes the reverse time to generate samples, bounded by $\partial_t \left\|Y_t - \bar{Y}_t\right\|^2$ in W_2 distance. However, if we use the Wasserstein analysis trivially, the distance between distributions grows exponentially with time t.

To deal with this problem, the algorithm introduces a ULD corrector stage to inject some suitable noise. The corrector allows the use of the data-processing inequality, which restarts the coupling at each stage to replace the exponential T with linear T.

The exponential-decay predictor stepsize. In this part, we show why the exponential-decay version algorithm is specifically for VESDE and introduces an additional $\exp(T)$ dependence for VPSDE. As an example, we run one predictor step in the reverse time $t \in [t'_k, t'_{k+1}]$ starting from the same distribution. For VPSDE, as shown in Chen et al. (2023c), the one step predictor error $\partial_t \left\|Y_t - \bar{Y}_t\right\|^2$ is equal to

$$2\langle Y_t - \bar{Y}_t, \frac{1}{2}(\nabla \ln q_{T-t}(Y_t) - s_{T-t'_k}(\bar{Y}_{t'_k}))\rangle + 2\|Y_t - \bar{Y}_t\|^2,$$

which is bounded by

$$\Big(2 + \frac{1}{h'_k}\Big)\|Y_t - \bar{Y}_t\|^2 + \frac{h'_k}{4}\|\nabla \ln q_{T-t}(Y_t) - s_{T-t'_k}(\bar{Y}_{t'_k})\|^2.$$

Hence, the one predictor step W_2^2 error is bounded by

$$\exp\big((\frac{1}{h'_k} + 2)h'_k\big) \times \int_{t'_{k-1}}^{t'_k} h'_k \mathbb{E}\Big[\|\nabla \ln q_{T-t}(Y_t) - s_{T-t'_k}(\bar{Y}_{t'_k})\|^2\Big] \mathrm{d}t\,.$$

Due to the drifted term of VPSDE, there is a $\exp(2h'_k)$, which leads to an additional $\exp(T)$ when using exponential decay stepsize $h_k = r t_k$. However, when considering VESDE, we know that

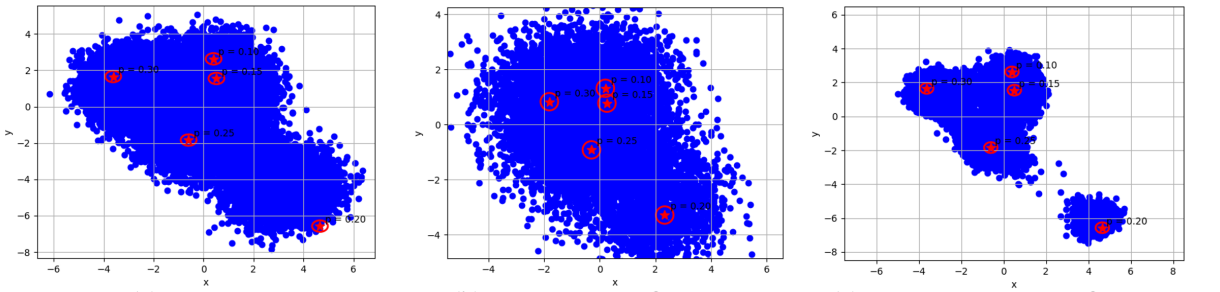
$$\dot{Y}_t = \frac{g(T-t)^2}{2} \nabla \ln q_{T-t}(Y_t)\,,$$

which indicates the W_2^2 error is bounded by

$$\int_{t'_k}^{t'_{k+1}} g(T-t)^4 h'_k \mathbb{E}\left[\|\nabla \ln q_{T-t}(Y_t) - s_{T-t'_k}(\bar{Y}_{t'_k})\|^2\right] \mathrm{d}t\,.$$

The above result indicates that the exponential dependence of h'_k is removed for VESDE, and the only requirement for h'_k is the discretization term of the corrector stage can be well controlled. Hence, we can use exponential decay stepsize to obtain better complexity.

The time-dependent corrector stage. After the predictor stage, Chen et al. (2023c) introduce an uniform corrector time $T_{\text{corr}} = 1/\sqrt{L_{\max}}$ to do the ULD



(a) DPUM (b) Pure PFODE for VESDE (c) PFODE-Corrector for VESDE

Figure 1: The experiments for the PFODE with algorithm (VESDE).

corrector. We show that the time-dependent corrector stage is more suitable for the exponential-decay predictor. When considering the exponential-decay predictor, the W_2 error of one predictor stage $W_2(p,q)$ for VESDE (SOTA) is bounded by $dh_k'^{1.5}/\sqrt{T-t_k'}$. Note that $h_0' \approx rT$, which indicate $h_0'^{1.5}/\sqrt{T-t_0'} \approx r^{1.5}T$. If we choose uniform time corrector $T_{\text{corr}} = 1/\sqrt{L_{\max}}$, the short regularization result $\text{TV}\left(\boldsymbol{p}P_{\text{ULD}}^{t_{k+1}',T_{\text{corr}}}, \boldsymbol{q}\right) \lesssim L_{\max}^{1/2} W_2(p,q)$, which indicates that the influence of T cannot be eliminated. However, with a time-dependent corrector time $T_{\text{corr}} = 1/\sqrt{L_{t_k'}}$, we have that

$$L_{t_k'}^{1/2} W_2(p,q) \lesssim dRh_k'^{1.5}/(\delta(T-t_k')^{1.5}),$$

which avoids the influence of T.

6 Experiments

Section 5 shows the exponential-decay PFODE-corrector algorithm enjoys a better predictor complexity compared with DPUM (an algorithm for VPSDE with reverse PFODE predictor and ULD corrector, Chen et al. (2023c)) and pure PFODE predictor for VESDE. In this part, we further provide simulation experiments to support our theoretical results.

Setting. The target distribution is a mixture of five Gaussian with $d = 2$, which is highly non-log-concave. Following Chen et al. (2023c), we use the closed-form score function of the mixture of Gaussian, use the uniform predictor stepsize at each stage $h_{\text{pred}} = T/K_0$, and set 3 corrector steps of the underdamped Langevin algorithm at each corrector stage (Details in Appendix G).

Observation. From the qualitative perspective, the experiments show that the PFODE-Corrector algorithm for VESDE can generate each clustering of the mixture of Gaussian with limited predictor steps even if the target distribution is a highly non-log-concave distribution (Fig. 1).

K_0	DPUM	Pure PFODE for VESDE	Pure -Corrector for VESDE
10	123.77	118.87	**29.73**
20	69.61	84.87	**21.67**
30	50.15	64.90	**20.44**
100	19.08	32.49	19.82

Table 2: KL divergence for generated and target data

From the quantitative perspective (Table 2), we show the DPUM algorithm and pure PFODE algorithm have a significantly larger error compared with our PFODE-Corrector algorithm for VESDE. For the DPUM algorithm, this phenomenon is due to the large predictor stepsize introduce a $\exp(T)$ dependence (Section 5.3). For the pure PFODE predictor, as shown in Table 1, this phenomenon is due to its exponential dependence guarantee. We also note that as K_0 becomes larger, the results of both different algorithms improve.

7 Conclusion

In this work, we analyze VE-based models and achieve SOTA polynomial iteration complexity for the SDE-based and PFODE-based implementations. For the SDE-based implementation, we prove the first polynomial results under bounded support assumption and realistic EDM stepsize. During the analysis, we explain why the current SOTA VE-based models perform better. After that, we further analyze the low-dimensional linear subspace data situation and obtain improved results, which makes the first step to explain the great performance of VESDE under the manifold data. For the PFODE-based implementation, we first propose a basic uniform PFODE-Corrector algorithm and provide the first polynomial iteration complexity for VESDE with a PFODE predictor. As a next step, by analyzing the property of VESDE, we design an exponential-decay PFODE-Corrector algorithm with a larger predictor stepsize to improve the results.

References

Iskander Azangulov, George Deligiannidis, and Judith Rousseau. Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv preprint arXiv:2409.18804*, 2024.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Patrick Cattiaux, Giovanni Conforti, Ivan Gentil, and Christian Léonard. Time reversal of diffusion processes under a finite entropy condition. *arXiv preprint arXiv:2104.07708*, 2021.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. *arXiv preprint arXiv:2211.01916*, 2022a.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2023a.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. *arXiv preprint arXiv:2302.07194*, 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. *arXiv preprint arXiv:2209.11215*, 2022b.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ode is provably fast. *arXiv preprint arXiv:2305.11798*, 2023c.

Sitan Chen, Giannis Daras, and Alexandros G Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. *arXiv preprint arXiv:2303.03384*, 2023d.

Xinlei Chen, Zhuang Liu, Saining Xie, and Kaiming He. Deconstructing denoising diffusion models for self-supervised learning, 2024.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2208.05314*, 2022.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

James Foster, Terry Lyons, and Harald Oberhauser. The shifted ode method for underdamped langevin mcmc. *arXiv preprint arXiv:2101.03446*, 2021.

Xuefeng Gao and Lingjiong Zhu. Convergence analysis for general probability flow odes of diffusion models in wasserstein distances. *arXiv preprint arXiv:2401.17958*, 2024.

Xuefeng Gao, Hoang M Nguyen, and Lingjiong Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. *arXiv preprint arXiv:2311.11003*, 2023.

Yingqing Guo, Hui Yuan, Yukang Yang, Minshuo Chen, and Mengdi Wang. Gradient guidance for diffusion models: An optimization perspective. *arXiv preprint arXiv:2404.14743*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022.

Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. *arXiv preprint arXiv:2106.05527*, 2021.

Dongjun Kim, Yeongmin Kim, Wanmo Kang, and Il-Chul Moon. Refining generative process with discriminator guidance in score-based diffusion models. *arXiv preprint arXiv:2211.17091*, 2022.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. *arXiv preprint arXiv:2310.02279*, 2023.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *arXiv preprint arXiv:2206.06227*, 2022.

Gen Li and Yuling Yan. Adapting to unknown low-dimensional structures in score-based diffusion models. *arXiv preprint arXiv:2405.14861*, 2024a.

Gen Li and Yuling Yan. $o(d/t)$ convergence theory for diffusion probabilistic models under minimal assumptions. *arXiv preprint arXiv:2409.18959*, 2024b.

Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv preprint arXiv:2306.09251*, 2023.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. A sharp convergence theory for the probability flow odes of diffusion models. *arXiv preprint arXiv:2408.02320*, 2024.

Yuchen Liang, Peizhong Ju, Yingbin Liang, and Ness Shroff. Non-asymptotic convergence of discrete-time diffusion models: New approach and improved rate. *arXiv preprint arXiv:2402.13901*, 2024.

Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input perturbation reduces exposure bias in diffusion models. *arXiv preprint arXiv:2301.11706*, 2023.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

Paul Thierry Yves Rolland. Predicting in uncertain environments: methods for robust machine learning. Technical report, EPFL, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

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. *arXiv preprint arXiv:2011.13456*, 2020b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Jan Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

Rong Tang and Yun Yang. Adaptivity of diffusion models to manifold structures. In *International Conference on Artificial Intelligence and Statistics*, pages 1648–1656. PMLR, 2024.

Jiayan Teng, Wendi Zheng, Ming Ding, Wenyi Hong, Jianqiao Wangni, Zhuoyi Yang, and Jie Tang. Relay diffusion: Unifying diffusion process across resolutions for image synthesis. *arXiv preprint arXiv:2309.03350*, 2023.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23 (7):1661–1674, 2011.

Yuqing Wang, Ye He, and Molei Tao. Evaluating the design space of diffusion-based generative models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Ruofeng Yang, Zhijie Wang, Bo Jiang, and Shuai Li. The convergence of variance exploding diffusion models under the manifold hypothesis, 2024. URL `https://openreview.net/forum?id=tD4NOxYTfg`.

Hui Yuan, Kaixuan Huang, Chengzhuo Ni, Minshuo Chen, and Mengdi Wang. Reward-directed conditional diffusion: Provable distribution estimation and reward improvement. *arXiv preprint arXiv:2307.07055*, 2023.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]

   Justification: The sampling algorithms have been clearly introduced in Section 3 and Algorithm 1.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable]
   (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable]
   (c) Clear explanations of any assumptions. [Yes/No/Not Applicable]

   Justification: All assumptions, Theorem, Corollary, and proof sketch have been clearly stated in the main content. The detailed proof appears in the appendix.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable]

   Justification: The experiments details are presented in Appendix G and the code is provided in a GitHub repository.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes/No/Not Applicable]
   (b) The license information of the assets, if applicable. [Yes/No/Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable]
   (d) Information about consent from data providers/curators. [Yes/No/Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable]

Appendix

A Notation

We define by q_0 the target data distribution with dimension d and $\{q_t\}_{t\in[0,T]}$ the distribution of X_t (the random variable of the forward process). For the timeline, we define by $\delta = t_0 \le t_1 \le \cdots \le t_K = T$ the discretization points in the forward timeline and $h_k := t_k - t_{k-1}$ as the step size. When consider the reverse process, we define $t'_k = T - t_{K-k}$ and $h'_k = h_{K-k}$.

The reverse SDE setting. we define by $p_{T-\delta}$ be the output of the discrete reverse SDE process with the approximated score (Eq. (5), K predictor steps). The iteration complexity of the reverse SDE is the number of K to guarantee $\mathrm{KL}\left(p_{T-\delta}, q_\delta\right) \le \epsilon_{\mathrm{KL}}^2$ and $W_2(q_0, q_\delta) \le \epsilon_{W_2}$.

The PFODE-Corrector algorithm. We define by h_{pred} and h_{corr} the stepsize of predictor and corrector and K_0 the total number of stages of Algorithm 1. Let $p_{T-\delta}^{\mathrm{ULMC}}$ be the output of Algorithm 1. The iteration complexity contains two parts: the predictor and the corrector iteration complexity. The goal is to guarantee $\mathrm{TV}(p_{T-\delta}^{\mathrm{ULMC}}, q_\delta) \le \epsilon_{\mathrm{TV}}$ and $W_2(q_0, q_\delta) \le \epsilon_{W_2}$.

- The predictor iteration complexity: $K_0 T_{\mathrm{pred}}/h_{\mathrm{pred}} = T/h_{\mathrm{pred}}$ (uniform version); K_0 (the exponential-decay version).
- The complexity of the iteration of the corrector: $K_0 T_{\mathrm{corr}}/h_{\mathrm{corr}}$ (since the corrector runs T_{corr} time at each stage).

For each step of the PFODE-Corrector algorithm, since we need to switch between the predictor and the corrector and restart repeatedly, we do not specify the beginning distribution of the following Markov kernels in this section. In the detailed proof, we will specify the starting distribution of these processes.

- $Q_{\mathrm{ODE}}^{t,h}$ is the output of running the continuous ODE for time h, starting at (reverse) time t.
- P_{ULD} is the output of running the continuous underdamped Langevin corrector for time h.
- $\bar{P}_{\mathrm{ODE}}^{t,h}$ and $\bar{P}_{\mathrm{ULMC}}$ are the corresponding implementable algorithm with estimated score.

