THE CONVERGENCE OF VARIANCE EXPLODING DIFFUSION MODELS UNDER THE MANIFOLD HYPOTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Variance Exploding (VE) based diffusion models, an important class of diffusion models, have empirically shown state-of-the-art performance in many tasks. However, there are only a few theoretical works on the VE-based models, and those works suffer from a worse forward process convergence rate $1/\text{poly}(T)$ than the $\exp(-T)$ results of Variance Preserving (VP) based models, where T is the forward diffusion time. The slow rate is due to the Brownian Motion without the drift term and introduces hardness in balancing the different error sources. In this work, we design a new forward VESDE process with a small drift term, which converts data into pure Gaussian noise while the variance explodes. Furthermore, unlike the previous theoretical works, we allow the diffusion coefficient to be unbounded instead of a constant, which is closer to the SOTA VE-based models. With an aggressive diffusion coefficient, the new forward process allows a faster $\exp(-T)$ rate. By exploiting this new process, we prove the first polynomial sample complexity for VE-based models with reverse SDE under the realistic manifold hypothesis. Then, we focus on a more general setting considering reverse SDE and probability flow ODE simultaneously and propose the unified tangent-based analysis framework for VE-based models. In this framework, we prove the first quantitative convergence guarantee for SOTA VE-based models with probability flow ODE. We also show the power of the new forward process in balancing different error sources on the synthetic experiments to support our theoretical results.

1 INTRODUCTION

In recent years, diffusion modeling has become an important paradigm for generative modeling and has shown SOTA performance in image generation, audio synthesis, and video generation (Rombach et al., 2022; Saharia et al., 2022; Popov et al., 2021; Ho et al., 2022). The core of diffusion models are two stochastic differential equations (SDE): the forward and reverse processes. The forward process can be described by an intermediate marginal distribution $\{q_t\}_{t\in[0,T]}$, which converts the data q_0 to Gaussian noise q_T. There are two types of forward SDE (Song et al., 2020b): (1) Variance Preserving (VP) SDE and (2) Variance Exploding (VE) SDE. The VPSDE corresponds to an Ornstein-Uhlenbeck (OU) process, and the stationary distribution is $\mathcal{N}(0, \mathbf{I})$. The VESDE corresponds to the Brownian motion with a (deterministic) time change, which has an exploding variance. In earlier times, VP-based models (Ho et al., 2020; Lu et al., 2022) provide an important boost for developing diffusion models. Recently, Karras et al. (2022) unify VPSDE and VESDE formula and show that the optimal parameters correspond to VESDE. Furthermore, VE-based models achieve SOTA performance in multi-step (Kim et al., 2022; Teng et al., 2023) and one-step image generation (Song et al., 2023).

After determining the forward SDE, the models reverse the forward SDE and generate samples by running the corresponding reverse SDE (Ho et al., 2020; Lu et al., 2022) or probability flow ODE (PFODE) (Song et al., 2020a; Karras et al., 2022). Before running the reverse process, three things need to be chosen: (1) a tractable reverse beginning distribution; (2) discretization scheme; (3) estimation of the score function $\nabla \log q_t$ (Ho et al., 2020). These procedures are error sources of the convergence guarantee, and we need to balance these terms to obtain the quantitative guarantee.

Despite the empirical success of diffusion models, only a few works focus on the convergence guarantee of these models. Furthermore, many previous works (Chen et al., 2023d;a;c; Lee et al., 2023; Benton et al., 2023) focus on VPSDE, and the analysis for VESDE is lacking. When considering the reverse beginning error, Lee et al. (2022) analyze a special VESDE with constant diffusion coefficient β_t and reverse SDE and obtain a $1/\sqrt{T}$ guarantee. This guarantee is worse than VP-based models since the reverse beginning error is $\exp(-T)$ for VP-based models. The above reverse beginning error introduces hardness to balance three error sources to achieve a great sample complexity, as shown in Section 5. To deal with this problem, De Bortoli et al. (2021) introduce a coefficient α to balance the drift and diffusion term. When choosing $\alpha = 1/T$, q_T is $\mathcal{N}(0, T\mathbf{I})$, and the reverse beginning error is $\exp\left(-\sqrt{T}\right)$. However, their framework only allows constant β_t, far from the SOTA choice (Karras et al., 2022), whose coefficient grows fast and is unbounded. Furthermore, De Bortoli et al. (2021) fail to balance the above three errors and has exponential dependence on problem parameters. Therefore, the following question remains open:

Is it possible to design a VESDE with a faster forward convergence rate than $1/poly(T)$ and balance error sources to achieve the polynomial complexity when the diffusion coefficient is unbounded?

In this work, for the first time, we propose a new forward VESDE, which allows unbounded coefficients and enjoys a faster forward convergence rate. We first show that the new process with a conservative β_t has similar trends but performs better compared to the original VESDE on synthetic data (Section 7). Then, we prove that the new process with a suitable aggressive β_t can balance reverse beginning, discretization and score function errors and achieve the first polynomial sample complexity for VE-based models. It is worth emphasizing that our results are achieved under the manifold hypothesis, which means the data q_0 is supported on a lower dimensional compact set $\mathcal{M}$. Different from the previously discussed works, which assume the score function is Lipschitz continuous (Chen et al., 2023e) or satisfy strong log-Sobelev inequality (LSI) assumption (Lee et al., 2022; Wibisono and Yang, 2022) to compulsorily guarantee that the score does not explode when $t \to 0$, the manifold hypothesis allows the expansion phenomenon of the score function (Kim et al., 2021; Pope et al., 2021). To our knowledge, only Pidstrigach (2022) analyze VESDE in the continuous reverse process under the manifold hypothesis, which misses the key discretization step.

Besides the polynomial sample complexity for VE-based models, another important point for VE-based models is the quantitative guarantee for VESDE with reverse PFODE. Recently, many works show that models with reverse PFODE have faster sample rate (Lu et al., 2022), and VE-based models with reverse PFODE (Song et al., 2020b; Karras et al., 2022) achieve great performance. Furthermore, the reverse PFODE can be applied in areas such as computing likelihoods (Song et al., 2020b) and one-step generation models (Salimans and Ho, 2022; Song et al., 2023). However, except for a few theoretical works for VPSDE (Chen et al., 2023e;c), the current works focus on reverse SDE. Hence, after achieving polynomial complexity for reverse SDE, we go one step further and propose the tangent-based unified framework for VE-based models, which contains reverse SDE and PFODE. We note that when considering VPSDE with reverse SDE, Bortoli (2022) also use the tangent-based method. However, as discussed in Section 6.2, the original tangent-based lemma can not deal with reverse PFODE even in VPSDE. We carefully control the tangent process to avoid additional $\exp(T)$ by using the exploding variance property of VESDE. Using this unified framework, we achieve the first quantitative convergence for the SOTA VE-based models with reverse PFODE (Karras et al., 2022). In conclusion, we accomplish the following results under the manifold hypothesis:

1. We propose a new forward VESDE with the unbounded coefficient β_t and a drift term that is typically small. With an aggressive β_t, the new process balances different error terms and achieves the first polynomial sample complexity for VE-based methods with reverse SDE.
2. When considering the general setting of VE-based methods, we propose the tangent-based unified framework, which contains reverse SDE and PFODE. Under this framework, we prove the first quantitative guarantee for SOTA VE-based models with reverse PFODE.
3. We show the power of our new forward process via synthetic experiments. For aggressive drift VESDE, we show it balances reverse beginning and discretization error. For conservative one, we show that it improves the quality of generated distribution without training.

2 Related Work

In the first two paragraphs of this section, we focus on works without the manifold hypothesis. In the last paragraph, we discuss the manifold hypothesis.

Analyses for VP-based models. De Bortoli et al. (2021) study VPSDE in TV distance and propose the first quantitative convergence guarantee with exponential dependence on problem parameters. Lee et al. (2022) achieve the first polynomial complexity with strong LSI assumption. Chen et al. (2023d) remove LSI assumption, assume the Lipschitz score and achieve polynomial complexity. Recently, Chen et al. (2023a) and Benton et al. (2023) further remove the Lipschitz score assumption, and Benton et al. (2023) achieve optimal dependence on d. When considering PFODE, Chen et al. (2023e) propose the first quantitative guarantee with exponential dependence on score Lipschitz constant. Chen et al. (2023c) achieve polynomial complexity by introducing a corrector component.

Analyses for VE-based models. When considering VESDE, most works focus on constant β_t and reverse SDE setting. De Bortoli et al. (2021) provide the first quantitative convergence guarantee with exponential dependence on problem parameters. Lee et al. (2022) analyze a constant diffusion coefficient VESDE and achieve polynomial sample complexity. However, the results of Lee et al. (2022) relies heavily on the LSI assumption, which is unrealistic (Remark 1). When considering the reverse PFODE and unbounded β_t, Chen et al. (2023e) only consider the discretization error and provide a quantitative convergence guarantee. Furthermore, as discussed in Section 6.2, their results introduce additional $\exp(T)$ compared to ours.

Analyses for diffusion models under the manifold hypothesis. There are other line works (Pidstrigach, 2022; Bortoli, 2022; Lee et al., 2023; Chen et al., 2023d;b) that consider the manifold hypothesis. Pidstrigach (2022) is the first work considering the guarantee of VESDE, VPSDE, and CLD in the continuous process. Bortoli (2022) study VPSDE with reverse SDE, and it is the most relevant work to our unified framework. However, as discussed in Section 1, we need a refined analysis of the tangent process for VESDE with reverser PFODE to avoid $\exp(T)$. Chen et al. (2023d) and Lee et al. (2023) also analyze VPSDE and achieve the polynomial complexity. Chen et al. (2023b) study the score learning process and distribution recovery on the low dimensional data.

3 The Variance Exploding (VE) SDE for Diffusion Models

In this section, we introduce the basic knowledge of VE-based diffusion models from the perspective of SDE and discuss the assumption on the balance coefficient τ and diffusion coefficient β_t.

3.1 The Forward Process

First, we denote by q_0 the data distribution, $\mathbf{X}_0 \sim q_0 \in \mathbb{R}^d$, and $\{\beta_t\}_{t\in[0,T]}$ a non-decreasing sequence in $[0,T]$. Then, we define the forward process:

$$\mathrm{d}\mathbf{X}_t = -\frac{1}{\tau}\beta_t \mathbf{X}_t \,\mathrm{d}t + \sqrt{2\beta_t}\,\mathrm{d}\mathbf{B}_t, \quad \mathbf{X}_0 \sim q_0\,, \tag{1}$$

where $(\mathbf{B}_t)_{t\geq 0}$ is a d-dimensional Brownian motion, and $\tau \in [T, T^2]$ is the coefficient to balance the drift and diffusion term. We also define q_t^τ as the marginal distribution of the forward OU process at time t. For well-defined β_t (**Assumption 1**), we have that

$$\mathbf{X}_t = m_t\mathbf{X}_0 + \sigma_t Z,\, m_t = \exp\left[-\int_0^t \beta_s/\tau \,\mathrm{d}s\right],\, \sigma_t^2 = \tau\left(1-\exp\left[-2\int_0^t \beta_s/\tau\,\mathrm{d}s\right]\right)\,, \tag{2}$$

where $Z \sim \mathcal{N}(0,\mathbf{I})$. Later, we will discuss the choice of β_t, which depends on the type of reverse process and the choice of τ. When the previous works consider VESDE, they usually consider the forward process without the drift term:

$$\mathrm{d}\mathbf{X}_t = \sqrt{\mathrm{d}\sigma_t^2/\,\mathrm{d}t}\,\mathrm{d}\mathbf{B}_t, \quad \mathbf{X}_0 \sim q_0\,, \tag{3}$$

where σ_t^2 is a non-decreasing variance sequence. There are two common choices for VESDE. The first choice (Chen et al., 2023e) is $\sigma_t^2 = t$, whose marginal distribution q_T is $\mathcal{N}(\mathbf{X}_0, T\mathbf{I})$. This choice is the setting that most theoretical works focus on (De Bortoli et al., 2021; Lee et al., 2022) and is similar to Eq. (1) with $\tau = T^2$ and $\beta_t = \frac{1}{2}, \forall t \in [T]$ since $q_T^\tau = \mathcal{N}(\exp(\frac{2}{T})\mathbf{X}_0, (1-\exp(\frac{1}{T}))T^2\mathbf{I})$. The second SOTA choice (Karras et al., 2022; Song et al., 2023) is $\sigma_t^2 = t^2$, which corresponds to $\tau = T^2$, $\beta_t = t$ and $q_T^\tau = \mathcal{N}(e^{-1/2}\mathbb{E}[q_0], e^{-1}\mathrm{Cov}[q_0] + (1-e^{-1})T^2\mathbf{I})$. We note that this q_T^τ still contains mean and variance information and almost identical to Eq. (3) with $\sigma_t^2 = t^2$. To support our argumentation, we do simulation experiments and show that these two setting have similar trend (Fig. 1) . Hence, Eq. (1) is representative enough to represent current VESDE. Furthermore, the general forward SDE retaining the drift term will lead to a series of VESDE, which is helpful in choosing the tractable Gaussian distribution q_∞^τ and enjoys a greater sample complexity (Section 5).