B The Detailed Calculation of Previous work

In this section, we show the detailed calculation to obtain the results of VPSDE in this paper.

VPSDE with reverse SDE. In this part, we prove the iteration complexity of VPSDE with reverse SDE under the EDM stepsize. As shown in Lemma 17 of Chen et al. (2023a), the discretization error for VPSDE is

$$\sum_{k=1}^{K} \int_{t_{k-1}}^{t_k} \mathbb{E}\left\|\nabla \log q_t\left(X_t\right) - \nabla \log q_{t_k}\left(X_{t_k}\right)\right\|^2 \, \mathrm{d}t \lesssim d^2 \sum_{k=1}^{K} \frac{h_k^2}{\sigma_{t_{k-1}}^4} .$$

When t_{k-1} is close to δ, $\sigma_{t_{k-1}}^4 = t_{k-1}^2$, which has the same order compared to Theorem 1. When t_{k-1} is close to T, $\sigma_{t_{k-1}}^4$ is close to 1 and the discretization in this interval becomes $d^2 h_k^2$, which is not a dominated term [4]. Hence, the discretization error for VPSDE is $d^2(T/\delta)^{\frac{1}{a}}/K$, which is exactly the same as the results of VESDE. For T, as shown in Lemma 9 of Chen et al. (2023a), it is a logarithmic term and can be ignored. For the early stopping δ, Corollary 3 of Chen et al. (2022b) shows that $\delta \asymp \varepsilon_{W_2}^2/(\sqrt{d}(R \vee \sqrt{d}))$ for VPSDE. Then, we achieve the final results by requiring $d^2(T/\delta)^{\frac{1}{a}}/K$ smaller than ϵ_{KL}^2.

[4] The augmentation is exactly the same with Lemma 18 of Chen et al. (2023a).

VPSDE with reverse PFODE. Theorem 3 of Chen et al. (2023c) achieve $\widetilde{\Theta}\left(\frac{L^2 d^{1/2}}{\epsilon_{\mathrm{TV}}}\right)$ iteration complexity for DPUM algorithm under the L-Lipschitz score function. As we shown in Lemma 12, under the bounded support assumption, the Lipschitz constant at time t is $\left(1+R^2\right)/\sigma_t^4$. Since Chen et al. (2023c) assume L constant holds for all t, we choose $L = L_{\max} = (1+R^2)/\sigma_\delta^4 = (1+R^2)/\delta^2$. As shown in the above paragraph, we need to choose $\delta \asymp \varepsilon_{W_2}^2/(\sqrt{d}(R \vee \sqrt{d}))$ to obtain the final results.

C The Proof for VESDE with Reverse SDE

Similar to Chen et al. (2023a), we can achieve the KL divergence instead of TV distance by using the chain rule of the KL divergence. Before introducing the KL version of the Girsanov lemma, we also need to check some properties of the reverse process.

Lemma 1. *For $0 \le k \le K-1$, consider the reverse SDE starting from $Y_{t'_k} = a$*

$$\mathrm{d}Y_t = g(T-t)^2 \nabla \log q_{T-t}\left(Y_t\right)\mathrm{d}t + g(T-t)\mathrm{d}B_t, \quad Y_{t'_k} = a\,,$$

and the discretization process in this small interval

$$\mathrm{d}\bar{Y}_t = g(T-t)^2 s_{T-t'_k}(a)\mathrm{d}t + g(T-t)\mathrm{d}B_t, \quad \bar{Y}_{t'_k} = a$$

for time $t \in \left(t'_k, t'_{k+1}\right]$. Let $\tilde{q}_{t|t'_k}$ be the density of Y_t given $Y_{t'_k}$ and $p_{t|t'_k}$ be density of $\bar{Y}_t$ given $\bar{Y}_{t'_k}$. Then we have

- *For any $a \in \mathbb{R}^d$, the two processes satisfy the uniqueness and regularity condition: the above two processes have a unique solution and $\tilde{q}_{t|t'_k}(\cdot|a), p_{t|t'_k}(\cdot|a) \in C^2\left(\mathbb{R}^d\right)$ for $t > t'_k$.*
- *For a.e. $a \in \mathbb{R}^d$ (with respect to the Lebesgue measure):*

$$\lim_{t \to t'_k +} \mathrm{KL}\left(\tilde{q}_{t|t'_k}(\cdot|a) \| p_{t|t'_k}(\cdot|a)\right) = 0$$

Proof. For the discretization process, since conditionally on $\bar{Y}_{t'_k}$, the next iterate $\bar{Y}_{t'_{k+1}}$ has an explicit Gaussian distribution and $g(T-t)^2 = 1$ and $T-t$ for VESDE (SMLD) and VESDE (SOTA) respectively, the uniqueness and regularity holds. For the continuous process, similar to Chen et al. (2023a), the uniqueness is guaranteed by the local Lipshitz property of $\nabla \log q_{T-t}$ and the definition of $g^2(T-t)$ since $\tilde{q}_t \in C^2\left(\mathbb{R}^d\right)$ is supported on $\mathbb{R}^d$. For the regularity, we have that

$$\tilde{q}_{t|t'_k}(x|a) = q_{T-t|T-t'_k}(x|a) = \frac{q_{T-t}(x) q_{T-t'_k|T-t}(a|x)}{q_{T-t'_k}(a)}\,.$$

Since $q_{T-t'_k|T-t}(a|x)$ has distribution $\mathcal{N}\left(x, \left(\sigma^2_{T-t'_k} - \sigma^2_{T-t}\right) I\right)$, it is smooth for any $a \in \mathbb{R}^d$, and we know that $\tilde{q}_{t|t'_k}(x|a) \in C^2\left(\mathbb{R}^d\right)$. For the second property, we define $\mathbb{Q}_{[t'_k,t]}$ and $\mathbb{P}_{[t'_k,t]}$ the path measure of $(Y_s)_{t'_k \le s \le t}$ and $\left(\bar{Y}_s\right)_{t'_k \le s \le t}$. Then by the augmentation of Lemma 7 of Chen et al. (2023a), we then check the following inequality hold for a.e. $a \in \mathbb{R}^d$ when $t - t'_k$ is sufficient small

$$\mathbb{E}\left[\exp\left(\int_{t'_k}^t g(T-s)^2 \left\|\nabla \log q_{T-s}\left(Y_s\right)\right\|^2 \,\mathrm{d}s\right) | Y_{t'_k} = a\right] < \infty\,.$$

By Lemma 14, we know that $\left\|\nabla \log q_t\left(X_t\right)\right\|_{\psi_2} \lesssim \sqrt{\frac{d}{\sigma_t^2}}$. Hence, we know that

$$\begin{aligned}\left\|\int_{t'_k}^t g(T-s)^2 \|\nabla \log q_{T-s}\left(Y_s\right)\|^2 \,\mathrm{d}s\right\|_{\psi_1} &\le \int_{t'_k}^t g(T-s)^2 \left\|\nabla \log q_{T-s}\left(Y_s\right)\right\|_{\psi_2}^2 \,\mathrm{d}s \\ &\lesssim \frac{d}{T-t}\left(t - t'_k\right)\,.\end{aligned}$$

The last inequality follows that $\sigma^2_{T-t} = T-t$ and $g(T-t)^2 = 1$ for VESDE (SMLD) and $\sigma^2_{T-t} = (T-t)^2$ and $g(T-t)^2 = T-t$ for VESDE (SOTA). ■

After obtaining the above lemma, we obtain the Girsanov Lemma for VESDE.

Lemma 2. *Assume* **Assumption 2** *holds, we have that*

$$\begin{aligned}&\mathrm{KL}\left(q_\delta \| p_{T-\delta}\right)\\ &\lesssim \mathrm{KL}\left(q_T \| q_\infty\right)+\sum_{k=1}^{K} \int_{t_{k-1}}^{t_k} g(t)^2 \mathbb{E}\left\|\nabla \log q_t\left(X_t\right)-\nabla \log q_{t_k}\left(X_{t_k}\right)\right\|^2 \mathrm{~d} t+\sum_{k=1}^{K} \frac{h_k g(t_k)^2 \epsilon_{\text{score}}^2}{\sigma_{t_k}^2} .\end{aligned}$$

Proof. For $t_k' < t \le t_{k+1}'$, let $\tilde{q}_{t|t_k'}$ be the distribution of Y_t given $Y_{t_k'}$ and $p_{t|t_k'}$ be the distribution of $\bar{Y}_t$ given $\bar{Y}_{t_k'}$. After checking the two property in Lemma 1, we can use Lemma 6 and proposition 8 of Chen et al. (2023a) to achieve the following inequality for any $a \in \mathbb{R}^d$ and $t > t_k'$.

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d} t} \mathrm{KL}\left(\tilde{q}_{t|t_k'}(\cdot|a) \| p_{t|t_k'}(\cdot|a)\right)=-g(T-t)^2 \frac{1}{2} \mathbb{E}_{\tilde{q}_{t|t_k'}(y|a)}\left\|\nabla \log \frac{\tilde{q}_{t|t_k'}(y|a)}{p_{t|t_k'}(y|a)}\right\|^2\\ &+\mathbb{E}_{\tilde{q}'_{t|t_k'}(y|a)}\left[\left\langle g(T-t)\left(\nabla \log \tilde{q}_t(y)-s_{t_{K-k}}(a)\right), g(T-t) \nabla \log \frac{\tilde{q}_{t|t_k'}(y|a)}{p_{t|t_k'}(y|a)}\right\rangle\right]\\ &\le \frac{1}{2} g(T-t)^2 \mathbb{E}_{\tilde{q}_{t|t_k'}(y|a)}\left\|s_{t_{K-k}}(a)-\nabla \log \tilde{q}_t(y)\right\|^2 ,\end{aligned}$$

where the last inequality follows the fact that $\langle v, w\rangle \le \frac{1}{2}\|v\|^2+\frac{1}{2}\|w\|^2$. Then, by using the chain rule of KL divergence and similar augmentation compared to Chen et al. (2023a), we have that

$$\begin{aligned}&\mathrm{KL}\left(\tilde{q}_{t'_{k+1}} \| p_{t'_{k+1}}\right)\\ &\le \mathbb{E}_{\tilde{q}_{t_k'}(a)} \mathrm{KL}\left(\tilde{q}_{t'_{k+1}|t_k'}(\cdot|a) \| p_{t'_{k+1}|t_k'}(\cdot|a)\right)+\mathrm{KL}\left(\tilde{q}_{t_k'} \| p_{t_k'}\right)\\ &\le \mathrm{KL}\left(\tilde{q}_{t_k'} \| p_{t_k'}\right)+\frac{1}{2} \int_{t_k'}^{t'_{k+1}} g(T-t)^2 \mathbb{E}\left\|s_{T-t_k'}\left(Y_{t_k'}\right)-\nabla \log q_{T-t}\left(Y_t\right)\right\|^2 \mathrm{~d} t .\end{aligned}$$

Summing over $k=0,1, \ldots, K-1$ and using $q_t=\tilde{q}_{T-t}$ to convert the reverse time to the forward time, we obtain

$$\begin{aligned}&\mathrm{KL}\left(q_\delta \| p_{T-\delta}\right)\\ &\le \mathrm{KL}\left(q_T \| q_\infty\right)+\frac{1}{2} \sum_{k=0}^{K-1} \int_{t_k'}^{t'_{k+1}} g(T-t)^2 \mathbb{E}\left\|s_{T-t_k'}\left(Y_{t_k'}\right)-\nabla \log q_{T-t}\left(Y_t\right)\right\|^2 \mathrm{~d} t\\ &\le \mathrm{KL}\left(q_T \| q_\infty\right)+\frac{1}{2} \sum_{k=1}^{K} \int_{t_{k-1}}^{t_k} g(t)^2\left\|s_{t_k}(X_{t_k})-\nabla \log q_t\left(X_t\right)\right\|^2 \mathrm{~d} t\\ &\le \mathrm{KL}\left(q_T \| q_\infty\right)+\sum_{k=1}^{K} \int_{t_{k-1}}^{t_k} g(t)^2\left\|s_{t_k}(X_{t_k})-\nabla \log q_{t_k}\left(X_{t_k}\right)\right\|^2 \mathrm{~d} t\\ &+\sum_{k=1}^{K} \int_{t_{k-1}}^{t_k} g(t)^2\left\|\nabla \log q_{t_k}\left(X_{t_k}\right)-\nabla \log q_t\left(X_t\right)\right\|^2 \mathrm{~d} t\\ &\le \mathrm{KL}\left(q_T \| q_\infty\right)+\sum_{k=1}^{K} \int_{t_{k-1}}^{t_k} g(t)^2\left\|\nabla \log q_{t_k}\left(X_{t_k}\right)-\nabla \log q_t\left(X_t\right)\right\|^2 \mathrm{~d} t+\sum_{k=1}^{K} \frac{h_k \epsilon_{\text{score}}^2 g(t_k)^2}{\sigma_{t_k}^2} ,\end{aligned}$$

where the last inequality follows **Assumption 2**. ■

Theorem 1. *Assume Assumption 1 and 2 hold. Let $p_{T-\delta}$ be the output of Eq.* (5)*. If using the EDM stepsize (Eq.* (4)*), then we have that*

$$\mathrm{KL}\left(p_{T-\delta}, q_\delta\right) \le \bar{D}^2/\sigma_T^2+d^2(T/\delta)^{\frac{1}{a}}/K+\epsilon_{\text{score}}^2 \log(T/\delta) ,$$

for VESDE (SMLD) and VESDE (SOTA), where c is the eigenvalue of $\mathrm{Cov}\left[q_0\right]$ *with the largest absolute value and $\bar{D}=d|c|+\mathbb{E}[q_0]+R$.*

Furthermore, by choosing $\sigma_T^2 \geq \bar{D}^2/\epsilon_{\mathrm{KL}}^2$, $\sigma_\delta^2 = \epsilon_{W_2}^2/d$, *the output is* $\widetilde{O}\left(\epsilon_{\mathrm{KL}}^2 + \epsilon_{\mathrm{score}}^2\right)$*-close to* q_δ, *which is* $\epsilon_{W_2}^2$*-close to* q_0, *with iteration complexity*

$$K = \Theta\left(\frac{d^{2+\frac{1}{2a}}\bar{D}^{1/a}}{\epsilon_{\mathrm{KL}}^{2+1/a}\epsilon_{W_2}^{1/a}}\right).$$

for VESDE (SOTA). For VESDE (SMLD), we require $K = \Theta\left(d^{2+\frac{1}{a}}\bar{D}^{\frac{2}{a}}/\left(\epsilon_{\mathrm{KL}}^{2+2/a}\epsilon_{W_2}^{2/a}\right)\right)$ *to achieve the same guarantee.*

Proof. By using Lemma 15 and Lemma 16, we know that

$$\begin{aligned}
&\sum_{k=1}^{K}\int_{t_{k-1}}^{t_k} g(t)^2\mathbb{E}\left\|\nabla\log q_t\left(X_t\right) - \nabla\log q_{t_k}\left(X_{t_k}\right)\right\|^2 \,\mathrm{d}t \\
&\lesssim \sum_{k=1}^{K}\int_{t_{k-1}}^{t_k}\frac{d^2 g(t)^2\left(\sigma_{t_k}^2 - \sigma_t^2\right)}{\sigma_t^4}\,\mathrm{d}t \lesssim \sum_{k=1}^{K}\int_{t_{k-1}}^{t_k}\frac{d^2\left(t_k - t\right)}{t_{k-1}^2}\,\mathrm{d}t \lesssim d^2\sum_{k=1}^{K}\frac{h_k^2}{t_{k-1}^2}.
\end{aligned}$$

The second inequality follows the fact that (1) for VESDE (SMLD), $g(t)^2 = 1, \sigma_t^2 = t$; (2) for VESDE (SOTA), $g(t)^2 = 2t, \sigma_t^2 = t^2$ and $\frac{t_k}{t_{k-1}} \leq 2$. For the rest of this theorem, we consider two stepsize (1) the exponential decay stepsize; (2) EDM stepsize.