3.2 The Reverse Process

Reversing the forward SDE, we obtain the reverse process $(\mathbf{Y}_t)_{t\in[0,T]} = (\mathbf{X}_{T-t})_{t\in[0,T]}$:

$$\mathrm{d}\mathbf{Y}_t = \beta_{T-t}\left\{\mathbf{Y}_t/\tau + (1+\eta^2)\nabla \log q_{T-t}\left(\mathbf{Y}_t\right)\right\}\mathrm{d}t + \eta\sqrt{2\beta_{T-t}}\,\mathrm{d}\mathbf{B}_t\,, \tag{4}$$

where now $(\mathbf{B}_t)_{t\geq 0}$ is the reversed Brownian Motion and $\eta \in [0,1]$. When $\eta = 1$, the reverse process corresponds to reverse SDE. When $\eta = 0$, the process corresponds to PFODE. Since $\nabla \log q_t$ can not be computed exactly, we need to use a score function $\mathbf{s}(t,\cdot)$ to approximate them. Then, we introduce the continuous-time reverse process $(\widehat{\mathbf{Y}}_t)_{t\in[0,T]}$ with approximated score :

$$\mathrm{d}\widehat{\mathbf{Y}}_t = \beta_{T-t}\big\{\widehat{\mathbf{Y}}_t/\tau + (1+\eta^2)\mathbf{s}(T-t,\widehat{\mathbf{Y}}_t)\big\}\mathrm{d}t + \eta\sqrt{2\beta_{T-t}}\,\mathrm{d}\mathbf{B}_t, \quad \widehat{\mathbf{Y}}_0 \sim q_\infty^\tau\,, \tag{5}$$

where q_∞^τ is the reverse beginning distribution, which always is a tractable Gaussian distribution. In this work, similar to Karras et al. (2022) and Song et al. (2023), we choose $q_\infty^\tau = \mathcal{N}(0,\sigma_T^2\mathbf{I})$. The last step is to discrete the continuous process. We define $\{\gamma_k\}_{k\in\{0,\ldots,K\}}$ as the stepsize and $t_{k+1} = \sum_{j=0}^{k}\gamma_j$. In this work, we adapt the early stopping technique $t_K = T-\delta$, which has been widely used in practice (Ho et al., 2020; Kim et al., 2021; Karras et al., 2022). In this work, we consider the exponential integrator (EI) discretization (Zhang and Chen, 2022), which freezes the score function at time t_k and defines the new SDE for small interval $t \in [t_k, t_{k+1}]$:

$$\mathrm{d}\widetilde{\mathbf{Y}}_t = \beta_{T-t}\big\{\widetilde{\mathbf{Y}}_t/\tau + (1+\eta^2)\mathbf{s}(T-t_k,\widetilde{\mathbf{Y}}_t)\big\}\mathrm{d}t + \eta\sqrt{2\beta_{T-t}}\,\mathrm{d}\mathbf{B}_t, \quad t\in[t_k,t_{k+1}]\,. \tag{6}$$

Compared to the Euler–Maruyama (EM) discretization, EI has better experimental performance, and the results for EI can transfer to EM discretization (Bortoli, 2022). In this work, we consider β_t can increase rapidly, for example, $\beta_t = t^2$, instead of a constant (Chen et al., 2023d) or in a small interval $[1/\bar{\beta},\bar{\beta}]$ (Bortoli, 2022). Hence, we make the following assumption on β_t.

Assumption 1. *Define $t \mapsto \beta_t$ as a continuous, non-decreasing sequence. For any $\tau \in [T,T^2]$, there exists constants $\bar{\beta}$ and G, which are independent of t, such that for any $t\in[0,T]$: (1) if $\eta = 1$, then $1/\bar{\beta} \leq \beta_t \leq \max\{\bar{\beta},t^2\}$; (2) if $\eta\in[0,1)$, then $1/\bar{\beta}\leq\beta_t\leq\max\{\bar{\beta},t\}$ and $\int_0^T \beta_t/\tau\,\mathrm{d}t \leq G$.*

This assumption rules out cases where β_t grows too fast, such as e^t. We emphasize that β_t grows slower than t in the real world and satisfies $\int_0^T \beta_t/\tau\,\mathrm{d}t \leq G$ (Song et al., 2020b; Karras et al., 2022). However, our assumption is more general since β_t depends on τ instead of at most linearly. For example, when $\eta = 1$ and $\tau = T^2$, we can choose $\beta_t = t^2$, which has the same order compared to τ.

Notations. For $x\in\mathbb{R}^d$ and $A\in\mathbb{R}^{d\times d}$, we denote by $\|x\|$ and $\|A\|$ the Euclidean norm for vector and the spectral norm for matrix. We denote by q_0P_T the distribution of $\mathbf{X}_T$, $Q_{t_K}^{q_\infty^\tau}$ the distribution of $\mathbf{Y}_{t_K}$, $R_K^{q_\infty^\tau}$ the distribution of $\widetilde{\mathbf{Y}}_{t_K}$ and $Q_{t_K}^{q_0P_T}$ the distribution which does forward process, then does reverse process (Eq. (4)). We denote by W_1 and W_2 the Wasserstein distance of order one and two.

4 The Faster Forward Convergence Guarantee for the VESDE

Since the previous VESDE (Eq. (3)) does not converge to a stationary distribution, we usually choose a normal distribution $\mathcal{N}(\bar{m}_T,V_T)$ to approximate q_T. Pidstrigach (2022) show that the optimal solution is $\bar{m}_T = \mathbb{E}[q_0]$ and $V_T = \mathrm{Cov}[q_0]+\sigma_T^2\mathbf{I}$, which leads $1/\sigma_T^2$ forward convergence rate. To obtain a faster forward convergence rate, we introduce a new process, which allows $\mathbb{E}[q_0]$ and $\mathrm{Cov}[q_0]$ decay, as well as the variance explodes. In particular, we have the following lemma.

Lemma 1. *The minimization problem $\min_{\bar{m}_t,V_t} KL\left(q_t \mid \mathcal{N}(\bar{m}_t,V_t)\right)$ is minimized by $\bar{m}_t = m_t\mathbb{E}\left[q_0\right]$ and $V_t = m_t^2\,\mathrm{Cov}\left[q_0\right]+\sigma_t^2\mathbf{I}$, where m_t and σ_t defined in Eq. (2).*

Lemma 1 is a general version compare to Pidstrigach (2022) since the existence of variable m_t instead of constant one. Theorem 1 shows that variable m_t allows a faster forward convergence rate. Before introducing the results, we introduce the manifold hypothesis on the data distribution.

Assumption 2. *q_0 is supported on a compact set $\mathcal{M}$ and $0 \in \mathcal{M}$.*

We denote R the diameter of the manifold by $R = \sup\{\|x - y\| : x, y \in \mathcal{M}\}$ and assume $R > 1$. The manifold hypothesis is supported by much empirical evidence (Bengio et al., 2013; Fefferman et al., 2016; Pope et al., 2021) and is naturally satisfied by the image datasets since each channel of images is bounded. Furthermore, different from the Lipschitz score assumption, this assumption allows the expansion phenomenon of the score Kim et al. (2021) and is studied in the VPSDE setting. As shown in Bortoli (2022), this assumption also encompasses distributions which admit a continuous density on a lower dimensional manifold, and R contains the dimension information. For example, considering a hypercube $\mathcal{M} = [-1/2, 1/2]^p$ with $p \leq d$, the $R = \sqrt{p}$, corresponding to the latent dimension p. Then, we obtain the following guarantee for the new forward process.

Theorem 1. *Let q_T be the marginal distribution of the forward process, and $q_\infty^\tau = \mathcal{N}(0, \sigma_T^2 \mathbf{I})$ be the reverse beginning distribution. With m_T, σ_T defined in Eq. (2), we have*

$$\|q_T - q_\infty^\tau\|_{TV} \leq \sqrt{m_T}\bar{D}/\sigma_T \,,$$

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

Recall that $m_T = \exp[-\int_0^T \beta_t/\tau \,\mathrm{d}t]$, the previous VESDE (Song et al., 2020b; Karras et al., 2022; Lee et al., 2022) choose a conservative β_t satisfies $\int_0^T \beta_t/\tau \,\mathrm{d}t \leq G$. However, if we allow an aggressive β_t, the forward process will have a faster convergence rate. To illustrate the accelerated forward process, we use $\tau = T^2$ as an example and discuss the influence of different $\beta_t = t^{\alpha_1}, \alpha_1 \in [1, 2]$ under large enough T. Due to the definition of σ_T, $\sigma_T \approx T$, and the forward convergence rate mainly depends on $\sqrt{m_T}$. When $\alpha_1 = 1$ is conservative, m_T is a constant, and the convergence rate is $1/T$. When $\alpha_1 = 1 + \ln(2r\ln(T))/\ln(T)$ is slightly aggressive, the convergence rate is $1/T^{r+1}$ for $r > 0$. When $\alpha_1 \geq 1 + \ln(T - \ln(T))/\ln(T)$ is aggressive, the forward convergence rate is faster than $\exp(-T)$. When $\alpha_1 = 2$ is the most aggressive choice, the convergence rate is $\exp(-T)/T$. In our analysis, whether β_t can be aggressive in our unified framework depends on the form of the reverse process, as the discussion in Section 6. In the following section, we show that when choosing aggressive β_t (reverse SDE setting), the new process balances the reverse beginning, discretization, and approximated score errors and achieves the first polynomial sample complexity for VE-based models under the manifold hypothesis.

5 The Polynomial Complexity for VESDE with Reverse SDE

In this section, we first pay attention to VESDE with reverse SDE to show the power of our new forward process and aggressive β_t. In this section, we assume an uniform L_2-accuracy assumption on scores, which is exactly the same compared to Chen et al. (2023d); Benton et al. (2023).

Assumption 3 (Approximated score)**.** *For all $k = 1, \ldots, K, \mathbb{E}_{q_{t_k}}\left[\|s_{t_k} - \nabla \ln q_{t_k}\|^2\right] \leq \epsilon_{score}^2$.*

We show that introducing aggressive β_t only slightly affects the discretization error (additional logarithmic factors) and significantly benefits in balancing reverse beginning and discretization errors. Hence, we can obtain a polynomial sample complexity for VE-based models with unbounded β_t.

Corollary 1. *Assume* **Assumption 1***, 2, 3. Let $\gamma_K = \delta$, $\bar{\gamma}_K = \text{argmax}_{k\in\{0,\ldots,K-1\}}\gamma_k$, $\tau = T^2$ and $\beta_t = t^2$. Then, TV$\left(R_K^{q_\infty^\tau}, q_0\right)$ is bounded by*

$$\mathrm{TV}\left(R_K^{q_\infty^\tau}, q_0\right) \leq \frac{\bar{D}\exp(-T/2)}{T} + \frac{R^2\sqrt{d}}{\delta^6}\sqrt{\bar{\gamma}_K T^5} + \epsilon_{score}\sqrt{T^3} \,,$$

where $\bar{D} = d|c| + \mathbb{E}\left[q_0\right] + R$. Furthermore, by choosing $\delta \leq \frac{\epsilon_{W_2}^{2/3}}{(d+R\sqrt{d})^{1/3}}$, $T \geq 2\ln\frac{\bar{D}}{\epsilon_{TV}}$, maximum stepsize $\bar{\gamma}_K \leq \delta^{12}\epsilon_{TV}^2\ln^5(\bar{D}/\epsilon_{TV})/R^4 d$ and assuming $\epsilon_{score} \leq \widetilde{O}(\epsilon_{TV})$, the output of $R_K^{q_\infty^\tau}$ is $(\epsilon_{TV} + \epsilon_{score})$ close to q_δ, which is ϵ_{W_2} close to q_0, with sample complexity (hiding logarithmic factors) is

$$K \leq \widetilde{O}\left(\frac{dR^4(d + R\sqrt{d})^4}{\epsilon_{W_2}^8 \epsilon_{TV}^2}\right) .$$

For choice $\beta_t = t$ and $\tau = T$, by choosing $\delta \leq \frac{\epsilon_{W_2}}{(d+R\sqrt{d})^{1/2}}$ and $\bar{\gamma}_K \leq \frac{\delta^8 \epsilon_{TV}^2 \ln^3(\bar{D}/\epsilon_{TV})}{R^4 d}$, we obtain the same sample complexity.

First, we discuss the power of our new process and aggressive coefficient under the setting $\tau = T^2$, which is closer to the SOTA setting. Then, we discuss the results of $\tau = T$ and compare them to existing polynomial sample complexity results in Remark 1. For $\tau = T^2$, the above results show that our new process with aggressive $\beta_t = t^2$ balances the reverse beginning, discretization scheme and approximated score errors. If choosing a conservative $\beta_t = t$, term $\exp(-T/2)$ will be removed. Then, the guarantee has the form $1/T + \sqrt{\bar{\gamma}_K T^5}/\delta^6 + \epsilon_{\text{score}}\sqrt{T^3}$, which means if $T \geq 1/\epsilon_{\text{TV}}$, then $\epsilon_{\text{score}}\sqrt{T^3} \geq \epsilon_{\text{score}}/\sqrt{\epsilon_{\text{TV}}^3}$. Hence, it is hard to achieve non-asymptotic results for conservative β_t. However, by choosing aggressive β_t, T becomes the logarithmic factor, and these error sources are balanced. In Section 7.1, we do experiments on 2-D Gaussian to support our above augmentation.

Remark 1. *Lee et al. (2022) consider pure VESDE (Eq. (3)) with $\sigma_t^2 = t$ and reverse SDE under the LSI assumption with parameter C_{LS}. The LSI assumption does not allow the presence of substantial non-convexity and is far away from the multi-modal real-world distribution. Furthermore, they use unrealistic assumption $\epsilon_{score} \leq 1/(C_{LS} + T)$ to avoid the effect of the approximated score, which is stronger than* **Assumption 3***. Under the above strong assumption on data and approximated score function, Lee et al. (2022) achieve the polynomial sample complexity $\tilde{O}(L^2 d(d|c| + R)^2/\epsilon_{TV}^4)$. Under the manifold hypothesis and choosing $L = R^2 d^2/\epsilon_{W_2}^4$, the sample complexity is $\tilde{O}(R^4 d^5 (d|c| + R)^2/\epsilon_{W_2}^8 \epsilon_{TV}^4)$, which worse than Corollary 1.*

6 The Unified Analysis Framework for VE-based Methods

In this section, we go beyond the reverse SDE setting and introduce the unified analysis framework for VESDE with reverse SDE and PFODE. In Section 6.1, we show the convergence results in our unified framework. In Section 6.2, we introduce the detail of our tangent-based unified framework and discuss the variance exploding property of VESDE, which allows analyzing reverse PFODE.

6.1 The Convergence Guarantee for VESDE

In this part, we consider the reverse beginning and discretization errors and assume an accurate score. This setting is similar to Chen et al. (2023e), which mainly considers the reverse PFODE. It is a meaningful step to analyze VESDE with reverse SDE and PFODE in a unified framework and show the property of VESDE instead of the property of reverse process.

Theorem 2. *Assume* **Assumption 1** *and* **2***, $\gamma_k \sup_{v \in [T-t_{k+1}, T-t_k]} \beta_v/\sigma_v^2 \leq 1/28$ for $\forall k \in \{0, ..., K-1\}$, and $\delta \leq 1/32$. Let $\bar{\gamma}_K = argmax_{k \in \{0,...,K-1\}} \gamma_k$ and $\gamma_K = \delta$. Then, for $\forall \tau \in [T, T^2]$, we have the following convergence guarantee.*

(1) If $\eta = 1$ (the reverse SDE), choosing an aggressive $\beta_t = t^2$, we have

$$W_1\left(R_K^{q_\infty^\tau}, q_0\right) \leq C_1(\tau) T \exp\left[\frac{R^2}{2}(\frac{\bar{\beta}}{\delta^3} + \frac{1}{\tau})\right][\kappa_1^2(\tau)(\frac{\bar{\beta}}{\delta^3} + \frac{1}{\tau})\bar{\gamma}_K^{1/2} + \kappa_1^2(\tau)]\bar{\gamma}_K^{1/2} + \exp\left[\frac{R^2}{2}(\frac{\bar{\beta}}{\delta^3} + \frac{1}{\tau})\right]\frac{\bar{D}\exp(-T/2)}{\sqrt{\tau}} + 2(\frac{R}{\tau} + \sqrt{d})\sqrt{\delta}\,,$$

where $C_1(\tau)$ is linear in τ^2, $\kappa_1(\tau) = \max\{\bar{\beta}, T^2\}(1/\tau + \bar{\beta}/\delta^3)$.

(2) If $\eta = 0$ (the reverse PFODE), choosing a conservative β_t satisfies **Assumption 1***, we have*

$$W_1\left(R_K^{q_\infty^\tau}, q_0\right) \leq C_2(\tau) T \exp\left[\frac{R^2}{2}(\frac{\bar{\beta}}{\delta^2} + \frac{1}{\tau}) + \frac{1}{2}\right][\kappa_2^2(\tau)(\frac{\bar{\beta}}{\delta^2} + \frac{1}{\tau})\bar{\gamma}_K^{1/2} + \kappa_2^2(\tau)]\bar{\gamma}_K^{1/2} + \exp\left[\frac{R^2}{2}(\frac{\bar{\beta}}{\delta^2} + \frac{1}{\tau})\right]\frac{\bar{D}}{\sqrt{\tau}} + 2(\frac{R}{\tau} + \sqrt{d})\sqrt{\delta}\,,$$

where $C_2(\tau)$ is linear in τ^2, $\kappa_2(\tau) = \max\{\bar{\beta}, T\}(1/\tau + \bar{\beta}/\delta^2)$.

Theorem 2 proves the first quantitative guarantee for VE-based models with reverse PFODE in the unified tangent-based method (Section 6.2). We also show that this framework can deal with reverse SDE. As shown in Theorem 2, for different reverse processes, our unified framework chooses different β_t to achieve a quantitative convergence guarantee, which shows the power of our framework. Correspondingly, the Girsanov-based method (Chen et al., 2023d;a; Benton et al., 2023) can not obtain the guarantee of reverse PFODE since the diffusion term for reverse process is not well-defined.

For the reverse PFODE, Chen et al. (2023e) employ the Restoration-Degradation framework to analyze the VESDE. Since their result has an exponential dependence on the Lipschitz constant of the score, their results also have exponential dependence on R and δ. Furthermore, their results have exponential dependence on the growth rate of β_t (g_{max} in (Chen et al., 2023e)), which corresponds to τ of VESDE. However, our dependence on τ appears in the polynomial term. Hence, our framework is a suitable unified framework for VE-based methods. Furthermore, we emphasize that our tangent-based unified framework is not a simple extension of Bortoli (2022). We carefully control the tangent process according to the variance exploding property of VESDE to avoid $\exp(T)$ term when considering PFODE, as discussed in Section 6.2.

Corollary 2. *Assume* **Assumption 1** *and* **2**. *Let* $\epsilon \in (0, 1/32), \tau \in [T, T^2]$, $\gamma_K = \delta = \epsilon^2, \bar{\gamma}_K = \text{argmax}_{k\in\{0,\dots,K-1\}}\gamma_k$,

(1) if $\eta = 1$, *with an aggressive* $\beta_t = t^2$, $T \geq \bar{\beta}(R^2+1)/\epsilon^6$, $\bar{\gamma}_K \leq \exp(-T)/(T^{\frac{20}{3}}\tau^4 C_1^2(\tau))$*;*

(2) if $\eta = 0$, *with a conservative* β_t (**Assumption 1**), $\tau \geq \exp(\frac{R^2\bar{\beta}}{\epsilon^4})/\epsilon^2$ *and* $\bar{\gamma}_K \leq \frac{1}{\tau T^6 \ln^2(T) C_2^2(\tau)}$*:*

$$W_1\left(R_K^{q_\infty^\tau}, q_0\right) \leq (\bar{D} + 2R^2 + 2\sqrt{d})\epsilon\,,$$

where $\bar{D}$ *is defined in Theorem 1, and* $\max\{C_1(\tau), C_2(\tau)\} \leq (16+\bar{\beta}^{\frac{3}{2}}+1)(2+R^2)(12R+4\tau^2\sqrt{d})$.

We note that Theorem 2 has exponential dependence on R and δ, which is introduced by the tangent process. Similar to Bortoli (2022), if we assume the Hessian $\left\|\nabla^2 \log q_t(x_t)\right\| \leq \Gamma/\sigma_t^2$, we obtain a better control on the tangent process and replace the exponential dependence on δ by a polynomial dependence on δ and exponential dependence on Γ when considering reverse PFODE.

Corollary 3. *Assume* **Assumption 1**, **Assumption 2** *and* $\left\|\nabla^2 \log q_t(x_t)\right\| \leq \Gamma/\sigma_t^2$. *Let* $\eta = 0$ *(reverse PFODE),* $\epsilon \in (0, 1/32), \tau = T^2$, $\beta_t = t$, $\bar{\gamma}_K = \text{argmax}_{k\in\{0,\dots,K-1\}}\gamma_k, \gamma_K = \delta$, *we have*

$$W_1\left(R_K^{q_\infty^\tau}, q_0\right) \leq C_2(\tau)T\frac{\bar{\beta}^{\frac{\Gamma}{2}}}{\delta^\Gamma}\exp\left[\frac{\Gamma+2}{2}\right][\kappa_2^2(\tau)(\frac{\bar{\beta}}{\delta^2}+\frac{1}{\tau})\bar{\gamma}_K^{1/2}+\kappa_2^2(\tau)]\bar{\gamma}_K^{1/2}$$
$$+\frac{\bar{\beta}^{\frac{\Gamma}{2}}}{\delta^\Gamma}\exp\left[\frac{\Gamma+2}{2}\right]\frac{\bar{D}}{\sqrt{\tau}}+2(\frac{R}{\tau}+\sqrt{d})\sqrt{\delta}\,,$$

where $C_2(\tau)$ *is linear in* τ^2, $\kappa_2(\tau) = \max\{\bar{\beta}, T\}\left(\frac{1}{\tau}+\frac{\bar{\beta}}{\delta^2}\right)$.

Though the Γ/σ_δ^2 bound is stronger than $(1+R^2)/\sigma_\delta^4$ in Theorem 2 and Chen et al. (2023d), there are many special cases such as hypercube $\mathcal{M} = [-1/2, 1/2]^p$ with $p \leq d$ satisfy this assumption.

Remark 2. *Since the reverse PFODE setting does not involve the aggressive* $\beta_t = t^2$*, our analysis still holds if we consider original VESDE Eq.* (3) *with* $\sigma_t^2 = t^2$ *and reverse PFODE, which means our analysis can explain the results in (Karras et al., 2022). For consistency, we use Eq.* (1) *in this work.*

Remark 3. *In this section, similar to Chen et al. (2023e), we assume an accurate score as the first step. When considering the approximated score, our guarantee has an additional* $\epsilon_{score}T$ *term. As discussed in Remark 1, we can eliminate the effect of* ϵ_{score} *by adding strong assumption. However, if assuming* **Assumption 3***, since* T *in Corollary 2 is not a logarithmic factor, we can not ignore it. One future work is considering an approximated score in PFODE and achieving a polynomial complexity.*

6.2 The Dicussion on the Unified Framework

In this section, we introduce the unified tangent-based framework for reverse SDE and PFODE and discuss key steps to achieve the quantitative guarantee for PFODE. Firstly, we decompose the goal:

$$W_1\left(R_K^{q_\infty^\tau}, q_0\right) \leq W_1\left(R_K^{q_\infty^\tau}, Q_{t_K}^{q_\infty^\tau}\right) + W_1\left(Q_{t_K}^{q_\infty^\tau}, Q_{t_K}^{q_0 P_T}\right) + W_1\left(Q_{t_K}^{q_0 P_T}, q_0\right)\,.$$

These terms correspond to the discretization scheme, reverse beginning distribution, and the early stopping parameter δ. We focus on most difficult discretization term and first recall the stochastic flow of the reverse process for any $x \in \mathbb{R}^d$ and $s, t \in [0, T]$ with $t \geq s$:

$$\mathrm{d}\mathbf{Y}_{s,t}^x = \beta_{T-t}\left\{\mathbf{Y}_{s,t}^x/\tau + \left(1+\eta^2\right)\nabla \log q_{T-t}\left(\mathbf{Y}_{s,t}^x\right)\right\}\mathrm{d}t + \eta\sqrt{2\beta_{T-t}}\mathrm{d}\mathbf{B}_t, \qquad \mathbf{Y}_{s,s}^x = x\,,$$

and the corresponding tangent process

$$\mathrm{d}\nabla\mathbf{Y}_{s,t}^x = \beta_{T-t}\left\{\mathbf{I}/\tau + \left(1+\eta^2\right)\nabla^2 \log q_{T-t}(\mathbf{Y}_{s,t}^x)\right\}\nabla\mathbf{Y}_{s,t}^x\mathrm{d}t, \qquad \nabla\mathbf{Y}_{s,s}^x = \mathbf{I}\,, \tag{7}$$

which is used to control the discretization error in Bortoli (2022). Then, the discretization error is bounded by time and space discretization error for a small interval $[t_k, t_{k+1}]$ and the tangent process $\left\|\nabla\mathbf{Y}_{s,t_K}^x\right\|$ for $\forall s \in [0, t_K]$. For the first two terms, we control the Lipshctiz constant $\left\|\nabla^2 \log q_t\left(x_t\right)\right\|$ and score perturbation $\left\|\partial_t \nabla \log q_t\left(x_t\right)\right\|$ at time $t \in [t_k, t_{k+1}]$. For the key $\left\|\nabla\mathbf{Y}_{s,t_K}^x\right\|$, we consider the reverse SDE and PFODE simultaneously and propose a general version of Bortoli (2022). Since the bound of tangent process depend on $\sigma_{T-t_K}^{-2}$, which corresponding to β_t, we introduce an indicator $i \in \{1, 2\}$ for σ_{T-t_K}. We use $\tau = T^2$ as an example. When $\beta_t = t^2$ is aggressive, $i = 1$, $\eta = 1$ and $\sigma_{T-t_K}^{-2}(i = 1) \leq 1/\tau + \bar{\beta}/\delta^3$. When $\beta_t = t$ is conservative, $i = 2$, $\eta \in [0, 1)$ and $\sigma_{T-t_K}^{-2}(i = 2) \leq 1/\tau + \bar{\beta}/\delta^2$. Then, we obtain the following guarantee for the tangent process.

Lemma 2. *Assume* **Assumption 1** *and* **2**. *Then, for $\forall s \in [0, t_K], x \in \mathbb{R}^d$, and $i \in \{1, 2\}$, we have*

$$\left\|\nabla\mathbf{Y}_{s,t_K,i}^x\right\| \leq \exp\left[\frac{R^2}{2\sigma_{T-t_K}^2(i)} + \frac{(1-\eta^2)}{2}\int_0^{t_K}\frac{\beta_{T-u}}{\tau}\mathrm{d}u\right].$$

Furthermore, if assuming $\left\|\nabla^2 \log q_t\left(x_t\right)\right\| \leq \Gamma/\sigma_t^2$, we have that

$$\left\|\nabla\mathbf{Y}_{s,t_K,i}^x\right\| \leq \sigma_{T-t_K}^{-(1+\eta^2)\Gamma}(i)\exp\left[\left(\left(1+\eta^2\right)\Gamma + 2\right)\int_0^{t_K}\frac{\beta_{T-u}}{\tau}\mathrm{d}u\right].$$

We emphasize that the general bound for the tangent process is the key to achieving the guarantee for VESDE with the reverse ODE. Recall that in the original lemma for the tangent processes, since τ is independent of T and β_t is bounded in a small interval $[1/\bar{\beta}, \bar{\beta}]$, $\int_0^{t_K}\beta_{T-u}/\tau\mathrm{d}u = \Theta(T)$, which means there is an additional $\exp\left(T\right)$ when considering VPSDE with revere PFODE. However, our tangent-based lemma makes use of the variance exploding property of VESDE to guarantee that $\int_0^T \beta_t/\tau\mathrm{d}t \leq G$ with a conservative $\beta_t = t$ when considering reverse PFODE. When considering reverse SDE ($\eta = 1$), we can choose aggressive $\beta_t = t^2$ since the choice of β_t does not affect the bound of the tangent process. Then, we control the discretization error for $\eta \in [0, 1]$. For the remaining two term, we know that the early stopping terms is smaller than $2(R/\tau + \sqrt{d})\sqrt{\delta}$ and

$$W_1\left(Q_{t_K}^{q_\infty^\tau}, Q_{t_K}^{q_0 P_T}\right) \leq \frac{\sqrt{m_T}\bar{D}}{\sigma_T}\exp\left[\frac{R^2}{2\sigma_{T-t_K}^2(i)} + \frac{(1-\eta^2)}{2}\int_0^{t_K}\frac{\beta_{T-u}}{\tau}\mathrm{d}u\right],$$

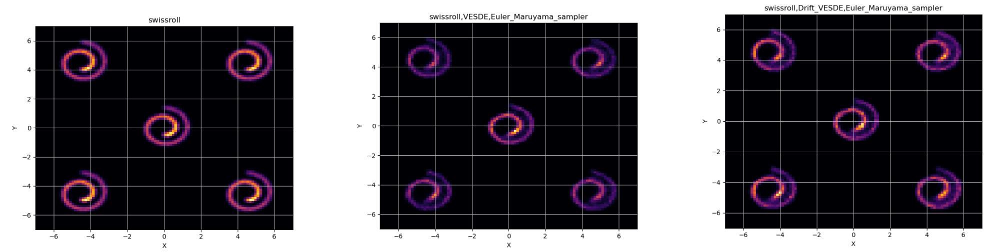
The reason why an exponential dependence in reverse beginning term is that we can not use the data processing inequality in Wasserstein distance. One future work is introducing the short regularization technique (Chen et al., 2023c) and suitable corrector to remove this exponential dependence.