Similar to Chen et al. (2023a), for exponential decay stepsize $h_k := t_k - t_{k-1} = rt_k$, where $r \leq 1/2d$, we have that

$$\sum_k \frac{h_k^2}{t_{k-1}^2} = \sum_k \frac{r^2 t_k^2}{t_{k-1}^2} \asymp r^2\frac{\log(T/\delta)}{r} = r\log(T/\delta).$$

We also know that for the exponential decay stepsize, the number of stepsize $K \lesssim \frac{1}{r}\log(T/\delta)$. Then, we complete our prove for the exponential decay stepsize. For $\epsilon_{\mathrm{score}}^2$ term,

$$\sum_{k=1}^{K}\frac{\epsilon_{\mathrm{score}}^2 h_k}{t_k} \lesssim \epsilon_{\mathrm{score}}^2\log(T/\delta).$$

The above results show that with the exponential decay stepsize, VPSDE, VESDE (SMLD), and VESDE (SOTA) have similar iteration complexity. However, as mentioned in Chen et al. (2023a), the exponential decay stepsize has not yet been experimentally verified to be effective. Hence, we then focus on the EDM stepsize Corollary 1, which shows SOTA performance:

$$t_k = (\delta + kh)^a, h = \frac{T^{\frac{1}{a}} - \delta}{K}.$$

In this timestep, we know that $\frac{h_k}{h} \asymp t_k^{\frac{a-1}{a}}$ and

$$\sum_{k=1}^{K}\frac{h_k^2}{t_{k-1}^2} \asymp h\sum_{k=1}^{K}\frac{h_k}{t_k^{\frac{a+1}{a}}} \asymp h\int_\delta^T \frac{1}{t^{\frac{a+1}{a}}}\,\mathrm{d}t \asymp h\delta^{-\frac{1}{a}} \asymp \frac{(T/\delta)^{\frac{1}{a}}}{K}.$$

To achieve the total bound is smaller than ϵ_0^2 when considering KL divergence, we need

$$K = \frac{d^2(T/\delta)^{1/a}}{\epsilon_{\mathrm{KL}}^2},$$

For VESDE (SMLD), we choose $T \geq \frac{(d|c| + \mathbb{E}[q_0] + R)^2}{\epsilon_{\mathrm{KL}}^2}$ and $\delta = \frac{\epsilon_{W_2}^2}{d}$, then the final iteration complexity is

$$K = \frac{d^{2+\frac{1}{a}}\left(d|c| + \mathbb{E}\left[q_0\right] + R\right)^{\frac{2}{a}}}{\epsilon_{\mathrm{KL}}^{2+\frac{2}{a}}\epsilon_{W_2}^{\frac{2}{a}}}.$$

For VESDE (SOTA), we choose $T \geq \frac{d|c|+\mathbb{E}[q_0]+R}{\epsilon_{\mathrm{KL}}}$ and $\delta = \frac{\epsilon_{W_2}}{\sqrt{d}}$, then the final iteration complexity is

$$K = \frac{d^{2+\frac{1}{2a}} \left(d|c| + \mathbb{E}\left[q_0\right] + R\right)^{\frac{1}{a}}}{\epsilon_{\mathrm{KL}}^{2+\frac{1}{a}} \epsilon_{W_2}^{\frac{1}{a}}} .$$

We can also use a similar technique mentioned in Chen et al. (2023a) to obtain the iteration complexity for VPSDE under EDM stepsize. For VPSDE, choosing $T \geq \log\left(\frac{d+R}{\epsilon_{\mathrm{KL}}^2}\right)$ and $\delta = \frac{\epsilon_{W_2}^2}{d}$, the iteration complexity is

$$K = \frac{d^{2+\frac{1}{a}} \log\left(\frac{d+R}{\epsilon_{\mathrm{KL}}^2}\right)^{\frac{1}{a}}}{\epsilon_{\mathrm{KL}}^2 \epsilon_{W_2}^{\frac{2}{a}}} .$$

■

At the end of this section, we use the projection technique to achieve pure W_2 guarantee for VESDE with reverse SDE by using similar technique compared to Chen et al. (2022b).

Corollary 2. *Defined by $p_{T-\delta,R_0}$ the output $p_{T-\delta}$ of Corollary 1 projected onto $\mathrm{B}\left(0,R_0\right)$ for $R_0 = \widetilde{\Theta}(R)$. Then, $W_2\left(p_{T-\delta,R_0}, q_0\right) \leq \epsilon_{W_2}$ with iteration complexity $K = \widetilde{\Theta}\left(d^2R^4\log^2(T/\delta)/\epsilon_{W_2}^4\right)$ and $\epsilon_{\mathrm{score}} \leq \widetilde{O}(\epsilon_{W_2})$.*

Proof. For $R_0 > 0$, let Π_{R_0} denote the projection onto $\mathrm{B}\left(0,R_0\right)$. To achieve pure W_2 guarantee, we do the following decomposition:

$$W_2\left(\left(\Pi_{R_0}\right)_{\#} p_{T-\delta}, q_0\right) \leq W_2\left(\left(\Pi_{R_0}\right)_{\#} p_{T-\delta}, \left(\Pi_{R_0}\right)_{\#} q_\delta\right) + W_2\left(\left(\Pi_{R_0}\right)_{\#} q_\delta, q_0\right) .$$

For the first term, we can upper bound the Wasserstein distance by the KL divergence:

$$\begin{aligned} W_2\left(\left(\Pi_{R_0}\right)_{\#} p_{T-\delta}, \left(\Pi_{R_0}\right)_{\#} q_\delta\right) &\lesssim R_0\sqrt{\mathrm{TV}\left(\left(\Pi_{R_0}\right)_{\#} p_{T-\delta}, \left(\Pi_{R_0}\right)_{\#} q_\delta\right)} + R_0\exp\left(-R_0\right) \\ &\leq R_0\left(\mathrm{KL}\left(\left(\Pi_{R_0}\right)_{\#} p_{T-\delta}, \left(\Pi_{R_0}\right)_{\#} q_\delta\right)\right)^{1/4} + R_0\exp\left(-R_0\right) , \end{aligned}$$

where the first inequality follows Lemma 9 of Rolland (2022) and the second inequality follows the pinsker's inequality. By the data-processing inequality, we know that

$$\mathrm{KL}\left(\left(\Pi_{R_0}\right)_{\#} p_{T-\delta}, \left(\Pi_{R_0}\right)_{\#} q_\delta\right) \leq \mathrm{KL}\left(p_{T-\delta}, q_\delta\right) \leq \epsilon_{\mathrm{KL}}^2 ,$$

where the last inequality comes from Corollary 1. Then, we take $R_0 \geq R$ so that $\left(\Pi_{R_0}\right)\#q_0 = q_0$. Since Π_{R_0} is 1-Lipschitz, we have that

$$W_2\left(\left(\Pi_{R_0}\right)_{\#} q_\delta, q_0\right) = W_2\left(\left(\Pi_{R_0}\right)_{\#} q_\delta, \left(\Pi_{R_0}\right)_{\#} q_0\right) \leq W_2\left(q_\delta, q_0\right) \leq \varepsilon_{W_2} .$$

Combined with these two terms, we know that

$$W_2\left(\left(\Pi_{R_0}\right)_{\#} p_{T-\delta}, q_\delta\right) \lesssim R_0\sqrt{\epsilon_{\mathrm{KL}}} + R_0\exp\left(-R_0\right) + \epsilon_{W_2}.$$

Then, we only need choose $R_0 = \widetilde{\Theta}(R)$, and $\epsilon_{\mathrm{KL}} = \widetilde{\Theta}\left(\epsilon_{W_2}^2/R^2\right)$ to obtain the final results. ■

C.1 The Proof for Low Dimensional Data

In this part, we first show that the diffusion process happens in the latent space. As mentioned in Chen et al. (2023b), when assuming linear distribution, the ground-truth score function is decomposed into the latent score function $\nabla \log q_t^{\mathrm{LD}}(Z')$ and linear encoder and decoder:

$$\nabla \log q_t(X) = A\nabla \log q_t^{\mathrm{LD}}\left(A^\top X\right) - \frac{1}{\sigma_t^2}\left(I_d - AA^\top\right)X ,$$

where $q_t^{\mathrm{LD}}(Z') = \int q_t(Z'|Z)\, q_z(Z)\mathrm{d}Z$ and $q_t(\cdot|Z) = \mathcal{N}(Z, \sigma_t^2 I_{d'})$. This form indicates that the diffusion process happens in the latent space, which is the key of better iteration complexity. Then, we provide a detail proof of the iteration complexity of VESDE when considering the low dimensional data to explain why VESDE is more suitable for the manifold data

Theorem 2. *Assume* **Assumption 2** *and 3 holds. If choosing* $\sigma_T^2 \geq \bar{D}'^2/\epsilon_{\mathrm{KL}}^2$, $\sigma_\delta^2 = \epsilon_{W_2}^2/d$ *and using exponential decay stepsize with* $r = \epsilon_{\mathrm{KL}}^2/\log(T/\delta)$, *then for VESDE (SMLD) and VESDE (SOTA), the output is* $\widetilde{O}\left(\epsilon_{\mathrm{KL}}^2 + \epsilon_{\mathrm{score}}^2\right)$*-close to* q_δ, *which is* $\epsilon_{W_2}^2$*-close to* q_0, *with iteration complexity*

$$K = \widetilde{\Theta}\left(\left(d'^3 + d - d'\right)\log^2(T/\delta)/\epsilon_{\mathrm{KL}}^2\right).$$

Proof. With similar proof idea compared to Theorem 1, we first control the discretization error term. By using the second part of Lemma 15, we know that

$$\begin{aligned}
&\sum_{k=1}^{K}\int_{t_{k-1}}^{t_k} g(t)^2\mathbb{E}\left\|\nabla\log q_t(X_t) - \nabla\log q_{t_k}(X_{t_k})\right\|^2\,\mathrm{d}t \\
&\lesssim \sum_{k=1}^{K}\int_{t_{k-1}}^{t_k}\frac{\left(d'^3 + \left\|I_d - AA^\top\right\|_F^2\right)g(t)^2\left(\sigma_{t_k}^2 - \sigma_t^2\right)}{\sigma_t^4}\,\mathrm{d}t\,.
\end{aligned}$$

This results only uses $d'^3 + \left\|I_d - AA^\top\right\|_F^2$ to replace d^2. The rest proof is exactly the same compared to Corollary 1. ■

D The Proof for Uniform PFODE-Corrector Algorithm

D.1 The Predictor Stage

In the section, we first prove that when starting at the same distribution q at time t_0, the W_2 distance for reverse PFODE can be controlled under a suitable middle predictor step (consist of many h_{pred}). First, we define $L_{\max} = \frac{1+R^2}{\sigma_\delta^4}$, which corresponding to maximum middle predictor step $T_{\mathrm{pred}} = 1/L_{\max}$. Then, we prove that if the middle predictor is decomposed into many h_{pred}, the W_2 error is still controlled. In the rest of this subsection, we omit the index pred.

Lemma 3. *Suppose* **Assumption 1** *and* **Assumption 2** *hold and choosing* $L_{\max} = \frac{1+R^2}{\sigma_\delta^4}$. *Assume that* $h \lesssim 1/L_{\max}$. *Then*

$$W_2\left(qQ_{\mathrm{ODE}}^{t_0,h}, q\bar{P}_{\mathrm{ODE}}^{t_0,h}\right) \lesssim h^2R^2(2R+\sqrt{d})\frac{g(T-t_0-h)^4}{\sigma_{T-t_0-h}^6} + h\frac{g(T-t_0-h)^2\epsilon_{\mathrm{score}}}{\sigma_{T-t_0-h}}\,.$$

Proof. We know that the reverse PFODE is

$$\begin{aligned}
\dot{Y}_t &= \frac{g(T-t)^2}{2}\nabla\ln q_{T-t}(Y_t)\,, \\
\dot{\bar{Y}}_t &= \frac{g(T-t)^2}{2}s_{T-t_0}\left(\bar{Y}_{t_0}\right)\,,
\end{aligned}$$

for $t_0 \leq t \leq t_0 + h$, with $Y_{t_0} = \bar{Y}_{t_0} \sim q, Y_{t_0+h} \sim qQ_{\mathrm{ODE}}$, and $\bar{Y}_{t_0+h} \sim q\bar{P}_{\mathrm{ODE}}$. Then, we have that

$$\begin{aligned}
\partial_t\left\|Y_t - \bar{Y}_t\right\|^2 &= 2\left\langle Y_t - \bar{Y}_t, \dot{Y}_t - \dot{\bar{Y}}_t\right\rangle \\
&= 2\left\langle Y_t - \bar{Y}_t, \frac{g(T-t)^2}{2}\left(\nabla\ln q_{T-t}(Y_t) - s_{T-t_0}\left(\bar{Y}_{t_0}\right)\right)\right\rangle \\
&\leq \frac{1}{h}\left\|Y_t - \bar{Y}_t\right\|^2 + \frac{hg(T-t)^4}{4}\left\|\nabla\ln q_{T-t}(Y_t) - s_{T-t_0}\left(\bar{Y}_{t_0}\right)\right\|^2
\end{aligned}$$

By Grönwall's inequality, we have that

$$
\begin{aligned}
&\mathbb{E}\left[\left\|Y_{t_0+h}-\bar{Y}_{t_0+h}\right\|^2\right] \\
&\leq \int_{t_0}^{t_0+h} \frac{g(T-t)^4 h}{4} \mathbb{E}\left[\left\|\nabla \ln q_{T-t}\left(Y_t\right)-s_{T-t_0}\left(\bar{Y}_{t_0}\right)\right\|^2\right] \mathrm{d}t \\
&\lesssim h \int_{t_0}^{t_0+h} g(T-t)^4 \mathbb{E}\left[\left\|\nabla \ln q_{T-t}\left(Y_t\right)-s_{T-t_0}\left(\bar{Y}_{t_0}\right)\right\|^2\right] \mathrm{d}t \\
&\lesssim h \int_{t_0}^{t_0+h} h^2 g(T-t)^4\left\|\partial_t \nabla \log q_{T-t}\left(Y_{kh}\right)\right\|^2+\frac{g(T-t)^4 \epsilon_{\text{score}}^2}{\sigma_{T-t}^2} \mathrm{d}t \\
&\lesssim h^3 \int_{t_0}^{t_0+h} g(T-t)^4\left(\frac{g(T-t)^2}{\sigma_{T-t}^6} R^2(2R+\sqrt{d})\right)^2 \mathrm{d}t+h^2 \frac{g(T-t_0-h)^4 \epsilon_{\text{score}}^2}{\sigma_{T-t_0-h}^2} \\
&\lesssim h^4 R^4(2R+\sqrt{d})^2 \frac{g(T-t_0-h)^8}{\sigma_{T-t_0-h}^{12}}+h^2 \frac{g(T-t_0-h)^4 \epsilon_{\text{score}}^2}{\sigma_{T-t_0-h}^2},
\end{aligned}
$$

where the last two inequality holds since the dominant term in the final results is $1/\mathrm{Poly}(\delta)$, which appears when $t \to T-\delta$. ■

Before the following lemma, the reason why we do not consider the situation $T-t_0 \leq 1/L_{\max}=\frac{\sigma_\delta^4}{1+R^2}$ is that $\sigma_\delta^4 \leq \delta^2$ for two typical VESDE and $T-t_0 \geq \delta$.

Lemma 4. *Suppose that* **Assumption 1** *hold. Let* $N_{\text{pred}}=T_{\text{pred}}/h$*, where* $L_{\max}=\frac{1+R^2}{\sigma_\delta^4}$ *and* $T_{\text{pred}} \leq 1/L_{\max}$*. Then*

$$
\begin{aligned}
&W_2\left(qQ_{\text{ODE}}^{t_0,N_{\text{pred}}}, q\bar{P}_{\text{ODE}}^{t_0,N_{\text{pred}}}\right) \\
&\lesssim hT_{\text{pred}} R^2(2R+\sqrt{d}) \frac{g(T-t_0-T_{\text{pred}})^4}{\sigma_{T-t_0-T_{\text{pred}}}^6}+T_{\text{pred}} \frac{g\left(T-t_0-T_{\text{pred}}\right)^2 \epsilon_{\text{score}}}{\sigma_{T-t_0-T_{\text{pred}}}}
\end{aligned}
$$

Proof. Using the triangle inequality,

$$
\begin{aligned}
&W_2\left(qQ_{\text{ODE}}^{t_0,N_{\text{pred}}}, q\bar{P}_{\text{ODE}}^{t_0,N_{\text{pred}}}\right) \\
&\leq W_2\left(qQ_{\text{ODE}}^{t_0,N_{\text{pred}}}, qQ_{\text{ODE}}^{t_0,N_{\text{pred}}-1}\bar{P}_{\text{ODE}}\right)+W_2\left(qQ_{\text{ODE}}^{t_0,N_{\text{pred}}-1}\bar{P}_{\text{ODE}}, q\bar{P}_{\text{ODE}}^{t_0,N_{\text{pred}}}\right) \\
&\leq O\left(h^2 R^2(2R+\sqrt{d}) \frac{g(T-t_0-h)^4}{\sigma_{T-t_0-h}^6}+h \frac{g\left(T-t_0-h\right)^4 \epsilon_{\text{score}}^2}{\sigma_{T-t_0-h}^2}\right) \\
&\quad+\exp\left(O\left(L_{t_0+T_{\text{pred}}} h\right)\right) W_2\left(qP_{\text{ODE}}^{t_0,N_{\text{pred}}-1}, q\bar{P}_{\text{ODE}}^{t_0,N_{\text{pred}}-1}\right).
\end{aligned}
$$

By induction, we know that

$$
\begin{aligned}
&W_2\left(qQ_{\text{ODE}}^{t_0,N_{\text{pred}}}, q\bar{P}_{\text{ODE}}^{t_0,N_{\text{pred}}}\right) \\
&\lesssim \sum_{n=1}^{N_{\text{pred}}} h\left(hR^2(2R+\sqrt{d}) \frac{g(T-t_0-T_{\text{pred}})^4}{\sigma_{T-t_0-T_{\text{pred}}}^6}+\frac{g\left(T-t_0-T_{\text{pred}}\right)^4 \epsilon_{\text{score}}^2}{\sigma_{T-t_0-T_{\text{pred}}}^2}\right) \\
&\quad\times \exp\left(O\left(L_{t_0+T_{\text{pred}}} T_{\text{pred}}\right)\right) \\
&\lesssim hT_{\text{pred}} R^2(2R+\sqrt{d}) \frac{g(T-t_0-T_{\text{pred}})^4}{\sigma_{T-t_0-T_{\text{pred}}}^6}+T_{\text{pred}} \frac{g\left(T-t_0-T_{\text{pred}}\right)^2 \epsilon_{\text{score}}}{\sigma_{T-t_0-T_{\text{pred}}}}.
\end{aligned}
$$

The last inequality by the fact that $L_{t_0+T_{\text{pred}}} T_{\text{pred}} \leq O(1)$.

■

D.2 The Corrector Stage

In this section, we use the underdamped Langevin corrector to inject suitable noise to guarantee the data processing inequality holds. Since the underdamped Langevin corrector involves the vector term v_m, similar to Chen et al. (2023c), we use the following notation.

In Appendix D.1, starting at the same distribution q_{t_0}, we run the continuous process and discretization process for T_{pred}. In this section, we define the probability measures $p = q_{t_0}\bar{P}_{\text{ODE}}^{t_0,N_{\text{pred}}}$ and $q = q_{t_0}Q_{\text{ODE}}^{t_0,N_{\text{pred}}}$. We also write $\boldsymbol{p} := p \otimes \gamma_d$ and $\boldsymbol{q} := q \otimes \gamma_d$, where γ_d is the standard Gaussian measure in $\mathbb{R}^d$. Furthermore, we set the friction parameter to $\rho \asymp \sqrt{L_{\max}}$, where $L_{\max} = \frac{1+R^2}{\sigma_\delta^4}$.

To show that the TV error $\text{TV}\left(\boldsymbol{p}\bar{P}_{\text{LMC}}^{t_0+T_{\text{pred}},N_{\text{corr}}}, \boldsymbol{q}\right)$ is controlled after run Eq. (6), we need to upper bound $\text{TV}\left(\boldsymbol{p}P_{\text{ULD}}^{t_0+T_{\text{pred}},N_{\text{corr}}}, \boldsymbol{q}\right)$ and $\text{TV}\left(\boldsymbol{p}\bar{P}_{\text{ULMC}}^{t_0+T_{\text{pred}},N_{\text{corr}}}, \boldsymbol{p}P_{\text{ULD}}^{t_0+T_{\text{pred}},N_{\text{corr}}}\right)$, where $N_{\text{corr}} = \frac{T_{\text{corr}}}{h_{\text{corr}}}$. T_{corr} is the middle step of the corrector, which is defined later. We will omit the time index when it is clear from the context.

The following lemma comes from Lemma 9 of Chen et al. (2023c) and is used to control $\text{TV}\left(\boldsymbol{p}P_{\text{ULD}}^{t_0+T_{\text{pred}},N_{\text{corr}}}, \boldsymbol{q}\right)$.

Lemma 5. *If $T_{\text{corr}} \lesssim 1/\sqrt{L_{\max}}$, then*

$$\text{TV}\left(\boldsymbol{p}P_{\text{ULD}}^{t_0+T_{\text{pred}},N_{\text{corr}}}, \boldsymbol{q}\right) \lesssim \sqrt{\text{KL}\left(\boldsymbol{p}P_{\text{ULD}}^{t_0+T_{\text{pred}},N_{\text{corr}}} \| \boldsymbol{q}\right)} \lesssim \frac{W_2(p,q)}{L_{\max}^{1/4} T_{\text{corr}}^{3/2}}\,.$$

The remaining term is $\text{TV}\left(\boldsymbol{p}\bar{P}_{\text{ULMC}}^{t_0+T_{\text{pred}},N_{\text{corr}}}, \boldsymbol{p}P_{\text{ULD}}^{t_0+T_{\text{pred}},N_{\text{corr}}}\right)$, which correspond to the discretization error and Girsanov method. Since we run the underdamped corrector with a score function at a fixed time $t_0 + T_{\text{pred}}$, we need to introduce two benchmark processes, which correspond to the stationary distribution and continuous underdamped corrector at time $t = t_0 + T_{\text{pred}}$: $\mathrm{d}z_m^\circ = v_m^\circ \mathrm{d}m,\ \mathrm{d}z_m = v_m\ \mathrm{d}m,$

$$\begin{aligned}
\mathrm{d}v_m^\circ &= -\rho v_m^\circ \mathrm{d}m - \nabla U_{t_0+T_{\text{pred}}}(z_m^\circ)\,\mathrm{d}m + \sqrt{2\rho}\mathrm{d}B_m,\ (z_0^\circ, v_0^\circ) \sim \boldsymbol{q},\\
\mathrm{d}v_m &= -\rho v_m\ \mathrm{d}m - \nabla U_{t_0+T_{\text{pred}}}(z_m)\,\mathrm{d}m + \sqrt{2\rho}\mathrm{d}B_m,\ (z_0, v_0) \sim \boldsymbol{p}.
\end{aligned}$$

We know that for any integer $n_{\text{corr}} > 0$

$$(z_{n_{\text{corr}}h_{\text{corr}}}^\circ, v_{n_{\text{corr}}h_{\text{corr}}}^\circ) \sim \boldsymbol{q}P_{\text{ULD}}^{n_{\text{corr}}} = \boldsymbol{q}, \quad (z_{n_{\text{corr}}h_{\text{corr}}}, v_{n_{\text{corr}}h_{\text{corr}}}) \sim \boldsymbol{p}P_{\text{ULD}}^{n_{\text{corr}}}\,.$$

In this section, we omit the index of corr if it is clear from the context. The following lemma comes from Lemma 10 of Chen et al. (2023c) and controls the distance between the above two processes.

Lemma 6. *If $T_{\text{corr}} \lesssim 1/\sqrt{L_{\max}}$, then for all $0 \le m \le T_{\text{corr}}$,*

$$\mathbb{E}\left[\|z_m - z_m^\circ\|^2\right] \lesssim W_2^2(p,q)\,.$$

The following lemma controls the discretization error of the underdamped Langevin corrector.

Lemma 7. *Let $L_t = (R^2+1)/\sigma_{T-t}^4$. If $T_{\text{corr}} \lesssim 1/\sqrt{L_{\max}}$, then*

$$\begin{aligned}
&\text{TV}\left(\boldsymbol{p}\bar{P}_{\text{ULMC}}^{t_0+T_{\text{pred}},N_{\text{corr}}}, \boldsymbol{p}P_{\text{ULD}}^{t_0+T_{\text{pred}},N_{\text{corr}}}\right)\\
&\lesssim \sqrt{\text{KL}\left(\boldsymbol{p}P_{\text{ULD}}^{t_0+T_{\text{pred}},N_{\text{corr}}} \| \boldsymbol{p}\bar{P}_{\text{ULMC}}^{t_0+T_{\text{pred}},N_{\text{corr}}}\right)}\\
&\lesssim \frac{L_{t_0+T_{\text{pred}}} T_{\text{corr}}^{1/2}}{L_{\max}^{1/4}} W_2(p,q) + \frac{L_{t_0+T_{\text{pred}}} T_{\text{corr}}^{1/2}\sqrt{d}}{L_{\max}^{1/4}} h_{\text{corr}} + \frac{T_{\text{corr}}^{1/2}\epsilon_{\text{score}}}{L_{\max}^{1/4}\sigma_{T-t_0-T_{\text{pred}}}}\,.
\end{aligned}$$

Proof. With similar augmentation to Chen et al. (2023c) and Girsanov's theorem, we can obtain

$$\begin{aligned}
&\text{KL}\left(\boldsymbol{p}P_{\text{ULD}}^{N_{\text{corr}}} \| \boldsymbol{p}\bar{P}_{\text{ULMC}}^{N_{\text{corr}}}\right)\\
&\lesssim \frac{1}{\rho}\sum_{n=0}^{N_{\text{corr}}-1}\int_{nh_{\text{corr}}}^{(n+1)h_{\text{corr}}} \mathbb{E}\left[\left\|s_{T-t_0-T_{\text{pred}}}(z_{nh_{\text{corr}}}) - \nabla U_{T-t_0-T_{\text{pred}}}(z_u)\right\|^2\right]\mathrm{d}u
\end{aligned}$$

In the proof process of this lemma, we omit the time index $t_0 + T_{\text{pred}}$. Then, we have that

$$\begin{aligned}
\mathbb{E}\left[\left\|s\left(z_{nh_{\text{corr}}}\right) - \nabla U\left(z_u\right)\right\|^2\right] \lesssim \mathbb{E}\Bigg[& \left\|s\left(z_{nh_{\text{corr}}}\right) - s\left(z^{\circ}_{nh_{\text{corr}}}\right)\right\|^2 + \left\|s\left(z^{\circ}_{nh_{\text{corr}}}\right) - \nabla U\left(z^{\circ}_{nh_{\text{corr}}}\right)\right\|^2 \\
& + \left\|\nabla U\left(z^{\circ}_{nh_{\text{corr}}}\right) - \nabla U\left(z^{\circ}_u\right)\right\|^2 + \left\|\nabla U\left(z^{\circ}_u\right) - \nabla U\left(z_u\right)\right\|^2 \Bigg] \\
\lesssim\ & L^2_{t_0+T_{\text{pred}}}\mathbb{E}\left[\left\|z_{nh_{\text{corr}}} - z^{\circ}_{nh_{\text{corr}}}\right\|^2\right] + L^2_{t_0+T_{\text{pred}}}\mathbb{E}\left[\left\|z^{\circ}_{nh_{\text{corr}}} - z^{\circ}_u\right\|^2\right] \\
& + L^2_{t_0+T_{\text{pred}}}\mathbb{E}\left[\left\|z^{\circ}_u - z_u\right\|^2\right] + \frac{\epsilon^2_{\text{score}}}{\sigma^2_{T-t_0-T_{\text{pred}}}} \\
\lesssim\ & L^2_{t_0+T_{\text{pred}}} W_2^2(p,q) + L^2_{t_0+T_{\text{pred}}}\mathbb{E}\left[\left\|z^{\circ}_{nh_{\text{corr}}} - z^{\circ}_u\right\|^2\right] + \frac{\epsilon^2_{\text{score}}}{\sigma^2_{T-t_0-T_{\text{pred}}}}\,.
\end{aligned}$$

The second inequality by Assumption 4. We also know that

$$\mathbb{E}\left[\left\|z^{\circ}_{nh_{\text{corr}}} - z^{\circ}_u\right\|^2\right] = \mathbb{E}\left[\left\|\int_{nh_{\text{corr}}}^{u} v^{\circ}_s \mathrm{d}s\right\|^2\right] \le h_{\text{corr}}\int_{nh_{\text{corr}}}^{u}\mathbb{E}\left[\left\|v^{\circ}_s\right\|^2\right]\mathrm{d}s \le d h^2_{\text{corr}}\,.$$

The second inequality by the fact that $v^{\circ}_s \sim \gamma^d$. ■

D.3 The Detailed Proof for Theorem 2

Combined with the predictor step (Appendix D.1) and the corrector step (Appendix D.2), we proof Theorem 3. First, we provide the error guarantee after a middle step for predictor T_{pred} and corrector T_{corr}.