7 Synthetic Experiments

In this section, we do some synthetic experiments to show the power of the new forward process. In Section 7.1, we show that with aggressive β_t, the new process achieves good balance in different error terms. Furthermore, we consider the approximated score and show that the conservative drift VESDE can improve the quality of the generated distribution without training.

7.1 The Power of Aggressive VESDE in Balancing Different Error Sources

In this section, we do experiments on 2-D Gaussian to show that the drift VESDE with aggressive β_t has power in balancing the reverse beginning and discretization errors. Since the ground truth score of the Gaussian can be directly calculated, we use the accurate score function to discuss the balance between the other two error terms clearly. We show how to use approximated score in Section 7.2.



(a) Original Figure (b) Pure VESDE (Eq. (3), $T = 100$) (c) Drift VESDE ($T = 100$)

Figure 2: Experiment results of Swiss roll with Euler Maruyama Method (Reverse SDE)

As shown in Fig. 1, the process with aggressive $\beta_t = t^2$ achieves the best and second performance in EI and EM discretization, which support our theoretical result (Corollary 1). The third best process is conservative $\beta_t = t$ with the small drift term. The reason is that though it can not achieve a $\exp(-T)$ forward guarantee, it also has a constant decay on prior information, as shown in Section 3.1. This decay slightly reduces the effect of the reverse beginning error. The worst process is pure VESDE since it is hard to balance different error sources. Our experimental results also show that EI is better than EM discretization.

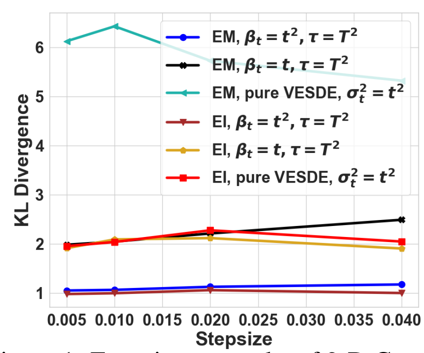


Figure 1: Experiment results of 2-D Gaussian

7.2 The Conservative Drift VESDE Benefits from Pure VESDE without Training

As shown in Fig. 1, the red and orange lines have similar trends. Hence, for conservative drift VESDE, which satisfies (2) of **Assumption 1**, we can directly use the models trained by pure VESDE to improve the quality of generated distribution. We confirm our intuition by training the model with pure VESDE (Eq. (3)) with $\sigma_t^2 = t$ and use the models to conservative drift VESDE with $\beta_t = 1$ and $\tau = T$. From the experimental results (Fig. 2), it is clear that pure VESDE has a low density on the Swiss roll except for the center one, which indicates pure VESDE can not deal with large dataset variance $\text{Cov}[q_0]$, as we discuss in Section 4. For conservative drift VESDE ($\beta_t = 1$ and $\tau = T$), as we discuss in the above section, it can reduce the influence of the dataset information. Fig. 2c support our augmentation and show that the density of the generated distribution is more uniform compared to pure VESDE, which means that the drift VESDE can deal with large dataset mean and variance.

There are more experiments on Swiss roll and 1D-GMM to explore different sampling methods (RK45, reverse PFODE) and different T. We refer to Appendix F for more details and discussion.

8 Conclusion

In this work, we analyze the VE-based models under the manifold hypothesis. Firstly, we propose a new forward VESDE process by introducing a small drift term, which enjoys a faster forward convergence rate than the Brownian Motion. Then, we show that with an aggressive β_t, the new process has the power to balance different error sources and achieve the first polynomial sample complexity for VE-based models with unbounded coefficient and reverse SDE.

After achieving the above results, we go beyond the reverse SDE and propose the tangent-based unified framework, which contains reverse SDE and PFODE. Under this framework, we make use of the variance exploding property of VESDE and achieve the first quantitative convergence guarantee for SOTA VE-based models with reverse PFODE. Finally, we do synthetic experiments to show the power of the new forward process.

Future Work. This work proposes the first unified framework for VE-based models with an accurate score. After that, we plan to consider the approximated score error and provide a polynomial sample complexity for the VE-based models with reverse PFODE under the manifold hypothesis.

REFERENCES

Vladimir M Alekseev. An estimate for the perturbations of the solutions of ordinary differential equations. *Vestn. Mosk. Univ. Ser. I. Math. Mekh*, 2:28–36, 1961.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

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.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Trans. Mach. Learn. Res.*, 2022, 2022. URL `https://openreview.net/forum?id=MhK5aXo3gB`.

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

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023d. URL `https://openreview.net/pdf?id=zyLVMgsZ0U_`.

Sitan Chen, Giannis Daras, and Alex Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. In *International Conference on Machine Learning*, pages 4462–4484. PMLR, 2023e.

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.

Pierre Del Moral and Sumeetpal Sidhu Singh. Backward itô–ventzell and stochastic interpolation formulae. *Stochastic Processes and their Applications*, 154:197–250, 2022.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 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.

Chieh-Hsin Lai, Yuhta Takida, Naoki Murata, Toshimitsu Uesaka, Yuki Mitsufuji, and Stefano Ermon. Fp-diffusion: Improving score-based diffusion models by enforcing the underlying score fokker-planck equation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 18365–18398. PMLR, 2023.

Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*. Springer, 2016.

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

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022.

Jakiw Pidstrigach. Score-based generative models detect manifolds. *arXiv preprint arXiv:2206.01018*, 2022.

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.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

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

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. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 32211–32252. PMLR, 2023.

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.

Andre Wibisono and Kaylee Yingxi Yang. Convergence in kl divergence of the inexact langevin algorithm with application to score-based generative models. *arXiv preprint arXiv:2211.01512*, 2022.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.

APPENDIX

A THE PROOF FOR THE FASTER FORWARD PROCESS

Lemma 1. *The minimization problem* $\min_{\bar{m}_t, V_t} KL\left(q_t \mid \mathcal{N}(\bar{m}_t, V_t)\right)$ *is minimized by* $\bar{m}_t = m_t\mathbb{E}\left[q_0\right]$ *and* $V_t = m_t^2 \operatorname{Cov}\left[q_0\right] + \sigma_t^2\mathbf{I}$*, where* m_t *and* σ_t *defined in Eq.* (2).

Proof. For simplicity, we denote the mean and covariance of q_0 by a and C. We also define the optimize variable $n_t = \mathcal{N}\left(\bar{m}_t, C_t\right)$. We can directly compute the KL divergence $KL(q_t|n_t)$:

$$
\begin{aligned}
KL\left(q_t|n_t\right) &= -H\left(q_t\right) - \int \log\left(n_t(x)\right) q_t(x)\mathrm{d}x \\
&= -H\left(q_t\right) + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log\left(\det\left(V_t\right)\right) + \frac{1}{2}\int (x-\bar{m}_t)^T V_t^{-1}(x-\bar{m}_t)q_t(x)\mathrm{d}x\,.
\end{aligned}
$$

For the last term, we directly compute

$$
\begin{aligned}
&\int \left(x-\bar{m}_t\right)^T V_t^{-1}\left(x-\bar{m}_t\right) p_t(x)\mathrm{d}x \\
=&\mathbb{E}\left[\left(X_t-\bar{m}_t\right)^T V_t^{-1}\left(X_t-\bar{m}_t\right)\right] = \mathbb{E}\left[\left(m_tX_0+\sigma_tZ-\bar{m}_t\right)^T V_t^{-1}\left(m_tX_0+\sigma_tZ-\bar{m}_t\right)\right] \\
=&\mathbb{E}\left[m_t^2\left(X_0-a\right)^T V_t^{-1}\left(X_0-a\right)\right] + \left(m_ta-\bar{m}_t\right)^T V_t^{-1}\left(m_ta-\bar{m}_t\right) + \sigma_t^2\mathbb{E}\left[Z^TV_t^{-1}Z\right] \\
=&m_t^2\operatorname{tr}\left(CV_t^{-1}\right) + \sigma_t^2\operatorname{tr}\left(V_t^{-1}\right) + \left(m_ta-\bar{m}_t\right)^T V_t^{-1}\left(m_ta-\bar{m}_t\right)\,,
\end{aligned}
$$

where the second inequality follows that $X_t = m_tX_0 + \sigma_tZ$. It is clear that the optimal solution of $\bar{m}_t$ is m_ta. In the next step, we focus on the optimization problem for V_t:

$$
\begin{aligned}
L\left(V_t^{-1}\right) &= \log\left(\det\left(V_t\right)\right) + \operatorname{tr}\left(\left(m_t^2C+\sigma_t^2\mathbf{I}\right)V_t^{-1}\right) \\
&= -\log\left(\det\left(V_t^{-1}\right)\right) + \operatorname{tr}\left(\left(m_t^2C+\sigma_t^2\mathbf{I}\right)V_t^{-1}\right)\,.
\end{aligned}
$$

Since the above optimization is a convex optimization problem, we use the method similar to Pidstrigach (2022), we obtain that the optimal solution of V_t is $m_t^2C + \sigma_t^2\mathbf{I}$. ∎

Lemma 3. *Let* $\bar{m}_t$ *and* V_t *be the optimal mean and covariance operator from Lemma 1. Then*

$$
\begin{aligned}
KL\left(q_t|\mathcal{N}(\bar{m}_t, V_t)\right) &\le \frac{1}{2}\log\left(\frac{\prod_{i=1}^d\left(m_t^2c_i+\sigma_t^2\right)}{(\sigma_t^2)^d}\right) + \frac{R^2m_t}{\sigma_t^2} \\
&\le \frac{dm_t^2c}{2\sigma_t^2} + \frac{R^2m_t}{\sigma_t^2} + o\left(\frac{m_t^2c}{\sigma_t^2}\right), \\
KL\left(\mathcal{N}(\bar{m}_t, V_t)|(\mathcal{N}(0,\sigma_t^2)\right) &\le \frac{m_t^2\sum_{i=1}^d c_i}{2\sigma_t^2} + \frac{m_t^2(\mathbb{E}[q_0])^2}{2\sigma_t^2} + \frac{1}{2}\log\left(\frac{(\sigma_t^2)^d}{\prod_{i=1}^d\left(m_t^2c_i+\sigma_t^2\right)}\right) \\
&\le \frac{m_t^2\sum_{i=1}^d c_i}{2\sigma_t^2} + \frac{m_t^2(\mathbb{E}[q_0])^2}{2\sigma_t^2} + \frac{dm_t^2c}{2\sigma_t^2} + o\left(\frac{m_t^2c}{\sigma_t^2}\right),
\end{aligned}
$$

where c_i *are the eigenvalues of* $\operatorname{Cov}\left[q_0\right]$*, and* c *is the eigenvalue with the largest absolute value.*

Proof. For $t \ge 0$, we directly calculate the KL divergence for this term:

$$
\begin{aligned}
\mathrm{KL}\left(q_t \mid \mathcal{N}(\bar{m}_t, V_t)\right) &= -H\left(q_t\right) + \frac{1}{2}\log\left(\det\left(2\pi V_t\right)\right) + \frac{1}{2}\operatorname{tr}\left(\left(m_t^2C+\sigma_t^2\mathbf{I}\right)V_t^{-1}\right) \\
&= -H\left(q_t\right) + \frac{1}{2}\log\left(\det\left(2\pi V_t\right)\right) + \frac{d}{2} \\
&= -H\left(q_t\right) + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log\left(\prod_{i=1}^d\left(m_t^2c_i+\sigma_t^2\right)\right) + \frac{d}{2}\,,
\end{aligned}
$$

where c_i are the eigenvalues of Cov $[q_0]$. Now, we only need to calculate $H(q_t)$:

$$-H\left(q_t\right)=\mathbb{E}_{X_t}\left[\log q_t\left(X_t\right)\right]=\mathbb{E}_{X_t}\left[\log\left(\mathbb{E}_{X_0}\left[(2\pi\sigma_t^2)^{-d/2}\exp\left(-\frac{1}{2\sigma_t^2}\left\|X_t-X_0\right\|^2\right)\right]\right)\right].$$

By **Assumption 2**, it is clear that

$$\exp\left(-\frac{1}{2\sigma_t^2}\left\|X_t-X_0\right\|^2\right)\leq\exp\left(-\frac{1}{2\sigma_t^2}\left(\left\|X_t\right\|^2+2\langle X_t,X_0\rangle\right)\right).$$

Then, we know that

$$\begin{aligned}&\mathbb{E}\left[\log\left(\mathbb{E}_{X_0}\left[(2\pi\sigma_t^2)^{-d/2}\exp\left(-\frac{1}{2\sigma_t^2}\left\|X_t-X_0\right\|^2\right)\right]\right)\right]\\ \leq&\mathbb{E}\left[\log\left((2\pi\sigma_t^2)^{-d/2}\right)-\frac{1}{2\sigma_t^2}\left(\left\|X_t\right\|^2+2\langle X_t,X_0\rangle\right)\right]\\ \leq&-\frac{d}{2}\log(2\pi)-\frac{1}{2}\log\left((\sigma_t^2)^d\right)-\frac{1}{2\sigma_t^2}\mathbb{E}\left[\left\|X_t\right\|^2\right]+\frac{R^2m_t}{\sigma_t^2}.\end{aligned}$$

we also know that

$$\mathbb{E}\left[\left\|X_t\right\|^2\right]=m_t^2\mathbb{E}\left[\left\|X_0\right\|^2\right]+\sigma_t^2\mathbb{E}\left[\|Z\|^2\right]=\mathbb{E}\left[\left\|X_0\right\|^2\right]+t\mathbb{E}\left[\|Z\|^2\right]=\bar{m}_0^2+V_0+\sigma_t^2 d.$$

Finally, put these terms together, we have:

$$\mathrm{KL}\left(q_t|\mathcal{N}(\bar{m}_t,V_t)\right)\leq\frac{1}{2}\log\left(\frac{\prod_{i=1}^d\left(m_t^2c_i+\sigma_t^2\right)}{(\sigma_t^2)^d}\right)+\frac{R^2m_t}{\sigma_t^2},$$

where c_i are the eigenvalues of Cov $[q_0]$. Then by choosing the largest absolute value eigenvalue largest absolute value, we can use the Taylor expansion to obtain the first results of this lemma. For the second result of this lemma, we directly compute the KL divergence between $\mathcal{N}(\bar{m}_t,V_t)$ and $\mathcal{N}(0,\sigma_t^2)$ to obtain the final results. ■

Theorem 1. *Let q_T be the marginal distribution of the forward process, and $q_\infty^\tau=\mathcal{N}(0,\sigma_T^2\mathbf{I})$ be the reverse beginning distribution. With m_T,σ_T defined in Eq. (2), we have*

$$\|q_T-q_\infty^\tau\|_{TV}\leq\sqrt{m_T}\bar{D}/\sigma_T,$$

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

Proof. We know that

$$\begin{aligned}&\|q_T-q_\infty^\tau\|_{\mathrm{TV}}\\ &\leq\|q_T-\mathcal{N}(m_T\mathbb{E}[q_0],m_T^2\mathrm{Cov}[q_0]+\sigma_T^2\mathbf{I})\|_{\mathrm{TV}}+\|\mathcal{N}(m_T\mathbb{E}[q_0],m_T^2\mathrm{Cov}[q_0]+\sigma_T^2\mathbf{I})-q_\infty^\tau\|_{\mathrm{TV}}.\end{aligned}$$

By directly using the Pinsker's inequality and Lemma 3, we complete the proof. ■

B The Proof of the Polynomial Complexity for Reverse SDE

In this section, we prove Corollary 1. First, we recall the Girsanov's Theorem (Le Gall, 2016) used in Chen et al. (2023d):

Lemma 4 (Girsanov's theorem)**.** *Let P_T and Q_T be two probability measures on path space $\mathcal{C}\left([0,T];\mathbb{R}^d\right)$. Suppose that under P_T, the process $(X_t)_{t\in[0,T]}$ follows*

$$\mathrm{d}X_t=\tilde{b}_t\,\mathrm{d}t+\alpha_t\,\mathrm{d}\tilde{B}_t$$

where $\tilde{B}$ is a P_T-Brownian motion, and under Q_T, the process $(X_t)_{t\in[0,T]}$ follows

$$\mathrm{d}X_t=b_t\,\mathrm{d}t+\alpha_t\,\mathrm{d}B_t$$

where B is a Q_T-Brownian motion. We assume that for each $t > 0, \alpha_t$ is a $d \times d$ symmetric positive definite matrix. Then, provided that Novikov's condition holds,

$$\mathbb{E}_{Q_T} \exp\left(\frac{1}{2}\int_0^T \left\|\alpha_t^{-1}\left(\tilde{b}_t - b_t\right)\right\|^2 \mathrm{d}t\right) < \infty,$$

we have that

$$\frac{\mathrm{d}P_T}{\mathrm{d}Q_T} = \exp\left(\int_0^T \alpha_t^{-1}\left(\tilde{b}_t - b_t\right)\mathrm{d}B_t - \frac{1}{2}\int_0^T \left\|\alpha_t^{-1}\left(\tilde{b}_t - b_t\right)\right\|^2 \mathrm{d}t\right).$$

If the Novikov's condition is satisfied, we apply the Girsanov theorem by choosing $P_T = R_K^{q_T^\tau}, Q_T = Q_{t_K}^{q_T^\tau}, \tilde{b}_t = \beta_{T-t}\left\{\frac{1}{\tau}\widetilde{\mathbf{Y}}_t + 2\mathbf{s}(T-t_k, \widetilde{\mathbf{Y}}_t)\right\}$ (for $t \in [t_k, t_{k+1}]$), $b_t = \beta_{T-t}\left\{\frac{1}{\tau}\mathbf{Y}_t + \left(1+\eta^2\right)\nabla \log q_{T-t}\left(\mathbf{Y}_t\right)\right\}$, and $\alpha_t = \sqrt{2\beta_{T-t}}\mathbf{I}_d$.

Then, similar to Chen et al. (2023d), we have the following lemma.

Lemma 5. *Assuming that $R_K^{q_T^\tau}$ and $Q_{t_K}^{q_T^\tau}$ satisfy Novikov's condition, it holds that*

$$\begin{aligned}&\mathrm{KL}\left(Q_{t_K}^{q_T^\tau} \| R_K^{q_T^\tau}\right)\\ &= \mathbb{E}_{Q_{t_K}^{q_T^\tau}} \ln\frac{\mathrm{d}Q_{t_K}^{q_T^\tau}}{\mathrm{d}R_K^{q_T^\tau}} = \sum_{k=0}^{K-1}\mathbb{E}_{Q_{t_K}^{q_T^\tau}}\int_{t_k}^{t_{k+1}} 2\beta_{T-t}\left\|\mathbf{s}\left(T-t_k, \mathbf{Y}_{t_k}\right) - \nabla \ln q_{T-t}\left(\mathbf{Y}_t\right)\right\|^2 \mathrm{d}t\,.\end{aligned}$$

Before using the Girsanov's Theorem, we need to check the Novikov's condition. The key proof of the Novikov's condition is Lemma 19 of Chen et al. (2023d). Since we assume the accurate score function in this work, this lemma need to control

$$\sup_{x^* \in \mathrm{B}(0,R), t^* \in [0, T-\delta]} 2\beta_{T-t^*}\left\|\nabla \ln q_{T-t^*}\left(x^*\right)\right\| =: B < \infty\,.$$

As we shown in Lemma 13, we know that with the early stopping parameter δ, $\left\|\nabla \ln q_{T-t^*}\left(x^*\right)\right\|$ is controlled. By using **Assumption 2**, we know that $\frac{1}{\beta_{T-t^*}} \leq \bar{\beta}$. Finally, with similar process to Chen et al. (2023d), we can proof that the Novikov's condition is satisfied. The following lemma show the discretization error for VESDE with reverse SDE.

Lemma 6 (Discretization)**.** *Suppose that* **Assumption 2** *and* **Assumption 3** *holds. Let $\bar{\gamma}_K = \mathrm{argmax}_{k\in\{0,\ldots,K-1\}}\gamma_k, \gamma_K = \delta$,*

(1) If $\tau = T^2$ and $\beta_t = t^2$, then with $Q_{t_K}^{q_T^\tau}$ and $R_K^{q_T^\tau}$ defined in Lemma 5,

$$\mathrm{TV}\left(R_K^{q_T^\tau}, Q_{t_K}^{q_T^\tau}\right)^2 \lesssim \frac{R^4 T^5 d}{\sigma_\delta^8}\bar{\gamma}_K + \frac{R^6 T^5}{\sigma_\delta^8}\bar{\gamma}_K^2 + \epsilon_{score}^2 T^3\,..$$

(2) If $\tau = T$ and $\beta_t = t$, we have

$$\mathrm{TV}\left(R_K^{q_T^\tau}, Q_{t_K}^{q_T^\tau}\right)^2 \lesssim \frac{R^4 T^3 d}{\sigma_\delta^8}\bar{\gamma}_K + \frac{R^6 T^3}{\sigma_\delta^8}\bar{\gamma}_K^2 + \epsilon_{score}^2 T^2\,.$$

Proof. First, we control the discretization error in an interval $t \in [t_k, t_{k+1}]$:

$$\begin{aligned}&\mathbb{E}_{Q_{t_K}^{q_T^\tau}}\left[\left\|\mathbf{s}\left(T-t_k, \mathbf{Y}_{t_k}\right) - \nabla \ln q_{T-t}\left(\mathbf{Y}_t\right)\right\|^2\right]\\ &\lesssim \epsilon_{\text{score}}^2 + \mathbb{E}_{Q_{t_K}^{q_T^\tau}}\left[\left\|\nabla \ln q_{T-t_k}\left(\mathbf{Y}_{t_k}\right) - \nabla \ln q_{T-t}\left(\mathbf{Y}_{t_k}\right)\right\|^2\right]\\ &\quad + \mathbb{E}_{Q_{t_K}^{q_T^\tau}}\left[\left\|\nabla \ln q_{T-t}\left(\mathbf{Y}_{t_k}\right) - \nabla \ln q_{T-t}\left(\mathbf{Y}_t\right)\right\|^2\right]\\ &\lesssim \mathbb{E}_{Q_{t_K}^{q_T^\tau}}\left[\left\|\nabla \ln\frac{q_{T-t_k}}{q_{T-t}}\left(\mathbf{Y}_{t_k}\right)\right\|^2\right] + L^2\mathbb{E}_{Q_{t_K}^{q_T^\tau}}\left[\left\|\mathbf{Y}_{t_k} - \mathbf{Y}_t\right\|^2\right] + \epsilon_{\text{score}}^2\\ &\lesssim (\tau + \beta_T)L^2 d\bar{\gamma}_K + \tau L^2\bar{\gamma}_K^2\left(d\tau + R^2\right) + \tau L^3\bar{\gamma}_K^2 + L^2(\beta_T d\bar{\gamma}_K + R^2\bar{\gamma}_K^2) + \epsilon_{\text{score}}^2\\ &\lesssim (\tau + \beta_T)L^2 d\bar{\gamma}_K + \tau L^2 R^2\bar{\gamma}_K^2 + \epsilon_{\text{score}}^2\,,\end{aligned}$$

where $L = \max_{t\in[0,T-\delta]} \left\|\nabla^2 \log q_{T-t}\left(\mathbf{Y}_t\right)\right\| \leq \left(1+R^2\right)/\sigma_\delta^4$ and the third inequality follows Lemma 17. Then, we know that for $\tau = T^2$ and $\beta_t = t^2$

$$\begin{aligned}
&\sum_{k=0}^{K-1} \mathbb{E}_{Q_{t_K}^{q_T^\tau}} \int_{t_k}^{t_{k+1}} 2\beta_{T-t} \left\|s\left(T-t_k, \mathbf{Y}_{t_k}\right) - \nabla \ln q_{T-t}\left(\mathbf{Y}_t\right)\right\|^2 \, \mathrm{d}t \\
&\lesssim T^5 L^2 d \bar{\gamma}_K + L^2 R^2 T^5 \bar{\gamma}_K^2 + \epsilon_{\text{score}}^2 T^3 \\
&\lesssim \frac{R^4 T^5 d}{\sigma_\delta^8} \bar{\gamma}_K + \frac{R^6 T^5}{\sigma_\delta^8} \bar{\gamma}_K^2 + \epsilon_{\text{score}}^2 T^3 \,.
\end{aligned}$$

For $\tau = T$ and $\beta_t = t$, we know that

$$\begin{aligned}
&\sum_{k=0}^{K-1} \mathbb{E}_{Q_{t_K}^{q_T^\tau}} \int_{t_k}^{t_{k+1}} 2\beta_{T-t} \left\|s\left(T-t_k, \mathbf{Y}_{t_k}\right) - \nabla \ln q_{T-t}\left(\mathbf{Y}_t\right)\right\|^2 \, \mathrm{d}t \\
&\lesssim T^3 L^2 d \bar{\gamma}_K + L^2 R^2 T^3 \bar{\gamma}_K^2 + \epsilon_{\text{score}}^2 T^2 \\
&\lesssim \frac{R^4 T^3 d}{\sigma_\delta^8} \bar{\gamma}_K + \frac{R^6 T^3}{\sigma_\delta^8} \bar{\gamma}_K^2 + \epsilon_{\text{score}}^2 T^2 \,.
\end{aligned}$$

■

Combined with the reversing beginning error controlled by Theorem 1, we can obtain the convergence guarantee for VESDE with reverse SDE.