Lemma 8. *Let $L_{\max} = \frac{1+R^2}{\sigma^4_\delta}$, $L_t = \left(1+R^2\right)/\sigma^4_{T-t}$ and the stationary distribution $q_{t_0}Q^{t_0,N_{\text{pred}}}_{\text{ODE}} = q_{t_0+T_{\text{pred}}}$. For the predictor, $T_{\text{pred}} \le 1/L_{\max}$ and $N_{\text{pred}} = T_{\text{pred}}/h_{\text{pred}}$. For the underdamped Langenvin, $T_{\text{corr}} \le 1/\sqrt{L_{\max}}$ and $N_{\text{corr}} = T_{\text{corr}}/h_{\text{corr}}$. Then*

$$\begin{aligned}
& \text{TV}\left(p\bar{P}^{t_0,N_{\text{pred}}}_{\text{ODE}}\bar{P}^{N_{\text{corr}}}_{\text{ULMC}}, q_{t_0+T_{\text{pred}}}\right) \\
& \le \text{TV}\left(p, q_{t_0}\right) + O\Bigg(\frac{R^2(2R+\sqrt{d})g(T-t_0-T_{\text{pred}})^4}{\sigma^6_{T-t_0-T_{\text{pred}}}L^{1/2}_{\max}}h_{\text{pred}} + \frac{L_{t_0+T_{\text{pred}}}\sqrt{d}}{L^{1/2}_{\max}}h_{\text{corr}} \\
& \quad + \frac{\epsilon_{\text{score}}}{L^{1/2}_{\max}\sigma_{T-t_0-T_{\text{pred}}}} + \frac{g\left(T-t_0-T_{\text{pred}}\right)^2\epsilon_{\text{score}}}{L^{1/2}_{\max}\sigma_{T-t_0-T_{\text{pred}}}}\Bigg).
\end{aligned}$$

Proof. After the corrector step, we achieve a TV guarantee. Hence, we can use data-processing inequality.

$$\begin{aligned}
& \text{TV}\left(p\bar{P}^{t_0,N_{\text{pred}}}_{\text{ODE}}\bar{P}^{N_{\text{corr}}}_{\text{LMC}}, q_{t_0+T_{\text{pred}}}\right) \\
& \le \text{TV}\left(p\bar{P}^{t_0,N_{\text{pred}}}_{\text{ODE}}\bar{P}^{N_{\text{corr}}}_{\text{LMC}}, q_{t_0}\bar{P}^{t_0,N_{\text{pred}}}_{\text{ODE}}\bar{P}^{N_{\text{corr}}}_{\text{LMC}}\right) + \text{TV}\left(q_{t_0}\bar{P}^{t_0,N_{\text{pred}}}_{\text{ODE}}\bar{P}^{N_{\text{corr}}}_{\text{LMC}}, q_{t_0+T_{\text{pred}}}\right) \\
& \le \text{TV}\left(p, q_{t_0}\right) + \text{TV}\left(q_{t_0}\bar{P}^{t_0,N_{\text{pred}}}_{\text{ODE}}\bar{P}^{N_{\text{corr}}}_{\text{LMC}}, q_{t_0+T_{\text{pred}}}\right)
\end{aligned}$$

For the last term, by using Lemma Lemma 5 and Lemma 7, we know that

$$\begin{aligned}
& \text{TV}\left(q_{t_0}\bar{P}^{t_0,N_{\text{pred}}}_{\text{ODE}}\bar{P}^{N_{\text{corr}}}_{\text{LMC}}, q_{t_0+T_{\text{pred}}}\right) \\
& \lesssim \left(\frac{L_{t_0+T_{\text{pred}}}T^{1/2}_{\text{corr}}}{L^{1/4}_{\max}} + L^{1/2}_{\max}\right)W_2(q_{t_0}\bar{P}^{t_0,N_{\text{pred}}}_{\text{ODE}}, q_{t_0+T_{\text{pred}}}) + \frac{L_{t_0+T_{\text{pred}}}T^{1/2}_{\text{corr}}\sqrt{d}}{L^{1/4}_{\max}}h_{\text{corr}} + \frac{T^{1/2}_{\text{corr}}\epsilon_{\text{score}}}{L^{1/4}_{\max}\sigma_{T-t_0-T_{\text{pred}}}} \\
& \lesssim L^{1/2}_{\max}W_2(q_{t_0}\bar{P}^{t_0,N_{\text{pred}}}_{\text{ODE}}, q_{t_0+T_{\text{pred}}}) + \frac{L_{t_0+T_{\text{pred}}}\sqrt{d}}{L^{1/2}_{\max}}h_{\text{corr}} + \frac{\epsilon_{\text{score}}}{L^{1/2}_{\max}\sigma_{T-t_0-T_{\text{pred}}}}
\end{aligned}$$

For the Wasserstein distance, by Lemma 3:

$$
\begin{aligned}
&W_2(q_{t_0}\bar{P}_{\mathrm{ODE}}^{t_0,N_{\mathrm{pred}}}, q_{t_0+T_{\mathrm{pred}}}) \\
&\lesssim h_{\mathrm{pred}} R^2(2R+\sqrt{d})\frac{g(T-t_0-T_{\mathrm{pred}})^4}{\sigma^6_{T-t_0-T_{\mathrm{pred}}}L_{\max}} + \frac{g\left(T-t_0-T_{\mathrm{pred}}\right)^2\epsilon_{\mathrm{score}}}{L_{\max}\sigma_{T-t_0-T_{\mathrm{pred}}}}\,.
\end{aligned}
$$

■

Theorem 3. *[Uniform Stepsize] Assume* **Assumption 1**, *2 and 4 hold. Let $p_{T-\delta}^{\mathrm{ULMC}}$ be the output of the uniform version algorithm. Then, if $\sigma_t^2 = t^2$ (VESDE (SOTA)), $\mathrm{TV}(p_{T-\delta}^{\mathrm{ULMC}}, q_\delta)$ is bounded by*

$$
\frac{\bar{D}}{T} + \frac{\sqrt{d}R^3 h_{\mathrm{corr}}}{\delta^5} + \frac{R^5(R+\sqrt{d})h_{\mathrm{pred}}}{\delta^5} + \frac{R\epsilon_{\mathrm{score}}T}{\delta^2}\,.
$$

Furthermore, for VESDE (SOTA), by ignoring the approximated score error and choosing $T \geq \bar{D}/\epsilon_{\mathrm{TV}}, \delta = \epsilon_{W_2}/\sqrt{d}, h_{\mathrm{corr}} \leq \delta^5\epsilon_{\mathrm{TV}}/(\sqrt{d}R^3), h_{\mathrm{pred}} \leq \delta^5\epsilon_{\mathrm{TV}}/(R^5(R+\sqrt{d}))$, the output is ϵ_{TV}-close to q_δ, which is ϵ_{W_2}-close to q_0, with total iteration complexity

$$
\max\left\{\frac{T}{h_{\mathrm{pred}}}, \frac{T\sqrt{L_{\max}}}{h_{\mathrm{corr}}}\right\} \leq \tilde{O}\left(\frac{\bar{D}R^5 d^{3.5}(R+\sqrt{d})}{\epsilon_{\mathrm{TV}}^2\epsilon_{W_2}^7}\right)\,.
$$

For VESDE (SMLD) setting ($\sigma_t^2 = t$), the total iteration complexity is $\tilde{O}\left(\frac{\bar{D}^2R^5d^3(R+\sqrt{d})}{\epsilon_{\mathrm{TV}}^3\epsilon_{W_2}^6}\right)$.

Proof. For the setting $\sigma_t^2 = t$ and $g(t) = 1, \forall t \in [\delta, T]$, the reverse beginning term has the following upper bound (Theorem 20):

$$
\mathrm{TV}(q_T, q_\infty) \leq \frac{d|c| + \mathbb{E}[q_0] + R}{\sqrt{T}}\,.
$$

The remaining term is the discretization term. Since this term is analyzed under the middle step, we assume that $T = \frac{K_0}{L_{\max}} + \delta$. Then we have

$$
\begin{aligned}
&\mathrm{TV}\left(p_{T-\delta}^{\mathrm{ULMC}}, q_\delta\right) \\
&\leq \mathrm{TV}\left(q_T, q_\infty\right) \\
&+ O\Bigg(\sum_{k=1}^{K_0=(T-\delta)L_{\max}}\Bigg(\frac{R^2(2R+\sqrt{d})g(T-\frac{k}{L_{\max}}-T_{\mathrm{pred}})^4}{\sigma^6_{T-\frac{k}{L_{\max}}-T_{\mathrm{pred}}}L_{\max}^{1/2}}h_{\mathrm{pred}} + \frac{L_{\frac{k}{L_{\max}}+T_{\mathrm{pred}}}\sqrt{d}}{L_{\max}^{1/2}}h_{\mathrm{corr}} \\
&\quad + \frac{\epsilon_{\mathrm{score}}}{L_{\max}^{1/2}\sigma_{T-\frac{k}{L_{\max}}-T_{\mathrm{pred}}}} + \frac{g\left(T-\frac{k}{L_{\max}}-T_{\mathrm{pred}}\right)^2\epsilon_{\mathrm{score}}}{L_{\max}^{\frac{1}{2}}\sigma_{T-\frac{k}{L_{\max}}-T_{\mathrm{pred}}}}\Bigg)\Bigg)\,.
\end{aligned}
$$

For the corrector stage, we have that

$$
\begin{aligned}
\sum_{k=1}^{K_0=(T-\delta)L_{\max}}\frac{L_{\frac{k}{L_{\max}}+T_{\mathrm{pred}}}\sqrt{d}}{L_{\max}^{1/2}}h_{\mathrm{corr}} &\lesssim \frac{\sqrt{d}(1+R^2)h_{\mathrm{corr}}}{L_{\max}^{1/2}}\sum_{k=1}^{K_0=(T-\delta)L_{\max}}\frac{1}{(T-\frac{k}{L_{\max}}-T_{\mathrm{pred}})^2} \\
&\lesssim \frac{\sqrt{d}(1+R^2)L_{\max}^{3/2}h_{\mathrm{corr}}}{\delta L_{\max}} \\
&\lesssim \frac{\sqrt{d}R^3h_{\mathrm{corr}}}{\delta^2}\,.
\end{aligned}
$$

For the predictor stage, we have that

$$\begin{aligned}
&\sum_{k=1}^{K_0=(T-\delta)L_{\max}} \frac{R^2(2R+\sqrt{d})}{\sigma^6_{T-k/L_{\max}-T_{\text{pred}}} L_{\max}^{1/2}} h_{\text{pred}} \\
&\lesssim \frac{R^2(1+R^2)(2R+\sqrt{d})h_{\text{pred}}}{L_{\max}^{1/2}} \sum_{k=1}^{K=(T-\delta)L_{\max}} \frac{1}{(T-\frac{k}{L_{\max}}-T_{\text{pred}})^3} \\
&\lesssim \frac{R^2(1+R^2)(2R+\sqrt{d})h_{\text{pred}} L_{\max}^{1/2}}{\delta^2} \lesssim \frac{R^2(1+R^2)^{3/2}(2R+\sqrt{d})h_{\text{pred}}}{\delta^3}
\end{aligned}$$

For the approximated score error, we have that

$$\sum_{k=1}^{K_0=(T-\delta)L_{\max}} \frac{\epsilon_{\text{score}}}{L_{\max}^{1/2}\sigma_{T-\frac{k}{L_{\max}}-T_{\text{pred}}}} \lesssim \frac{\epsilon_{\text{score}}}{L_{\max}^{1/2}} \sum_{k=1}^{K_0=(T-\delta)L_{\max}} \frac{1}{\sqrt{T-\frac{k}{L_{\max}}-T_{\text{pred}}}} \lesssim \epsilon_{\text{score}}\sqrt{T}\,.$$

For the setting $\sigma_t^2 = t^2$ and $g(t) = \sqrt{2t}, \forall t \in [\delta, T]$, the reverse beginning term has the following upper bound:

$$\text{TV}(q_T, q_\infty) \le \frac{d|c| + \mathbb{E}[q_0] + R}{T}\,.$$

For the discretization error term, we know that

$$\begin{aligned}
\sum_{k=1}^{K_0=(T-\delta)L_{\max}} \frac{L_{\frac{k}{L_{\max}}+T_{\text{pred}}}\sqrt{d}}{L_{\max}^{1/2}} h_{\text{corr}} &\lesssim \frac{\sqrt{d}(1+R^2)h_{\text{corr}}}{L_{\max}^{1/2}} \sum_{k=1}^{K_0=(T-\delta)L_{\max}} \frac{1}{(T-\frac{k}{L_{\max}}-T_{\text{pred}})^4} \\
&\lesssim \frac{\sqrt{d}(1+R^2)L_{\max}^{1/2}h_{\text{corr}}}{\delta^3} \lesssim \frac{\sqrt{d}R^3 h_{\text{corr}}}{\delta^5}\,.
\end{aligned}$$

$$\begin{aligned}
&\sum_{k=1}^{K_0=(T-\delta)L_{\max}} \frac{R^2(2R+\sqrt{d})g(T-k/L_{\max}-T_{\text{pred}})^4}{\sigma^6_{T-k/L_{\max}-T_{\text{pred}}} L_{\max}^{1/2}} h_{\text{pred}} \\
&\lesssim \frac{R^2\left(1+R^2\right)(2R+\sqrt{d})h_{\text{pred}}}{L_{\max}^{1/2}} \sum_{k=1}^{K_0=(T-\delta)L_{\max}} \frac{1}{(T-\frac{k}{L_{\max}}-T_{\text{pred}})^4} \\
&\lesssim \frac{R^5(R+\sqrt{d})h_{\text{pred}}}{\delta^5}
\end{aligned}$$

For the approximated score error, we have that

$$\begin{aligned}
\sum_{k=1}^{K_0=(T-\delta)L_{\max}} \frac{\epsilon_{\text{score}}}{L_{\max}^{1/2}\sigma_{T-\frac{k}{L_{\max}}-T_{\text{pred}}}} &\lesssim \frac{\epsilon_{\text{score}}}{L_{\max}^{1/2}} \sum_{k=1}^{K_0=(T-\delta)L_{\max}} \frac{1}{T-\frac{k}{L_{\max}}-T_{\text{pred}}} \\
&\lesssim \epsilon_{\text{score}} L_{\max}^{1/2} \log(\frac{1}{\delta L_{\max}}) = \frac{R\epsilon_{\text{score}}}{\delta^2}\log(\frac{R^2}{\delta^5})\,.
\end{aligned}$$

And,

$$\sum_{k=1}^{K_0=(T-\delta)L_{\max}} \frac{g\left(T-\frac{k}{L_{\max}}-T_{\text{pred}}\right)^2 \epsilon_{\text{score}}}{L_{\max}^{\frac{1}{2}}\sigma_{T-\frac{k}{L_{\max}}-T_{\text{pred}}}} \lesssim \epsilon_{\text{score}} L_{\max}^{1/2} T = \frac{R\epsilon_{\text{score}}T}{\delta^2}\,.$$

■

E The Proof For Exponential-decay PFODE-Corrector Algorithm

In this section, we provide the proof of Theorem 4, which uses exponential decay predictor stepsize and uniform corrector stepsize.