Corollary 1. *Assume* **Assumption 1**, *2, 3. Let* $\gamma_K = \delta$, $\bar{\gamma}_K = \mathit{argmax}_{k\in\{0,\ldots,K-1\}} \gamma_k$, $\tau = T^2$ *and* $\beta_t = t^2$. *Then,* $TV\left(R_K^{q_\infty^\tau}, q_0\right)$ *is bounded by*

$$\mathrm{TV}\left(R_K^{q_\infty^\tau}, q_0\right) \leq \frac{\bar{D}\exp(-T/2)}{T} + \frac{R^2\sqrt{d}}{\delta^6}\sqrt{\bar{\gamma}_K T^5} + \epsilon_{score}\sqrt{T^3}\,,$$

where $\bar{D} = d|c| + \mathbb{E}\left[q_0\right] + R$. *Furthermore, by choosing* $\delta \leq \frac{\epsilon_{W_2}^{2/3}}{(d+R\sqrt{d})^{1/3}}$, $T \geq 2\ln\frac{\bar{D}}{\epsilon_{TV}}$, *maximum stepsize* $\bar{\gamma}_K \leq \delta^{12}\epsilon_{TV}^2 \ln^5(\bar{D}/\epsilon_{TV})/R^4 d$ *and assuming* $\epsilon_{score} \leq \widetilde{O}(\epsilon_{TV})$, *the output of* $R_K^{q_\infty^\tau}$ *is* $(\epsilon_{TV} + \epsilon_{score})$ *close to* q_δ, *which is* ϵ_{W_2} *close to* q_0, *with sample complexity (hiding logarithmic factors) is*

$$K \leq \widetilde{O}\left(\frac{dR^4(d+R\sqrt{d})^4}{\epsilon_{W_2}^8 \epsilon_{TV}^2}\right) .$$

For choice $\beta_t = t$ *and* $\tau = T$, *by choosing* $\delta \leq \frac{\epsilon_{W_2}}{(d+R\sqrt{d})^{1/2}}$ *and* $\bar{\gamma}_K \leq \frac{\delta^8\epsilon_{TV}^2\ln^3(\bar{D}/\epsilon_{TV})}{R^4 d}$, *we obtain the same sample complexity.*

Proof. By the data processing inequality, we know that

$$\begin{aligned}
\mathrm{TV}\left(R_K^{q_\infty^\tau}, q_0\right) &\leq \mathrm{TV}\left(R_K^{q_\infty^\tau}, R_K^{q_T^\tau}\right) + \mathrm{TV}\left(R_K^{q_T^\tau}, Q_{t_K}^{q_T^\tau}\right) \\
&\leq \mathrm{TV}\left(q_T^\tau, q_\infty^\tau\right) + \mathrm{TV}\left(R_K^{q_\infty^\tau}, Q_{t_K}^{q_T^\tau}\right) \,.
\end{aligned}$$

Then we have that for $\tau = T^2$ and $\beta_t = t^2$

$$\begin{aligned}
\mathrm{TV}\left(R_K^{q_\infty^\tau}, q_0\right) &\lesssim \frac{\bar{D}\exp(-T/2)}{T} + \frac{R^2\sqrt{d}}{\sigma_\delta^4}\sqrt{\bar{\gamma}_K T^5} + \epsilon_{\text{score}}\sqrt{T^3} \\
&\lesssim \frac{\bar{D}\exp(-T/2)}{T} + \frac{R^2\sqrt{d}}{\delta^6}\sqrt{\bar{\gamma}_K T^5} + \epsilon_{\text{score}}\sqrt{T^3}\,,
\end{aligned}$$

where $\bar{D} = d|c| + \mathbb{E}\left[q_0\right] + R$. The last inequality by the fact that Lemma 19. We can also use similar process to obtain the guarantee for $\tau = T$ and $\beta_t = t$

$$\mathrm{TV}\left(R_K^{q_\infty^\tau}, q_0\right) \lesssim \frac{\bar{D}\exp(-T/2)}{\sqrt{T}} + \frac{R^2\sqrt{d}}{\delta^4}\sqrt{\bar{\gamma}_K T^3} + \epsilon_{\text{score}}\sqrt{T^2}\,.$$

■

C THE PROOF OF THE CONVERGENCE GUARANTEE IN THE UNIFIED FRAMEWORK

In this work, we introduce an indicator $i \in \{1, 2\}$ for σ_{T-t_K}. We use $\tau = T^2$ as an example. When $\beta_t = t^2$ is aggressive, we choose $i = 1$, $\eta = 1$ and $\sigma_{T-t_K}^{-2}(i = 1) \leq \frac{1}{\tau} + \frac{\bar{\beta}}{\delta^3}$. When $\beta_t = t$ is conservative, we choose $i = 2$, $\eta \in [0, 1)$ and $\sigma_{T-t_K}^{-2}(i = 2) \leq \frac{1}{\tau} + \frac{\bar{\beta}}{\delta^2}$. In the proof process of Lemma 2, Lemma 7, Lemma 8 and Lemma 9, we ignore the indicator i since this lemma does not involve the specific value of $\sigma_{T-t_K}^2(i)$. Before the proof of this section, we first recall the stochastic flow of the reverse process for any $x \in \mathbb{R}^d$ and $s, t \in [0, T]$ with $t \geq s$:

$$\mathrm{d}\mathbf{Y}_{s,t}^x = \beta_{T-t}\left\{\mathbf{Y}_{s,t}^x/\tau + \left(1+\eta^2\right)\nabla \log q_{T-t}\left(\mathbf{Y}_{s,t}^x\right)\right\}\mathrm{d}t + \eta\sqrt{2\beta_{T-t}}\mathrm{d}\mathbf{B}_t, \qquad \mathbf{Y}_{s,s}^x = x\,,$$

and the interpolation of its discretization for any $k \in \{0, ..., K\}$ and $t \in [s_k, t_{k+1})$:

$$\mathrm{d}\bar{\mathbf{Y}}_{s,t}^x(k) = \beta_{T-t}\left\{\bar{\mathbf{Y}}_{s,t}^x/\tau + \left(1+\eta^2\right)\mathbf{s}\left(T-s_k, \bar{\mathbf{Y}}_{s,t}^x\right)\right\}\mathrm{d}t + \eta\sqrt{2\beta_{T-t}}\mathrm{d}\mathbf{B}_t, \qquad \bar{\mathbf{Y}}_{s,s}^x = x\,,$$

where $s_k = \max(s, t_k)$. To deal with the discretization error, we use the approximation technique used in Bortoli (2022). Hence, we introduce the tangent process:

$$\mathrm{d}\nabla\mathbf{Y}_{s,t}^x = \beta_{T-t}\left\{\mathbf{I}/\tau + \left(1+\eta^2\right)\nabla^2 \log q_{T-t}(\mathbf{Y}_{s,t}^x)\right\}\nabla\mathbf{Y}_{s,t}^x\mathrm{d}t, \qquad \nabla\mathbf{Y}_{s,s}^x = \mathbf{I}\,.$$

Then, we discuss the interpolation formula, which is used to control the discretization error.

Proposition 1. *For $s, t \in [0, T)$ with $s < t$, any $k \in \{0, ..., K\}$ and $(\omega_v)_{v\in[s,T]}$, we define that*

$$b_u(\omega) = \beta_{T-u}\left(\frac{1}{\tau}\omega_u + (1+\eta^2)\nabla \log q_{T-u}(\omega_u)\right),$$
$$\bar{b}_u(\omega) = \beta_{T-u}\left(\frac{1}{\tau}\omega_u + (1+\eta^2)\mathbf{s}(T-s_k, \omega_{s_k})\right),\ \Delta b_u(\omega) = b_u(\omega) - \bar{b}_u(\omega)\,,$$

where $s_k = \max(s, t_k)$ and $u \in [s_k, t_{k+1})$. Then, for any $x \in \mathbb{R}^d$, we have that

$$\mathbf{Y}_{s,t}^x - \bar{\mathbf{Y}}_{s,t}^x = \int_s^t \nabla\mathbf{Y}_{u,t}^x\left(\bar{\mathbf{Y}}_{s,u}^x\right)^\top \Delta b_u\left(\left(\bar{\mathbf{Y}}_{s,v}^x\right)_{v\in[s,T]}\right)\mathrm{d}u,$$

where for any $u \in [0, T)$, there exists a $k \in \{0, ..., K\}$ satisfies $u \in [s_k, t_{k+1})$.

For reverse SDE, the augmentation is similar to Bortoli (2022) (Appendix E). When $\eta = 0$, the stochastic extension of the Alekseev–Gröbner formula (Del Moral and Singh, 2022) degenerates into the original version (Alekseev, 1961). After that, we control the tangent process.

Lemma 2. *Assume* **Assumption 1** *and 2. Then, for $\forall s \in [0, t_K], x \in \mathbb{R}^d$, and $i \in \{1, 2\}$, we have*

$$\|\nabla\mathbf{Y}_{s,t_K,i}^x\| \leq \exp\left[\frac{R^2}{2\sigma_{T-t_K}^2(i)} + \frac{(1-\eta^2)}{2}\int_0^{t_K}\frac{\beta_{T-u}}{\tau}\mathrm{d}u\right].$$

Furthermore, if assuming $\left\|\nabla^2 \log q_t\left(x_t\right)\right\| \leq \Gamma/\sigma_t^2$, we have that

$$\|\nabla\mathbf{Y}_{s,t_K,i}^x\| \leq \sigma_{T-t_K}^{-(1+\eta^2)\Gamma}(i)\exp\left[\left(\left(1+\eta^2\right)\Gamma + 2\right)\int_0^{t_K}\frac{\beta_{T-u}}{\tau}\mathrm{d}u\right].$$

Proof. Using Eq. (7) and Lemma 13, we have

$$\begin{aligned}
&\mathrm{d}\left\|\nabla\mathbf{Y}_{s,t}^x\right\|^2 \\
&\leq 2\beta_{T-t}\left(\frac{1}{\tau}\left\|\nabla\mathbf{Y}_{s,t}^x\right\|^2 - \left(1+\eta^2\right)\left(1 - m_{T-t}^2R^2/\left(2\sigma_{T-t}^2\right)\right)/\sigma_{T-t}^2\left\|\nabla\mathbf{Y}_{s,t}^x\right\|^2\right)\mathrm{d}t\,.
\end{aligned}$$

Using Lemma 18, we have

$$\begin{aligned}
&\int_s^t \beta_{T-u}\left(\frac{1}{\tau} - \left(1+\eta^2\right)/\sigma_{T-u}^2 + \left(1+\eta^2\right)m_{T-u}^2R^2/2\sigma_{T-u}^4\right)\mathrm{d}u \\
&\leq \left(\left(1+\eta^2\right)R^2/4\right)\left(\sigma_{T-t}^{-2} - \sigma_{T-s}^{-2}\right) + \frac{1-\eta^2}{2}\int_s^t\frac{\beta_{T-u}}{\tau}\mathrm{d}u \\
&\leq \frac{\left(1+\eta^2\right)R^2}{4\sigma_{T-t}^2} + \frac{1-\eta^2}{2}\int_s^t\frac{\beta_{T-u}}{\tau}\mathrm{d}u\,.
\end{aligned}$$

Note that $\nabla \mathbf{Y}_{s,s} = \mathbf{I}$, we get

$$\|\nabla \mathbf{Y}_{s,t_K}^x\|^2 \leq \exp\left[\frac{\left(1+\eta^2\right) R^2}{2\sigma_{T-t}^2} + \left(1-\eta^2\right)\int_0^{t_K} \frac{\beta_{T-u}}{\tau} \mathrm{d}u\right].$$

When we assume $\left\|\nabla \log q_t^2\left(x_t\right)\right\| \leq \Gamma/\sigma_t^2$, we know that

$$\mathrm{d}\left\|\nabla \mathbf{Y}_{s,t}^x\right\|^2 \leq 2\beta_{T-t}\left(\frac{1}{\tau} - \frac{\left(1+\eta^2\right)\Gamma}{\sigma_{T-t}^2}\right)\left\|\nabla \mathbf{Y}_{s,t}^x\right\|^2 \mathrm{d}t\,.$$

Using Lemma 18, we have

$$\begin{aligned}
&2\int_s^t \beta_{T-u}/\sigma_{T-u}^2 \mathrm{d}u \\
&\leq \log\left(\exp\left[2\int_0^{T-s} \frac{\beta_{T-u}}{\tau}\,\mathrm{d}u\right] - 1\right) - \log\left(\exp\left[2\int_0^{T-t} \frac{\beta_{T-u}}{\tau}\,\mathrm{d}u\right] - 1\right) \\
&\leq \log\left(\sigma_{T-s}^2\right) - \log\left(\sigma_{T-t}^2\right) + \int_{T-t}^{T-s} \frac{\beta_u}{\tau}\,\mathrm{d}u\,.
\end{aligned}$$

Then we have

$$\|\nabla \mathbf{Y}_{s,t_K}^x\|^2 \leq \sigma_{T-t_K}^{-(1+\eta^2)\Gamma} \exp\left[\left(\left(1+\eta^2\right)\Gamma + 2\right)\int_0^{t_K} \frac{\beta_{T-u}}{\tau}\mathrm{d}u\right].$$

Thus we complete our proof. ■

After bounding the gradient of the tangent process, the remaining term is $\|\Delta b\|$:

$$\|\Delta b\| \leq \|\Delta^{(a,b)} b\| + \|\Delta^{(b,c)} b\| + \|\Delta^{(c,d)} b\|\,, \tag{8}$$

where $b^{(a)} = b$ and $b^{(d)} = \bar{b}$. Moreover,

$$\begin{aligned}
b_u^{(b)}(\omega) &= \beta_{T-u}(\frac{1}{\tau}\omega_u + (1+\eta^2)\nabla \log q_{T-s_k}(\omega_u))\,, \\
b_u^{(c)}(\omega) &= \beta_{T-u}(\frac{1}{\tau}\omega_u + (1+\eta^2)\nabla \log q_{T-s_k}(\omega_{s_k}))\,, \\
\Delta_b^{a,b} &= b^{(a)} - b^{(b)},\ \Delta_b^{b,c} = b^{(b)} - b^{(c)},\ \Delta_b^{c,d} = b^{(c)} - b^{(d)}\,.
\end{aligned}$$

We then control $\|\Delta^{(a,b)} b\|, \|\Delta^{(b,c)} b\|, \|\Delta^{(c,d)} b\|$ separately. In this section, $\|\Delta^{(c,d)} b\| = 0$ since we assume that the accurate score function is achieved. For $\left\|\Delta^{(a,b)} b_u(\omega)\right\|$, we have the following lemma.