E.1 The Predictor Step with exponential decay predictor stepsize

In the section, we first show that when starting at the same distribution q at time t_0, the W_2 distance for reverse PFODE can be controlled under a suitable step. Note that in the following lemma, we do not require the $h_k \leq L_{\max} = \frac{1+R^2}{\sigma_\delta^4}$ (which is used in VPSDE Chen et al. (2023c)) to control one step predictor error since VESDE do not have additional drift term in the reverse PFODE. The requirement of h_k is introduced by the discretization error Lemma 15.

Lemma 9. *Suppose* **Assumption 1** *and* **Assumption 2** *hold and assuming* $h_k := t_k - t_{k-1} = rt_k$, *where* $r \leq 1/2d$, *then for the small interval* $t \in [t'_k, t'_{k+1}]$ *for* $\forall k \in [0, K-1]$, *we have that*
(1) For VESDE (SMLD)

$$W_2\left(qQ_{\mathrm{ODE}}^{t'_k,h'_k}, q\bar{P}_{\mathrm{ODE}}^{t'_k,h'_k}\right) \lesssim \frac{dh_k'^{1.5}}{T-t'_k} + \frac{h'_k\epsilon_{\mathrm{score}}}{\sqrt{T-t'_k}}\,.$$

(2) For VESDE (SOTA)

$$W_2\left(qQ_{\mathrm{ODE}}^{t'_k,h'_k}, q\bar{P}_{\mathrm{ODE}}^{t'_k,h'_k}\right) \lesssim \frac{dh_k'^{1.5}}{\sqrt{T-t'_k}} + h'_k\epsilon_{\mathrm{score}}\,.$$

Proof. For $t \in [t'_k, t'_{k+1}]$, the reverse PFODE is

$$\begin{aligned}
\dot{Y}_t &= \frac{g(T-t)^2}{2}\nabla \ln q_{T-t}\left(Y_t\right)\,,\\
\dot{\bar{Y}}_t &= \frac{g(T-t)^2}{2}s_{T-t'_k}\left(\bar{Y}_{t'_k}\right)\,,
\end{aligned}$$

for $t'_k \leq t \leq t'_{k+1}$, with $Y_{t'_k} = \bar{Y}_{t'_k} \sim q, Y_{t'_k+h'_k} \sim qQ_{\mathrm{ODE}}$, and $\bar{Y}_{t'_k+h'_k} \sim q\bar{P}_{\mathrm{ODE}}$. Then, we have that

$$\begin{aligned}
\partial_t\left\|Y_t - \bar{Y}_t\right\|^2 &= 2\left\langle Y_t - \bar{Y}_t, \dot{Y}_t - \dot{\bar{Y}}_t\right\rangle\\
&= 2\left\langle Y_t - \bar{Y}_t, \frac{g(T-t)^2}{2}\left(\nabla \ln q_{T-t}\left(Y_t\right) + s_{T-t'_k}\left(\bar{Y}_{t'_k}\right)\right)\right\rangle\\
&\leq \frac{1}{h'_k}\left\|Y_t - \bar{Y}_t\right\|^2 + \frac{h'_k g(T-t)^4}{4}\left\|\nabla \ln q_{T-t}\left(Y_t\right) - s_{T-t'_k}\left(\bar{Y}_{t'_k}\right)\right\|^2
\end{aligned}$$

By Grönwall's inequality, we have that

$$\begin{aligned}
&\mathbb{E}\left[\left\|Y_{t'_k+h'_k} - \bar{Y}_{t'_k+h'_k}\right\|^2\right]\\
&\leq \int_{t'_k}^{t'_k+h'_k} g(T-t)^4 h'_k\mathbb{E}\left[\left\|\nabla \ln q_{T-t}\left(Y_t\right) - s_{T-t'_k}\left(\bar{Y}_{t'_k}\right)\right\|^2\right]\mathrm{d}t\\
&\lesssim \int_{t'_k}^{t'_k+h'_k} \frac{d^2 g(T-t)^4 h'_k(\sigma^2_{T-t'_k} - \sigma_t^2)}{\sigma^4_{T-t}} + \frac{h'_k g(T-t)^4\epsilon^2_{\mathrm{score}}}{\sigma^2_{T-t}}\mathrm{d}t\,,
\end{aligned}$$

where the last inequality follows Lemma 15. Then, for VESDE (SMLD), we have that

$$\mathbb{E}\left[\left\|Y_{t'_k+h'_k} - \bar{Y}_{t'_k+h'_k}\right\|^2\right] \lesssim \frac{d^2 h_k'^3}{(T-t'_k)^2} + \frac{h_k'^2\epsilon^2_{\mathrm{score}}}{T-t'_k}\,.$$

For VESDE (SOTA), we have that

$$\mathbb{E}\left[\left\|Y_{t'_k+h'_k} - \bar{Y}_{t'_k+h'_k}\right\|^2\right] \lesssim \frac{d^2 h_k'^3}{T-t'_k} + h_k'^2\epsilon^2_{\mathrm{score}}\,.$$

■

E.2 The Corresponding Corrector Step

Similar to Appendix D.2, we inject noise into the process using the ULD corrector. The proof process is almost the same compared to Appendix D.2, except the exponential decay predictor stepsizes. We provide the proof of this part for completeness.

Different from Appendix D.1, the exponential decay stepsize predictor process starts at the same distribution $q_{t'_k}$, we run continuous process and discretization process for one predictor step h'_k instead of run uniform h_{pred} for time T_{pred}. We define the probability measures $p = q_{t'_k}\bar{P}_{\text{ODE}}^{t'_k,h'_k}$ and $q = q_{t'_k}Q_{\text{ODE}}^{t'_k,h'_k}$. We also write $\boldsymbol{p} := p \otimes \gamma_d$ and $\boldsymbol{q} := q \otimes \gamma_d$, where γ_d is the standard Gaussian measure in $\mathbb{R}^d$.

For the corrector stage, we set the friction parameter to $\rho \asymp \sqrt{L'_{t_{k+1}}}$ to match the exponential decay predictor stepsize. For the corrector stepsize, we still use the uniform stepsize and define $N_{\text{corr}} = T_{\text{corr}}/h_{\text{corr}}$ which is the same compared to Appendix D.2. Similar to Appendix D.2, we need to control $\text{TV}\left(\boldsymbol{p}P_{\text{ULD}}^{t'_{k+1},N_{\text{corr}}}, \boldsymbol{q}\right)$ and $\text{TV}\left(\boldsymbol{p}\bar{P}_{\text{ULMC}}^{t'_{k+1},N_{\text{corr}}}, \boldsymbol{p}P_{\text{ULD}}^{t'_{k+1},N_{\text{corr}}}\right)$. For these two terms, we just need to replace $t_0 + T_{\text{pred}}$ to obtain the final results.

The following lemma is a modification of Lemma 5, Lemma 6 and Lemma Lemma 7. The modified idea is that we only need local maximum Lipschitz constant at time t'_k since the corrector does not change the reverse timelin. Hence, we only need Lipschitz constant $L_{t'_k}$ at time t'_k.

Lemma 10. *If $T_{\text{corr}} \lesssim 1/\sqrt{L_{t'_{k+1}}}$, then*

$$\text{TV}\left(\boldsymbol{p}P_{\text{ULD}}^{t'_{k+1},N_{\text{corr}}}, \boldsymbol{q}\right) \lesssim \sqrt{\text{KL}\left(\boldsymbol{p}P_{\text{ULD}}^{t'_{k+1},N_{\text{corr}}} \| \boldsymbol{q}\right)} \lesssim \frac{W_2(p,q)}{L_{t'_{k+1}}^{1/4} T_{\text{corr}}^{3/2}},$$

$$\text{TV}\left(\boldsymbol{p}\bar{P}_{\text{ULMC}}^{t'_{k+1},N_{\text{corr}}}, \boldsymbol{p}P_{\text{ULD}}^{t'_{k+1},N_{\text{corr}}}\right) \lesssim \frac{L_{t'_{k+1}} T_{\text{corr}}^{1/2}}{L_{t'_{k+1}}^{1/4}} W_2(p,q) + \frac{L_{t'_{k+1}} T_{\text{corr}}^{1/2}\sqrt{d}}{L_{t'_{k+1}}^{1/4}} h_{\text{corr}} + \frac{T_{\text{corr}}^{1/2}\epsilon_{\text{score}}}{L_{t'_{k+1}}^{1/4}\sigma_{T-t'_{k+1}}}.$$

Similar to Lemma 8, after analyzing the predictor and corrector, we also provide the error guarantee after a step for predictor h'_k and a middle corrector step T_{corr}.

Lemma 11. *Let $L_t = \left(1+R^2\right)/\sigma_{T-t}^4$, the stationary distribution $q_{t'_k}Q_{\text{ODE}}^{t'_k,h'_k} = q_{t'_{k+1}}$ and $h_k := t_k - t_{k-1} = rt_k$, where $r \le 1/(2d)$. For the underdamped Langenvin, $T_{\text{corr}} \le 1/\sqrt{L_{t'_{k+1}}}$ and $N_{\text{corr}} = T_{\text{corr}}/h_{\text{corr}}$. Then, for VESDE (SMLD), we have that*

$$\begin{aligned}
&\text{TV}\left(p\bar{P}_{\text{ODE}}^{t'_k,h'_k}\bar{P}_{\text{ULMC}}^{N_{\text{corr}}}, q_{t'_{k+1}}\right)\\
&\le \text{TV}\left(p, q_{t'_k}\right) + O\left(\frac{dRh_k'^{1.5}}{(T-t'_k)^2} + \frac{R\sqrt{d}}{T-t'_k}h_{\text{corr}} + \frac{\sigma_{T-t'_k}\epsilon_{\text{score}}}{R} + \frac{R\epsilon_{\text{score}}h'_k}{(T-t'_k)^{1.5}}\right)\\
&\le \text{TV}\left(p, q_{t'_k}\right) + O\left(\frac{dRh_k'^{1.5}}{\sqrt{\delta}(T-t'_k)^{1.5}} + \frac{R\sqrt{d}}{\delta}h_{\text{corr}} + \frac{\sqrt{T}\epsilon_{\text{score}}}{R} + \frac{R\epsilon_{\text{score}}h'_k}{\sqrt{\delta}(T-t'_k)}\right)
\end{aligned}$$

For VESDE (SOTA), we have that

$$\begin{aligned}
&\text{TV}\left(p\bar{P}_{\text{ODE}}^{t'_k,h'_k}\bar{P}_{\text{ULMC}}^{N_{\text{corr}}}, q_{t'_{k+1}}\right)\\
&\le \text{TV}\left(p, q_{t'_k}\right) + O\left(\frac{dRh_k'^{1.5}}{\delta(T-t'_k)^{1.5}} + \frac{R\sqrt{d}}{\delta^2}h_{\text{corr}} + \frac{T\epsilon_{\text{score}}}{R} + \frac{R\epsilon_{\text{score}}h'_k}{\delta(T-t'_k)}\right).
\end{aligned}$$

Proof. After the corrector stage, we can use data-processing inequality.

$$\begin{aligned}
&\mathrm{TV}\left(p\bar{P}_{\mathrm{ODE}}^{t'_k,h'_k}\bar{P}_{\mathrm{LMC}}^{N_{\mathrm{corr}}}, q_{t'_k}\right)\\
&\leq \mathrm{TV}\left(p\bar{P}_{\mathrm{ODE}}^{t'_k,h'_k}\bar{P}_{\mathrm{LMC}}^{N_{\mathrm{corr}}}, q_{t'_k}\bar{P}_{\mathrm{ODE}}^{t'_k,h'_k}\bar{P}_{\mathrm{LMC}}^{N_{\mathrm{corr}}}\right) + \mathrm{TV}\left(q_{t'_k}\bar{P}_{\mathrm{ODE}}^{t'_k,h'_k}\bar{P}_{\mathrm{LMC}}^{N_{\mathrm{corr}}}, q_{t'_k+h'_k}\right)\\
&\leq \mathrm{TV}\left(p, q_{t'_k}\right) + \mathrm{TV}\left(q_{t'_k}\bar{P}_{\mathrm{ODE}}^{t'_k,h'_k}\bar{P}_{\mathrm{LMC}}^{N_{\mathrm{corr}}}, q_{t'_k+h'_k}\right).
\end{aligned}$$

For the last term, by using Lemma 10, we know that

$$\mathrm{TV}\left(q_{t'_k}\bar{P}_{\mathrm{ODE}}^{t'_k,h'_k}\bar{P}_{\mathrm{LMC}}^{N_{\mathrm{corr}}}, q_{t'_k+h'_k}\right) \lesssim L_{t'_k}^{1/2} W_2(q_{t'_k}\bar{P}_{\mathrm{ODE}}^{t'_k,h'_k}, q_{t'_{k+1}}) + \sqrt{dL_{t'_k}}h_{\mathrm{corr}} + \frac{\epsilon_{\mathrm{score}}}{L_{t'_k}^{1/2}\sigma_{T-t'_k}}.$$

For the Wasserstein distance, we use Lemma 9 to obtain the final results. ■

Theorem 4. *[Exponential-decay stepsize] Following the setting of Theorem 3, choosing the exponential-decay version algorithm with* $r = \frac{\epsilon_{\mathrm{TV}}^2\sigma_\delta^2}{d^2R^2\log^3(T/\delta)}$, $\bar{D}/\sigma_T \leq \epsilon_{\mathrm{TV}}$, $\sigma_\delta^2 = \epsilon_{W_2}^2/d$ *and considering VESDE (SOTA), the iteration complexity for the predictor is*

$$O\Big(\frac{d^3R^2\log^3(T/\delta)}{\epsilon_{\mathrm{TV}}^2\epsilon_{W_2}^2}\Big),$$

and the iteration complexity for the corrector is $O\Big(\frac{d^{7.5}R^5\bar{D}^2\log^6(T/\delta)}{\epsilon_{\mathrm{TV}}^7\epsilon_{W_2}^6}\Big)$. *The corrector iteration complexity dominates the total iteration complexity. We achieve the same result for VESDE (SMLD).*