Lemma 7. *For $s, u \in [0,T)$ such that $u \geq s$, $u \in [s_k, t_{k+1})$ and $\omega = (\omega_v)_{v\in[s,T]}$ we have*

$$\begin{aligned}
&\|\Delta^{(a,b)} b_u\left(\omega\right)\| \\
&\leq \left(1+\eta^2\right)\beta_{T-u} \sup_{v\in[T-u,T-t_k]} \left(\beta_v/\sigma_v^6\right)\left(2+R^2\right)\left(R + \|\omega_u\|\right)\gamma_k.
\end{aligned}$$

Proof. Without loss of generality, we assume $s \leq t_k$. Then

$$\begin{aligned}
\|\Delta^{(a,b)} b_u\left(\omega\right)\| &\leq \left(1+\eta^2\right)\beta_{T-u}\|\nabla \log q_{T-u}\left(\omega_u\right) - \nabla \log q_{T-t_k}\left(\omega_u\right)\| \\
&\leq \left(1+\eta^2\right)\beta_{T-u}\gamma_k \sup_{v\in[T-u,T-t_k]} \|\partial_v \nabla \log q_{T-v}\left(\omega_u\right)\|.
\end{aligned}$$

Then by Lemma 16, we have

$$\begin{aligned}
&\|\Delta^{(a,b)} b_u\left(\omega\right)\| \\
&\leq \left(1+\eta^2\right)\beta_{T-u} \sup_{v\in[T-u,T-t_k]} \left(\beta_v/\sigma_v^6\right)\left(2+R^2\right)\left(R + \|\omega_u\|\right)\gamma_k\,.
\end{aligned}$$

■

For $\left\|\Delta^{(b,c)} b_u(\omega)\right\|$, we have the following lemma.

Lemma 8. *For $s, u \in [0, T)$ such that $u \geq s$, $u \in [s_k, t_{k+1})$ and $\omega = (\omega_v)_{v\in[s,T]}$ we have*

$$\|\Delta^{(b,c)} b_u\left(\omega\right)\| \leq \left(1+\eta^2\right)\left(\beta_{T-u}/\sigma_{T-u}^4\right)\left(1+R^2\right)\|\omega_u - \omega_{s_k}\|\,.$$

Proof. Without loss of generality, we assume $s \leq t_k$. In this case $s_k = t_k$, Then

$$\begin{aligned}\|\Delta^{(b,c)} b_u\left(\omega\right)\| &\leq \left(1+\eta^2\right)\beta_{T-u}\|\nabla \log q_{T-t_k}\left(\omega_{t_k}\right) - \nabla \log q_{T-t_k}\left(\omega_u\right)\| \\ &\leq \left(1+\eta^2\right)\beta_{T-u}\sup_{v\in[u,T-t_k]}\|\nabla^2 \log q_{T-t_k}\left(\omega_v\right)\|\|\omega_u - \omega_{t_k}\|\,.\end{aligned}$$

Using Lemma 14, we have that

$$\|\Delta^{(b,c)} b_u\left(\omega\right)\| \leq \left(1+\eta^2\right)\left(\beta_{T-u}/\sigma_{T-u}^4\right)\left(1+R^2\right)\|\omega_u - \omega_{t_k}\|\,.$$

Then the proof is complete. ■

We need to control the reverse process when dealing with Δb. The following lemma shows an upper bound for the reverse Y_k.

Lemma 9. *Assume* **Assumption 1** ,**Assumption 2**, *and there exists $\delta > 0$ such that $\frac{\gamma_k \beta_{T-t_k}}{\sigma_{T-t_k}^2} \leq \delta \leq 1/28$ for any $k \in \{0, \cdots, K\}$, then we have*

$$\mathbb{E}[\|Y_k\|^2] \leq U(\tau) = \tau d + B(1/A + \delta)\,,$$

where

$$\begin{aligned}A &= 4\eta^2 + 2 - 2\delta - 4(1+\eta^2)(1+\delta)\mu R \\ B &= 4(1+\eta^2)R^2\delta + 2(1+\eta^2)(1+\delta)\frac{R}{\mu} + 4\eta^2\tau d\end{aligned}$$

and μ is an arbitrary positive number which makes $A > 0$. In particular, if $\delta \leq 1/28$, then

$$\mathbb{E}[\|Y_k\|^2] \leq U_0(\tau) = 111R^2 + 13\tau d\,.$$

Proof. Recall the discretization of the backward process (the explicit form of Eq. (6))

$$\begin{aligned}Y_{k+1} &= Y_k + \gamma_{1,k}\left(\frac{1}{\tau}Y_k + (1+\eta^2)\mathbf{s}\left(T-t_k, Y_k\right)\right) + \eta\sqrt{2\gamma_{2,k}}Z_k\,, \\ \gamma_{1,k} &= \exp\left[\int_{T-t_{k+1}}^{T-t_k}\beta_s\ \mathrm{d}s\right] - 1, \quad \gamma_{2,k} = \left(\exp\left[2\int_{T-t_{k+1}}^{T-t_k}\beta_s\ \mathrm{d}s\right] - 1\right)/2\,,\end{aligned}$$

where $\{Z_k\}_{k\in K}$ are independent Gaussian random variables. It is clear that $\gamma_{1,k} \leq \gamma_{2,k} \leq 2\gamma_{1,k}$, and using Lemma 13 we have

$$\begin{aligned}\langle x_t, \mathbf{s}(t, x_t)\rangle &= \langle x_t, \nabla \log q_t(x_t)\rangle \\ &\leq -\|x_t\|^2/\sigma_t^2 + m_t R\|x_t\|/\sigma_t^2 \\ &\leq (-1 + \mu m_t R)\|x_t\|^2/\sigma_t^2 + (m_t R/\mu)/\sigma_t^2\,,\end{aligned}$$

where the first equality follows that we assume the accurate score function. For any $\mu > 0$. Again using Lemma 13, we have

$$\begin{aligned}\|\mathbf{s}(t, x_t)\|^2 &= \|\nabla \log q_t(x_t)\|^2 \\ &\leq 2\|x_t\|^2/\sigma_t^4 + 2m_t^2 R^2/\sigma_t^4\,.\end{aligned}$$

Combining the results above, we have

$$\begin{aligned}\mathbb{E}[\|Y_{k+1}\|^2] &= (1+\frac{\gamma_{1,k}}{\tau})^2\mathbb{E}[\|Y_k\|^2] + (1+\eta^2)^2\gamma_{1,k}^2\mathbb{E}[\|s(T-t_k, Y_k)\|^2] \\ &\quad + 2(1+\eta^2)(1+\frac{\gamma_{1,k}}{\tau})\gamma_{1,k}\mathbb{E}[\langle Y_k, s(T-t_k, Y_k)\rangle] + 2\eta^2\gamma_{2,k}d \\ &\leq ((1+\frac{\gamma_{1,k}}{\tau})^2 + 2(1+\eta^2)^2\gamma_{1,k}^2/\sigma_{T-t_k}^4 \\ &\quad + 2(1+\eta^2)(1+\frac{\gamma_{1,k}}{\tau})\gamma_{1,k}(-1+\mu m_{T-t_k}R)/\sigma_{T-t_k}^2)\mathbb{E}[\|Y_k\|^2] \\ &\quad + \frac{2m_{T-t_k}^2 R^2}{\sigma_{T-t_k}^4}(1+\eta^2)^2\gamma_{1,k}^2 + \frac{m_{T-t_k}R}{\mu\sigma_{T-t_k}^2}(1+\eta^2)(1+\frac{\gamma_{1,k}}{\tau})\gamma_{1,k} + 4\eta^2\gamma_{1,k}d\,.\end{aligned}$$

If we denote $\delta_k = \gamma_{1,k}/\sigma^2_{T-t_k}$ and notice the fact that $m_t \in [0,1], \sigma_t^2 \in [0,\tau], \eta \in [0,1]$, then we have

$$\mathbb{E}[\|Y_{k+1}\|^2] \leq (1 + 2\delta_k + \delta_k^2)\mathbb{E}[\|Y_k\|^2] + 8\delta_k^2\mathbb{E}[\|Y_k\|^2] + 2(1+\delta_k)\delta_k(-1+\mu R)\mathbb{E}[\|Y_k\|^2] + 8R^2\delta_k^2 + \frac{2R}{\mu}\delta_k(1+\delta_k) + 4\tau\delta_k d\,.$$

We also have that

$$\gamma_{1,k} = \exp[\int_{T-t_{k+1}}^{T-t_k} \beta_s \mathrm{d}s] - 1 \leq \exp[\beta_{T-t_k}\gamma_k] - 1 \leq 2\beta_{T-t_k}\gamma_k\,,$$

where the last inequality follows that $\gamma_k = \exp(-T)$, $\beta_{T-t_k}\gamma_k \leq 1/2$ for small enough stepsize, and $e^\omega - 1 \leq 2\omega$ for any $\omega \in [0,1/2]$. We get $\delta_k \leq 2\gamma_k\beta_{T-t_k}/\sigma^2_{T-t_k} \leq 2\delta$. Thus

$$\mathbb{E}[\|Y_{k+1}\|^2] \leq (1 + 2\delta_k + 2\delta_k\delta)\mathbb{E}[\|Y_k\|^2] + 16\delta_k\delta\mathbb{E}[\|Y_k\|^2] + 4(1+\delta)(-1+\mu R)\delta_k\mathbb{E}[\|Y_k\|^2] + 16R^2\delta_k\delta + 4(1+\delta)\frac{R}{\mu}\delta_k + 4\tau d\delta_k\,.$$

Hence, we have

$$\mathbb{E}[\|Y_{k+1}\|^2] \leq (1 + \delta_k[-2 + 14\delta + 4(1+\delta)\mu R])\mathbb{E}[\|Y_k\|^2] + \delta_k[16R^2\delta + 4(1+\delta)\frac{R}{\mu} + 4\tau d]\,.$$

We denote $A = 2 - 14\delta - 4(1+\delta)\mu R$ and $B = 16R^2\delta + 4(1+\delta)\frac{R}{\mu} + 4\tau d$, then

$$\mathbb{E}[\|Y_{k+1}\|^2] \leq (1 - \delta_k A)\mathbb{E}[\|Y_k\|^2] + \delta_k B\,.$$

Notice that $\mathbb{E}[\|Y_0\|^2] = d\tau$ and if $\mathbb{E}[\|Y_k\|^2] \geq B/A$ it is decreasing, if $\mathbb{E}[\|Y_k\|^2] \leq B/A$ we have $\mathbb{E}[\|Y_{k+1}\|^2] \leq B/A + \delta B$. so

$$\mathbb{E}[\|Y_k\|^2] \leq \tau d + B(1/A + \delta)\,.$$

Notice that when $\delta \leq 1/28$,if we choose $\mu = 1/(4(1+\delta)R)$, $A \geq 1/2$, and

$$B \leq 37R^2 + 4\tau d\,.$$

Then, the proof is complete. ■

The following lemma shows a discretization error in the k-the interval.

Lemma 10. *Assume* **Assumption 1**,**Assumption 2** *and* $\gamma_k\beta_{T-t_k}/\sigma^2_{T-t_k} \leq 1/28$ *for any* $k \in \{0,\cdots,K-1\}$. *Then for any* k, $t \in [t_k, t_{k+1}]$ *and* $i \in \{1,2\}$, *we have that*

$$\mathbb{E}[\|\bar{\mathbf{Y}}_t - \bar{\mathbf{Y}}_{t_k}\|^2] \leq L_i(\tau)\beta_{T-t_k}\gamma_k\,,$$

where $L_i(\tau) = \bar{\gamma}_K\kappa_i(\tau)(\frac{64}{\sigma^2_{T-t_K}(i)} + \frac{8}{\tau})U_0(\tau) + 64R^2\frac{\bar{\gamma}_K\kappa_i(\tau)}{\sigma^2_{T-t_K}(i)} + 4d$, $\bar{\gamma}_K$, $\kappa_i(\tau)$ *is defined in Lemma 11 and* $U_0(\tau)$ *is defined in Lemma 9.*

Proof. Recall the discretized backward process

$$\bar{\mathbf{Y}}_t = \bar{\mathbf{Y}}_{t_k} + (\exp[\int_{T-t}^{T-t_k}\beta_s\mathrm{d}s] - 1)(\frac{1}{\tau}\bar{\mathbf{Y}}_{t_k} + (1+\eta^2)\mathbf{s}(T-t_k,\bar{\mathbf{Y}}_{t_k})) + \eta(\exp[2\int_{T-t}^{T-t_k}\beta_s\mathrm{d}s - 1])^{1/2}Z\,,$$

where Z is a standard Gaussian random variable. By directly calculating, we have that

$$\mathbb{E}[\|\bar{\mathbf{Y}}_t - \bar{\mathbf{Y}}_{t_k}\|^2] = 2(\exp[\int_{T-t}^{T-t_k}\beta_s\mathrm{d}s] - 1)^2(\frac{1}{\tau^2}\mathbb{E}[\|\bar{\mathbf{Y}}_{t_k}\|^2] + (1+\eta^2)^2\mathbb{E}[\|s(T-t_k,\bar{\mathbf{Y}}_{t_k})\|^2]) + \eta^2(\exp[2\int_{T-t}^{T-t_k}\beta_s\mathrm{d}s] - 1)d\,.$$

By Lemma 13 and accurate score function assumption,

$$\|\mathbf{s}(T-t_k,\bar{\mathbf{Y}}_{t_k})\|^2 \le 2\|\bar{\mathbf{Y}}_{t_k}\|^2/\sigma^4_{T-t_k}(i) + 2m^2_{T-t_k}R^2/\sigma^4_{T-t_k}(i)\,.$$

So we have that

$$\mathbb{E}[\|\bar{\mathbf{Y}}_t - \bar{\mathbf{Y}}_{t_k}\|^2] \le 2(\exp[\int_{T-t}^{T-t_k}\beta_s \mathrm{d}s]-1)^2((\frac{8}{\sigma^4_{T-t_k}(i)}+\frac{1}{\tau^2})\mathbb{E}[\|\bar{\mathbf{Y}}_{t_k}\|^2] + \frac{8R^2}{\sigma^4_{T-t_k}(i)}) + (\exp[2\int_{T-t}^{T-t_k}\beta_s \mathrm{d}s]-1)d\,.$$