Proof. For the setting $\sigma_t^2 = t$ and $g(t) = 1, \forall t \in [\delta, T]$, the reverse beginning term has the following upper bound (Theorem 20):

$$\mathrm{TV}(q_T, q_\infty) \leq \frac{d|c| + \mathbb{E}[q_0] + R}{\sqrt{T}}.$$

Hence, we need $T \geq \frac{(d|c|+\mathbb{E}[q_0]+R)^2}{\epsilon_{TV}^2}$. The remaining term is the discretization term. As the exponential decay stepsize, the iteration complexity of the predictor K (the number of the middle corrector stepsize) is $\frac{1}{c}\log(T/\delta)$. Then we have that

$$\begin{aligned}
&\mathrm{TV}\left(p_{T-\delta}^{ULMC}, q_\delta\right)\\
&\leq \mathrm{TV}\left(q_T, q_\infty\right) + O\Bigg(\sum_{k=1}^{K=\frac{1}{c}\log(T/\delta)}\left(\frac{dRh_k^{1.5}}{\sqrt{\delta}t_k^{1.5}} + \frac{R\sqrt{d}}{\delta}h_{\mathrm{corr}} + \frac{\sqrt{\delta}\epsilon_{\mathrm{score}}}{R} + \frac{R\epsilon_{\mathrm{score}}h_k}{\sqrt{\delta}t_k}\right)\Bigg).
\end{aligned}$$

We also know that

$$\sum_{k=1}^{K=\frac{1}{c}\log(T/\delta)}\frac{dRh_k^{1.5}}{\sqrt{\delta}t_k^{1.5}} \lesssim \frac{dR\log^{1.5}(T/\delta)}{\sqrt{\delta K}}.$$

Hence, we need $K \geq \frac{d^2R^2\log^3(T/\delta)}{\epsilon_{TV}^2\delta}$ to guarantee the error smaller than ϵ_{TV}. For the corrector term, we have that

$$\sum_{k=1}^{K}\frac{R\sqrt{d}}{\delta}h_{\mathrm{corr}} \leq \frac{d^{2.5}R^3\log^3(T/\delta)}{\epsilon_{TV}^2\delta^2}h_{\mathrm{corr}}.$$

Hence, we need $h_{\mathrm{corr}} \leq \frac{\epsilon_{TV}^3\delta^2}{d^{2.5}R^3\log^3(T/\delta)}$ to satisfy accuracy guarantee. For the approximated error term, we have that

$$\sum_{k=1}^{K}\frac{\sqrt{T}\epsilon_{\mathrm{score}}}{R} + \frac{R\epsilon_{\mathrm{score}}h_k}{\sqrt{\delta}t_k} \leq \frac{\bar{D}d^2R\log^3(T/\delta)\epsilon_{\mathrm{score}}}{\epsilon_{TV}^3\delta} + \frac{R\log(T/\delta)\epsilon_{\mathrm{score}}}{\sqrt{\delta}}.$$

Hence, we need to assume $\epsilon_{\text{score}} \leq \frac{\epsilon_{TV}^4 \delta}{\bar{D} d^2 R \log^3(T/\delta)}$. Then, the iteration complexity for the predictor is

$$K = \frac{d^3 R^2 \log^3(T/\delta)}{\epsilon_{TV}^2 \epsilon_{W_2}^2}.$$

The iteration complexity for the corrector is

$$\frac{K}{\sqrt{L_0} h_{corr}} = \frac{KT}{h_{corr}} = \frac{d^{7.5} R^5 (d|c| + \mathbb{E}[q_0] + R)^2 \log^6(T/\delta)}{\epsilon_{TV}^7 \epsilon_{W_2}^6}$$

For the setting $\sigma_t^2 = t^2$ and $g(t) = \sqrt{2t}$, the reverse beginning term has the following upper bound (Lemma 20):

$$\text{TV}(q_T, q_\infty) \leq \frac{d|c| + \mathbb{E}[q_0] + R}{T}.$$

Hence, we need $T \geq \frac{d|c| + \mathbb{E}[q_0] + R}{\epsilon_{TV}}$. The rest proof process is exactly the same compared to VESDE (SMLD).

$$\sum_{k=1}^{K = \frac{1}{c}\log(T/\delta)} \frac{dR h_k^{1.5}}{\delta t_k^{1.5}} \lesssim \frac{dR \log^{1.5}(T/\delta)}{\delta \sqrt{K}}.$$

We need $K \geq \frac{d^2 R^2 \log^3(T/\delta)}{\epsilon_{TV}^2 \delta^2}$, $h_{\text{corr}} \leq \frac{\epsilon_{TV}^3 \delta^4}{d^{2.5} R^3 \log^3(T/\delta)}$ and assume $\epsilon_{\text{score}} \leq \frac{\epsilon_{TV}^4 \delta^2}{\bar{D} d^2 R \log^3(T/\delta)}$.

■

E.3 The Discussion on the Approximated Score

We note that the dependence of T and $1/\delta$ in the approximated score term is polynomial in Theorem 3 and Theorem 4. Hence, if ϵ_{score} is much smaller than ϵ_{TV} and ϵ_{W_2}, we can consider the approximated score. Using VESDE (SOTA) as an example, we require $\epsilon_{\text{score}} \leq \epsilon_{\text{TV}}^2 \epsilon_{W_2}^2 / (R\bar{D}d)$ for Theorem 3 and $\epsilon_{\text{score}} \leq \epsilon_{\text{TV}}^4 \epsilon_{W_2}^2 / (\bar{D} d^3 R \log^3(T/\delta))$ for Theorem 4. These requirements are common when considering the VESDE or PFODE-Corrector algorithm. More specifically, Lee et al. (2022) assume $\epsilon_{\text{score}} \leq \widetilde{O}(\epsilon_{\text{TV}}^3)$ for VESDE (SMLD) [5]. The DPUM algorithm requires $\epsilon_{\text{score}} \leq \epsilon_{\text{TV}} \epsilon_{W_2}^4 / R$.

F Auxiliary Lemmas

F.1 Lemmas for the Ground Truth Score Function

In this section, we first show the uniform bound for the ground truth score function used in the PFODE with corrector setting (Lemma 12 and Lemma 13). Lemma 12 comes form Lemma C.1 and C.2 of De Bortoli (2022) and Lemma 13 modifies Lemma C.3 by using the property of VESDE. Then, we provide a more refined control on the Hessian matrix $\left\|\nabla^2 \log q_\sigma(x)\right\|_{F,\psi_1}$, which is proposed by Chen et al. (2023a) and is used to control the discretization error in our reverse SDE setting and exponential-deacy PFODE-Corrector algorithm analysis.

Lemma 12. *Assume* **Assumption 1**. *Then for any $t \in (0, T]$ and $x_t \in \mathbb{R}^d$ we have that*

$$\langle \nabla \log q_t(x_t), x_t \rangle \leq -\|x_t\|^2/\sigma_t^2 + R\|x_t\|/\sigma_t^2.$$

In addition, we have

$$\|\nabla \log q_t(x_t)\|^2 \leq 2\|x_t\|^2/\sigma_t^4 + 2R^2/\sigma_t^4,$$

and

$$\left\|\nabla^2 \log q_t(x_t)\right\| \leq \left(1 + R^2\right)/\sigma_t^4.$$

[5] The absence of the early stopping parameter δ in Lee et al. (2022) is due to the LSI assumption on the data.

The following lemma is the score perturbation lemma, which is used to control the $\|\partial_t \nabla \log q_t(x_t)\|$.

Lemma 13. *Assume* **Assumption 1**. *Then for any $t \in (0,T]$ and $x_t \in \mathbb{R}^d$ we have*

$$\|\partial_t \nabla \log q_t(x_t)\| \leq \frac{g(t)^2}{\sigma_t^6} R^2 (R + \|x_t\|).$$

Proof. Let $N \in \mathbb{N}$ and $t \in (0,T]$. We denote for any $x \in \mathbb{R}^d$, $q_t^N(x) = \bar{q}_t^N(x) / \left(2\pi\sigma_t^2\right)^{d/2}$ with

$$\bar{q}_t^N(x) = (1/N) \sum_{k=1}^N e_t^k(x), \qquad e_t^k(x) = \exp\left[-\|x - X^k\|^2 / \left(2\sigma_t^2\right)\right].$$

Next we denote $f_t^k \triangleq \log e_t^k$. Then we have

$$\partial_t \log \bar{q}_t^N(x) \sum_{k=1}^N \partial_t f_t^k(x) e_t^k(x) / \sum_{k=1}^N e_t^k(x).$$

Therefore we have

$$\begin{aligned}
\partial_t \nabla \log \bar{q}_t^N(x) &= \sum_{k=1}^N \partial_t \nabla f_t^k(x) e_t^k(x) / \sum_{k=1}^N e_t^k(x) + \sum_{k=1}^N \partial_t f_t^k(x) \nabla f_t^k(x) e_t^k(x) / \sum_{k=1}^N e_t^k(x) \\
&\quad - \sum_{k,j=1}^N \partial_t f_t^k(x) \nabla f_t^j(x) e_t^k(x) e_t^j(x) / \sum_{k,j=1}^N e_t^k(x) e_t^j(x) \\
&= \frac{1}{2} \sum_{k,j=1}^N (\partial_t f_t^k(x) - \partial_t f_t^j(x))(\nabla f_t^k(x) - \nabla f_t^j(x)) e_t^k(x) e_t^j(x) / \sum_{k,j=1}^N e_t^k(x) e_t^j(x) \\
&\quad + \sum_{k=1}^N \partial_t \nabla f_t^k(x) e_t^k(x) / \sum_{k=1}^N e_t^k(x).
\end{aligned}$$

In what follows, we provide upper bounds for $|\partial_t f_t^k - \partial_t f_t^j|, \|\nabla f_t^k - \nabla f_t^j\|$ and $\partial_t \nabla f_t^k$. First we notice that $\nabla f_t^k(x) = -\left(x - X^k\right)/\sigma_t^2$, and using $m_t \leq 1$ we get

$$\|\nabla f_t^k(x) - \nabla f_t^j(x)\| \leq R/\sigma_t^2.$$

and

$$\partial_t f_t^k(x) = \partial_t \sigma_t^2 / \left(2\sigma_t^4\right) \|x - X^k\|^2. \tag{7}$$

Notice the fact that $\partial_t \sigma_t^2 = g(t)^2$, combined with Eq. (7), we know that

$$\partial_t f_t^k(x) = \frac{g(t)^2}{2\sigma_t^4} \|x - X^k\|^2 = \frac{g(t)^2}{2\sigma_t^4} \left(\|x\|^2 + \left\|X^k\right\|^2 - 2\left\langle x, X^k \right\rangle\right),$$

The rest of the proof is identical to the Lemma C.3 in De Bortoli (2022).

$$\left|\partial_t f_t^k(x) - \partial_t f_t^j(x)\right| \leq \frac{g(t)^2}{2\sigma_t^4} (R^2 + 2R\|x\|)$$

Now we compute $\nabla \partial_t f_t^k(x)$ for any $x \in \mathbb{R}^d$

$$\nabla \partial_t f_t^k(x) = \frac{g(t)^2}{\sigma_t^4} \|x - X^k\|.$$

So we can bound the norm of it by

$$\|\partial_t \nabla f_t^k(x)\| \leq \frac{g(t)^2}{\sigma_t^4} (R + \|x\|)$$

Combining results above we get for any $x \in \mathbb{R}^d$

$$\begin{aligned}\left\|\partial_t \nabla \log \bar{q}_t^N(x)\right\| &\leq \frac{g(t)^2}{\sigma_t^4}(R+\|x\|)+\frac{g(t)^2}{\sigma_t^6}R^2(R+\|x\|) \\ &\leq \frac{g(t)^2}{\sigma_t^6}R^2(R+\|x\|)\,.\end{aligned}$$

The last inequality is by the fact that the dominated time t for the iteration complexity is $t \to \delta$.

Note that

$$\lim_{N\to+\infty} \partial_t \nabla \log q_t^N(x_t) = \partial_t \nabla \log q_t$$

and the proof is complete. ■

In the rest of this section, similar to Chen et al. (2023a), we provide a more refined control on the Hessian Matrix $\nabla^2 \log q_t(X_t)$, where $X_t \sim q_t$ instead of the uniform bound. These auxiliary two lemmas are useful for the aggressive exponential decay stepsize. The following lemma is exactly the same compared to Chen et al. (2023a), and these lemmas can be used in the VESDE setting since these lemmas do not involve the specific process.

Lemma 14. *Let Q be a probability measure on $\mathbb{R}^d$. Consider the density its Gaussian perturbation $q_\sigma(x) \propto \int_{\mathbb{R}^d} \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \mathrm{d}Q(y)$. Then for $x \sim q_\sigma$, we have the sub-exponential norm bound*

$$\left\|\nabla^2 \log q_\sigma(x)\right\|_{F,\psi_1} \lesssim \frac{d}{\sigma^2}\,,$$

and

$$\left\|\nabla \log q_\sigma(x)\right\|_{\psi_2} \lesssim \sqrt{\frac{d}{\sigma^2}}\,.$$

where $\|\cdot\|_{F,\psi_1} = \left\|\|\cdot\|_F\right\|_{\psi_1}$ denote the sub-exponential norm of the Frobenius norm of a random matrix.

Then, we will show how to deal with the space discretization error for VESDE. The following lemma is almost identical compared to Lemma 13 of Chen et al. (2023a) (choosing $\alpha_{t,s} = 1$ since the VESDE setting) except the order of the diffusion (variance) term and the relationship between the variable stepsize and the variance term. For the sake of completeness, we also give the proof process of this part.

Lemma 15. *For $0 \leq t \leq s \leq T$, $\frac{\sigma_s^2-\sigma_t^2}{\sigma_t^2} \leq \frac{1}{2d}$ and the forward process Eq. (1), we have*

$$\mathbb{E}\left\|\nabla \log q_t(X_t) - \nabla \log q_t(X_s)\right\|^2 \lesssim \frac{d^2(\sigma_s^2-\sigma_t^2)}{\sigma_t^4}\,.$$

*When considering the low dimensional assumption (***Assumption 3***) and $\frac{\sigma_s^2-\sigma_t^2}{\sigma_t^2} \leq \frac{1}{2d'}$, we have that*

$$\mathbb{E}\left\|\nabla \log q_t(X_t) - \nabla \log q_t(X_s)\right\|^2 \lesssim \frac{(d'^3 + \left\|I_d - AA^\top\right\|_F^2)(\sigma_s^2-\sigma_t^2)}{\sigma_t^4}\,.$$

Proof. **The proof for the bounded support data.** To bound the above term by using the Hessian matrix, we have that

$$\begin{aligned}\nabla \log q_t(X_t) - \nabla \log q_t(X_s) &= \int_0^1 \nabla^2 \log q_t\left(X_t + a(X_s - X_t)\right)(X_s - X_t)\,\mathrm{d}a \\ \mathbb{E}\left\|\nabla \log q_t(X_t) - \nabla \log q_t(X_s)\right\|^2 &\leq \int_0^1 \mathbb{E}\left\|\nabla^2 \log q_t\left(X_t + a z_{t,s}\right) z_{t,s}\right\|^2 \,\mathrm{d}a\,,\end{aligned}$$

where $z_{t,s}$ is defined by $z_{t,s} = X_s - X_t \sim \mathcal{N}\left(0, (\sigma_s^2 - \sigma_t^2) I_d\right)$ and is independent of X_t. For random vectors X, Y, we use $P_{X,Y}$ to denote the joint probability measure of (X, Y) and $P_{X|Y}$ to denote the conditional probability measure of X given Y. Then for $0 \leq a \leq 1$, we use change of measure to bound $\mathbb{E}\left\|\nabla^2 \log q_t\left(X_t + a z_{t,s}\right) z_{t,s}\right\|^2$:

$$\begin{aligned}
\mathbb{E}\left\|\nabla^2 \log q_t\left(X_t + a z_{t,s}\right) z_{t,s}\right\|^2 &= \mathbb{E}\left[\left\|\nabla^2 \log q_t\left(X_t\right) z_{t,s}\right\|^2 \frac{\mathrm{d} P_{X_t + a z_{t,s}, z_{t,s}}\left(X_t, z_{t,s}\right)}{\mathrm{d} P_{X_t, z_{t,s}}\left(X_t, z_{t,s}\right)}\right] \\
&\lesssim \sqrt{\mathbb{E}\left\|\nabla^2 \log q_t\left(X_t\right) z_{t,s}\right\|^4 \mathbb{E}\left(\frac{\mathrm{d} P_{X_t + a z_{t,s}, z_{t,s}}\left(X_t, z_{t,s}\right)}{\mathrm{d} P_{X_t, z_{t,s}}\left(X_t, z_{t,s}\right)}\right)^2} .
\end{aligned}$$

Similar to Chen et al. (2023a), we define $M_t = \nabla^2 \log q_t\left(X_t\right)\left(\nabla^2 \log q_t\left(X_t\right)\right)^\top, Z_{t,s} = z_{t,s} z_{t,s}^\top$. For $A, B \in \mathbb{R}^{d \times d}$, define the tensor product $A \otimes B \in \left(\mathbb{R}^d\right)^{\otimes 4}$ as $(A \otimes B)_{i_1, i_2, i_3, i_4} = A_{i_1 i_2} B_{i_3 i_4}$. Since M_t and $Z_{t,s}$ are independent, then we can bound the two terms in the above inequality separately

$$\mathbb{E}\left\|\nabla^2 \log q_t\left(X_t\right) z_{t,s}\right\|^4 = \left\langle \mathbb{E} M_t \otimes M_t, \mathbb{E} Z_{t,s} \otimes Z_{t,s} \right\rangle .$$

Term $\mathbb{E} Z_{t,s} \otimes Z_{t,s}$ is purely determined by the diffusion (variance) term:

$$\mathbb{E}\left(Z_{t,s} \otimes Z_{t,s}\right)_{i_1, i_2, i_3, i_4} = \begin{cases} 3(\sigma_s^2 - \sigma_t^2)^2, & i_1 = i_2 = i_3 = i_4, \\ (\sigma_s^2 - \sigma_t^2)^2, & i_1 \neq i_2, (i_1, i_2) = (i_3, i_4) \text{ or } (i_1, i_2) = (i_4, i_3), \\ 0, \quad \text{else.} \end{cases}$$

Then, we can bound the inner product term by using exactly the same process compared to Chen et al. (2023a):

$$\begin{aligned}
\left\langle \mathbb{E} M_t \otimes M_t, \mathbb{E} Z_{t,s} \otimes Z_{t,s} \right\rangle &\lesssim \sigma_{s-t}^4 \left(\sum_{(i_1, i_2) = (i_3, i_4)} + \sum_{(i_1, i_2) = (i_4, i_3)} \right) \mathbb{E}\left(M_t \otimes M_t\right)_{i_1, i_2, i_3, i_4} \\
&\lesssim (\sigma_s^2 - \sigma_t^2)^2 \mathbb{E}\left\|\nabla^2 \log q_t\left(X_t\right)\right\|_F^4 \\
&\lesssim (\sigma_s^2 - \sigma_t^2)^2 \left(\frac{d}{\sigma_t^2}\right)^4
\end{aligned}$$

For the rest term, we have that

$$\begin{aligned}
\mathbb{E}\left(\frac{\mathrm{d} P_{X_t + a z_{t,s}, z_{t,s}}\left(X_t, z_{t,s}\right)}{\mathrm{d} P_{X_t, z_{t,s}}\left(X_t, z_{t,s}\right)}\right)^2 &= \mathbb{E}\left(\frac{\mathrm{d} P_{X_t + a z_{t,s} | z_{t,s}}\left(X_t | z_{t,s}\right)}{\mathrm{d} P_{X_t | z_{t,s}}\left(X_t | z_{t,s}\right)}\right)^2 \\
&\leq \mathbb{E}\left(\frac{\mathrm{d} P_{X_t + a z_{t,s} | z_{t,s}, x_0}\left(X_t | z_{t,s}, x_0\right)}{\mathrm{d} P_{X_t | x_0}\left(X_t | x_0\right)}\right)^2 .
\end{aligned}$$

We also know that $X_t + a z_{t,s} | \left(z_{t,s}, x_0\right) \sim \mathcal{N}\left(x_0 + a z_{t,s}, \sigma_t^2 I_d\right)$ and $X_t | x_0 \sim \mathcal{N}\left(x_0, \sigma_t^2 I_d\right)$. Then, the above term has the following equation by the chi-squared divergence explicity:

$$\mathbb{E}\left(\frac{\mathrm{d} P_{X_t + a z_{t,s} | z_{t,s}, x_0}\left(X_t | z_{t,s}, x_0\right)}{\mathrm{d} P_{X_t | x_0}\left(X_t | x_0\right)}\right)^2 = \mathbb{E} \exp\left(\frac{a^2 \left\|z_{t,s}\right\|^2}{\sigma_t^2}\right) .$$

Recall that we assume $\frac{\sigma_s^2 - \sigma_t^2}{\sigma_t^2} \leq \frac{1}{2d}$, we have that

$$\mathbb{E} \exp\left(\frac{a^2 \left\|z_{t,s}\right\|^2}{\sigma_t^2}\right) = \left(1 - 2\frac{a^2\left(\sigma_s^2 - \sigma_t^2\right)}{\sigma_t^2}\right)^{-d/2} \lesssim 1 ,$$

and

$$\mathbb{E}\left\|\nabla^2 \log q_t\left(X_t + a z_{t,s}\right) z_{t,s}\right\|^2 \lesssim \frac{d^2(\sigma_s^2 - \sigma_t^2)}{\sigma_t^4} .$$

The proof for the low dimensional data. In the rest of the proof, we show how the low-dimensional assumption (**Assumption 3**) affects the results. As mentioned in Appendix C.1, under **Assumption 3**, the score function has the following form

$$\nabla \log q_t(X) = A\nabla \log q_t^{\mathrm{LD}}\left(A^\top X\right) - \frac{1}{\sigma_t^2}\left(I_d - AA^\top\right)X\,,$$

and the diffusion process happens in the latent space. For our goal $\mathbb{E}\left\|\nabla \log q_t\left(X_t\right) - \nabla \log q_t\left(X_s\right)\right\|^2$, we know that

$$\mathbb{E}\left\|\nabla \log q_t\left(X_t\right) - \nabla \log q_t\left(X_s\right)\right\|^2 \le \left\|A\nabla \log q_t^{\mathrm{LD}}\left(A^\top X_t\right) - A\nabla \log q_t^{\mathrm{LD}}\left(A^\top X_s\right)\right\|^2 + \frac{\|I_d - AA^\top\|_F^2(\sigma_s^2 - \sigma_t^2)}{\sigma_t^4}$$

As in the first part of this proof, we need to control the Hessian matrix:

$$AA^\top \nabla^2 \log q_t^{\mathrm{LD}}(A^\top X_t)$$

It is easy to check that the proof before

$$\begin{aligned}\langle \mathbb{E}M_t \otimes M_t, \mathbb{E}Z_{t,s} \otimes Z_{t,s}\rangle &\lesssim \sigma_{s-t}^4 \left(\sum_{(i_1,i_2)=(i_3,i_4)} + \sum_{(i_1,i_2)=(i_4,i_3)}\right)\mathbb{E}\left(M_t \otimes M_t\right)_{i_1,i_2,i_3,i_4} \\ &\lesssim (\sigma_s^2 - \sigma_t^2)^2 \mathbb{E}\left\|AA^\top \nabla^2 \log q_t\left(AX_t\right)\right\|_F^4\,,\end{aligned}$$

is exactly the same with the first part of this lemma. For the latent score $\nabla^2 \log q_t^{\mathrm{LD}}\left(A^\top X_t\right)$, we have that

$$\mathbb{E}\left[\left\|\nabla^2 \log q_t^{\mathrm{LD}}\left(A^\top X_t\right)\right\|_F\right] \lesssim \frac{d'}{\sigma_t^2}\,.$$

Then, we know that

$$\mathbb{E}\left[\left\|AA^\top \nabla^2 \log q_t^{\mathrm{LD}}\left(A^\top X_t\right)\right\|_F^2\right] \le \|AA^\top\|_F^2 \frac{d'^2}{\sigma_t^4} = \frac{d'^3}{\sigma_t^4}\,,$$

where the last equality follows the fact that A is a column orthonormal matrix. Then, we complete the proof. ■

The discretization error consists of the space and time discretizaiton error and the above lemma control one step space discretization error. Then, the following lemma shows that the discretization error is dominated by the space error. We note that the proof process is exactly the same compared to Lemma 11 of Chen et al. (2023a) since we just set $\alpha_{t,s} = 1$ in Chen et al. (2023a).

Lemma 16. *For any $0 \le t \le s \le T$, the forward process 1 satisfies*

$$\mathbb{E}\left\|\nabla \log q_t\left(X_t\right) - \nabla \log q_s\left(X_s\right)\right\|^2 \le 4\mathbb{E}\left\|\nabla \log q_t\left(X_t\right) - \nabla \log q_t\left(X_s\right)\right\|^2\,.$$

F.2 The Lemmas for the Forward Process

As shown in Lemma 12, we also need to control $\mathbb{E}[\|X_t\|^2]$ in the forward process. The following lemma shows that this term is bounded by the R^2 and exploding variance.

Lemma 17. *Suppose that* **Assumption 1** *hold. Let $(X_t)_{t\in[0,T]}$ denote the forward process Eq.* (1)*. Then, for all $t \ge 0$,*

$$\mathbb{E}\left[\|X_t\|^2\right] \le d\sigma_t^2 + R^2$$

Proof. We know that $X_t = X_0 + \sigma_t Z$, where $Z \sim \mathcal{N}(0, I)$. Hence, we have

$$\mathbb{E}\left[\|X_t\|^2\right] = \mathbb{E}\left[\|X_0\|^2\right] + \sigma_t^2 d \le d\sigma_t^2 + R^2\,.$$

■

Lemma 18 (movement bound for VESDE)**.** *Let $(X_t)_{t\in[0,T]}$ denote the forward process Eq.* (1)*. For $0 \le s < t$ with $\delta := t - s$, if $\delta \le 1$, then*

$$\mathbb{E}\left[\|X_t - X_s\|^2\right] \lesssim g^2(t)d\delta\,.$$

Proof. If considering pure VESDE Eq. (1), since the movement is purely dominated by the variance term σ_t^2, we can write

$$\mathbb{E}\left[\|X_t - X_s\|^2\right] = \mathbb{E}\left[g^2(t)\,\|B_t - B_s\|^2\right] \lesssim g^2(t)\delta d\,.$$

■

Similar to Chen et al. (2022b), we can also show that if we do forward process for time δ, q_δ will be close to q_0 in W_2 distance. Note that different from Lemma 18, for VESDE (SOTA), we can choose large early stopping $\delta = \frac{\epsilon_{W_2}}{\sqrt{d}}$ since the forward time start at 0 instead of $s \ge 0$ in Lemma 19.

Lemma 19. *Suppose* **Assumption 1** *holds. Let $\epsilon_{W_2} > 0$. (1) If $\sigma_t^2 = t$, we choose the early stopping parameter $\delta = \frac{\epsilon_{W_2}^2}{d}$; (2) If $\sigma_t^2 = t^2$, we choose the early stopping parameter $\delta = \frac{\epsilon_{W_2}}{\sqrt{d}}$, then we have $W_2\left(q_\delta, q_0\right) \le \epsilon_{W_2}$.*

Proof. For the forward process Eq. (1), we know that $X_t := X_0 + \sigma_t z$, where $z \sim \text{normal}\left(0, I_d\right)$ is independent of X_0. Hence, for $\delta \lesssim 1$,

$$W_2^2\left(q_0, q_\delta\right) \le \mathbb{E}\left[\|\sigma_\delta Z\|^2\right] = \sigma_\delta^2 d\,.$$

Then, for the setting $\sigma_t^2 = t$, we can take $\delta \le \frac{\epsilon_{W_2}^2}{d}$. For the setting $\sigma_t^2 = t^2$, we can take $\delta \le \frac{\epsilon_{W_2}}{\sqrt{d}}$. ■

The following lemma is used to control the reverse beginning error of VESDE, which is proposed by Yang et al. (2024).

Lemma 20. *Assume* **Assumption 1**. *Let $q_\infty = \mathcal{N}(0, \sigma_T^2 I)$. Then, we have*

$$\text{TV}(q_T, q_\infty) \le \bar{D}/\sigma_T\,,$$

where $\bar{D} = d|c| + \mathbb{E}[q_0] + R$ and c is the eigenvalue of $Cov[q_0]$ with the largest absolute value.

G Experiments Detail

Setting. The target distribution is a mixture of five Gaussian with $d = 2$, which is highly non-log-concave. Following the setting of Chen et al. (2023c), we use the closed-form score function of the mixture of Gaussian and focus on the discretization and reverse beginning error. We set $T = 3$ for VESDE (SOTA).

Similar with Chen et al. (2023c), we use the uniform predictor stepsize at each stage $h_{\text{pred}} = T/K_0$. The corrector consists of 3 steps of the underdamped Langevin algorithm with stepsize 0.001 at each stage. The above experiments are conduct on a GeForce RTX 4090 and it take 3 minutes to generate 40000 datapoints with the closed-form score function.

Metric. We sample 40000 datapoints and calculate the KL divergence between the generated samples and the ground truth samples.