By $\mathrm{e}^{2w}-1 \le 1+4w$ for any $w \in [0,1/2]$ and $\gamma_k \sup_{v\in[T-t_{k+1},T-t_k]}\beta_v/\sigma_v^2 \le 1/28$ for any $k \in \{0,...,K-1\}$, we have

$$\exp[\rho\int_{T-t}^{T-t_k}\beta_s\mathrm{d}s]-1 \le 2\rho\beta_{T-t_k}\gamma_k.$$

for $\rho = 1,2$. And using Lemma 9 and Lemma 19 we have

$$\mathbb{E}[\|\bar{\mathbf{Y}}_t - \bar{\mathbf{Y}}_{t_k}\|^2] \le (\frac{64\gamma_k}{\sigma^4_{T-t_k}(i)} + \frac{8\beta_{T-t_k}\gamma_k}{\tau^2})U_0(\tau)\beta_{T-t_k}\gamma_k + 64R^2\frac{\gamma_k}{\sigma^4_{T-t_k}(i)}\beta_{T-t_k}\gamma_k + 4d\beta_{T-t_k}\gamma_k\,.$$

We denote $L_i(\tau) = \bar{\gamma}_K\kappa_i(\tau)(\frac{64}{\sigma^2_{T-t_K}(i)} + \frac{8}{\tau})U_0(\tau) + 64R^2\frac{\bar{\gamma}_K\kappa_i(\tau)}{\sigma^2_{T-t_K}(i)} + 4d$ for $i \in \{1,2\}$ and the proof is complete. ■

Lemma 11. *Assume* **Assumption 1** *and* **Assumption 2***,* $\gamma_k \sup_{v\in[T-t_{k+1},T-t_k]}\beta_v/\sigma_v^2 \le 1/28$ *for any* $k \in \{0,...,K-1\}$*. Let* $\bar{\gamma}_K = argmax_{k\in\{0,...,K-1\}}\gamma_k$*,* $\kappa_i(\tau) = \max\{\bar{\beta}, \frac{T^2}{T^{-1+i}}\}\sigma^{-2}_{T-t_K}(i)$*, and*

$$C_i(\tau) = 2(2+R^2)(R+U_0^{1/2}(\tau)) + 2L_i^{1/2}(\tau)\tau^{3/2}(1+R^2)\,,$$

for $i \in \{1,2\}$*. Then, for any* $s,u \in [0,t_K]$ *with* $u \ge s$ *and* $i \in \{1,2\}$*, we have*

$$\mathbb{E}[\|\Delta b_{u,i}((\bar{\mathbf{Y}}_{s,v})_{v\in[s,T]})\|] \le C_i(\tau)[\kappa_i^2(\tau)\sigma^{-2}_{T-t_K}(i)\bar{\gamma}_K^{1/2} + \kappa_i^2(\tau)]\bar{\gamma}_K^{1/2},$$

where $\bar{\mathbf{Y}}_{s,s} \sim \mathrm{N}(0,\mathbf{I})$.

Proof. Combining Lemma 7, Lemma 8 and the exact score function, we get

$$\|\Delta b_{u,i}(\omega)\| \le (1+\eta^2)\sup_{v\in[T-t_{k+1},T-t_k]}(\beta_v^2/\sigma_v^6(i))(2+R^2)(\mathrm{diam}(\mathcal{M}+\|\omega_u\|))\gamma_k + (1+\eta^2)(\beta_{T-u}/\sigma^4_{T-u}(i))(1+\mathrm{diam}(\mathcal{M}^2))\|\omega_u - \omega_{s_k}\|.$$

For any $u \in [T-t_K, T]$, using Lemma 20 we have $\beta_u/\sigma_u^2(i) \le \kappa_i(\tau)$. Hence,

$$\begin{aligned}\|\Delta b_{u,i}(\omega)\| &\le (1+\eta^2)\sup_{v\in[T-t_{k+1},T-t_k]}(\beta_v^2/\sigma_v^6(i))(2+\mathrm{diam}(\mathcal{M}^2))(R+\|\omega_u\|)\gamma_k \\ &\quad + (1+\eta^2)(\beta_{T-u}/\sigma^4_{T-u}(i))(1+\mathrm{diam}(\mathcal{M}^2))(\|\omega_u-\omega_{t_k}\|) \\ &\le (1+\eta^2)(\kappa_i^2(\tau)/\sigma^2_{T-t_{k+1}}(i))\gamma_k(2+\mathrm{diam}(\mathcal{M}^2))(R+\|\omega_u\|) \\ &\quad + (1+\eta^2)\kappa_i^2(\tau)(1+R^2)\|\omega_u-\omega_{t_k}\|/\beta_{T-u}.\end{aligned}$$

Combining this with Lemma 9 and Lemma 10,

$$\mathbb{E}[\|\Delta b_{u,i}((\bar{\mathbf{Y}}_{s,v})_{v\in[s,T]})\|] \le (1+\eta^2)(\kappa_i^2(\tau)/\sigma^2_{T-t_{k+1}}(i))\bar{\gamma}_K(2+R^2)(R+U_0^{1/2}(\tau)) + (1+\eta^2)\kappa_i^2(\tau)(1+R^2)L_i^{1/2}(\tau)\max\{\bar{\beta},\tau\}^{3/2}\bar{\gamma}_K^{1/2}.$$

We denote $C_i(\tau) = 2(2+R^2)(R+U_0^{1/2}(\tau)) + 2L_i^{1/2}(\tau)\tau^{3/2}(1+R^2)$, for $i \in \{1,2\}$, then we have

$$\mathbb{E}[\|\Delta b_{u,i}((\bar{\mathbf{Y}}_{s,v})_{v\in[s,T]})\|] \le C_i(\tau)((\kappa_i^2(\tau)/\sigma^2_{T-t_{k+1}})\bar{\gamma}_K + \kappa_i^2(\tau)\bar{\gamma}_K^{1/2}).$$

■

Lemma 12. *Assume* **Assumption 1** *and* **Assumption 2**, $\gamma_k \sup_{v\in[T-t_{k+1},T-t_k]} \beta_v/\sigma_v^2 \leq 1/28$ *for any* $k \in \{0,...,K-1\}$. *Let* $\bar{\gamma}_K = argmax_{k\in\{0,\ldots,K-1\}}\gamma_k$, $\gamma_K = \delta$, *and* $\delta \leq 1/32$. *Then*

$$W_1\left(R_K^{q_\infty^\tau}, Q_{t_K}^{q_\infty^\tau}\right) \leq C_i(\tau)\kappa_i^2(\tau)T \exp\left[\frac{R^2}{2\sigma_{T-t_K}^2(i)} + \frac{(1-\eta^2)}{2}\right][\frac{\bar{\gamma}_K^{1/2}}{\sigma_{T-t_K}^2(i)} + 1]\bar{\gamma}_K^{1/2},$$

where $C_i(\tau), \kappa_i(\tau)$ *for* $i \in \{1,2\}$ *are the same terms to Theorem 2.*

Proof. By **Proposition 1** we have

$$\|\mathbf{Y}_{t_K} - Y_K\| = \|\mathbf{Y}_{t_K} - \bar{\mathbf{Y}}_{t_K}\| \leq \int_0^{t_K} \|\nabla \mathbf{Y}_{u,t_K,i}(\bar{\mathbf{Y}}_{0,u})\|\|\Delta b_{u,i}((\bar{\mathbf{Y}}_{0,v})_{v\in[0,T]})\|\mathrm{d}u.$$

$$\begin{aligned}&\|\mathbf{Y}_{t_K} - Y_K\|\\ &\leq \exp\left[\frac{\left(1+\eta^2\right)R^2}{4\sigma_{T-t}^2(i)} + \frac{(1-\eta^2)}{2}\int_0^{t_K}\frac{\beta_{T-u}}{\tau}\mathrm{d}u\right]\int_0^{t_K}\|\Delta b_{u,i}((\bar{\mathbf{Y}}_{0,v})_{v\in[0,T]})\|\mathrm{d}u\,.\end{aligned}$$

Then by definition of Wasserstein distance, we have

$$\begin{aligned}&W_1(q_\infty Q_{t_K}, q_\infty R_K)\\ &\leq \mathbb{E}[\|\mathbf{Y}_{t_K} - Y_K\|]\\ &\leq \exp\left[\frac{\left(1+\eta^2\right)R^2}{4\sigma_{T-t_K}^2(i)} + \frac{(1-\eta^2)}{2}\int_0^{t_K}\frac{\beta_{T-u}}{\tau}\mathrm{d}u\right]\int_0^{t_K}\mathbb{E}[\|\Delta b_{u,i}((\bar{\mathbf{Y}}_{0,v})_{v\in[0,T]})\|]\mathrm{d}u\\ &\leq C_i(\tau)T\exp\left[\frac{\left(1+\eta^2\right)R^2}{4\sigma_{T-t_K}^2(i)} + \frac{(1-\eta^2)}{2}\right][\kappa_i^2(\tau)\sigma_{T-t_K}^{-2}(i)\bar{\gamma}_K^{1/2} + \kappa_i^2(\tau)]\bar{\gamma}_K^{1/2}\,.\end{aligned}$$

■

Theorem 2. *Assume* **Assumption 1** *and* **2**, $\gamma_k \sup_{v\in[T-t_{k+1},T-t_k]} \beta_v/\sigma_v^2 \leq 1/28$ *for* $\forall k \in \{0,...,K-1\}$, *and* $\delta \leq 1/32$. *Let* $\bar{\gamma}_K = argmax_{k\in\{0,\ldots,K-1\}}\gamma_k$ *and* $\gamma_K = \delta$. *Then, for* $\forall\tau \in [T, T^2]$, *we have the following convergence guarantee.*

(1) If $\eta = 1$ *(the reverse SDE), choosing an aggressive* $\beta_t = t^2$, *we have*

$$\begin{aligned}W_1\left(R_K^{q_\infty^\tau}, q_0\right) \leq\;& C_1(\tau)T\exp\left[\frac{R^2}{2}(\frac{\bar{\beta}}{\delta^3} + \frac{1}{\tau})\right][\kappa_1^2(\tau)(\frac{\bar{\beta}}{\delta^3} + \frac{1}{\tau})\bar{\gamma}_K^{1/2} + \kappa_1^2(\tau)]\bar{\gamma}_K^{1/2}\\ &+ \exp\left[\frac{R^2}{2}(\frac{\bar{\beta}}{\delta^3} + \frac{1}{\tau})\right]\frac{\bar{D}\exp\left(-T/2\right)}{\sqrt{\tau}} + 2(\frac{R}{\tau} + \sqrt{d})\sqrt{\delta}\,,\end{aligned}$$

where $C_1(\tau)$ *is linear in* τ^2, $\kappa_1(\tau) = \max\left\{\bar{\beta}, T^2\right\}\left(1/\tau + \bar{\beta}/\delta^3\right)$.

(2) If $\eta = 0$ *(the reverse PFODE), choosing a conservative* β_t *satisfies* **Assumption 1**, *we have*

$$\begin{aligned}W_1\left(R_K^{q_\infty^\tau}, q_0\right) \leq\;& C_2(\tau)T\exp\left[\frac{R^2}{2}(\frac{\bar{\beta}}{\delta^2} + \frac{1}{\tau}) + \frac{1}{2}\right][\kappa_2^2(\tau)(\frac{\bar{\beta}}{\delta^2} + \frac{1}{\tau})\bar{\gamma}_K^{1/2} + \kappa_2^2(\tau)]\bar{\gamma}_K^{1/2}\\ &+ \exp\left[\frac{R^2}{2}(\frac{\bar{\beta}}{\delta^2} + \frac{1}{\tau})\right]\frac{\bar{D}}{\sqrt{\tau}} + 2(\frac{R}{\tau} + \sqrt{d})\sqrt{\delta}\,,\end{aligned}$$

where $C_2(\tau)$ *is linear in* τ^2, $\kappa_2(\tau) = \max\{\bar{\beta}, T\}\left(1/\tau + \bar{\beta}/\delta^2\right)$.

Proof. To obtain the convergence guarantee, we need to control three error terms:

$$W_1\left(R_K^{q_\infty^\tau}, q_0\right) \leq W_1\left(R_K^{q_\infty^\tau}, Q_{t_K}^{q_\infty^\tau}\right) + W_1\left(Q_{t_K}^{q_\infty^\tau}, Q_{t_K}^{q_0 P_T}\right) + W_1\left(Q_{t_K}^{q_0 P_T}, q_0\right)\,.$$

For term $W_1\left(R_K^{q_\infty^\tau}, Q_{t_K}^{q_\infty^\tau}\right)$, we use Lemma 12.

For the second term, we define $\left(\mathbf{Y}_{0,t}^x\right)_{t\in[0,T]}$ and $\left(\mathbf{Y}_{0,t}^y\right)_{t\in[0,T]}$ be the reverse processes with initial condition x and y. Then we have

$$\|\mathbf{Y}_{0,t}^x - \mathbf{Y}_{0,t}^y\| \leq \|x-y\| \int_0^1 \|\nabla \mathbf{Y}_{0,t}^{z_\lambda}\| d\lambda \,,$$

where $z_\lambda = \lambda x + (1-\lambda)y$. In this work, we choose $x \sim q_\infty^\tau$ and $y \sim q_0 P_T$. Combined with the above inequality, Theorem 1 and Lemma 2, we know that:

$$\begin{aligned}
&W_1\left(Q_{t_K}^{q_\infty^\tau}, Q_{t_K}^{q_0 P_T}\right) \\
&\leq \exp\left[\frac{R^2}{2\sigma_{T-t_K}^2(i)} + \frac{(1-\eta^2)}{2}\int_0^{t_K} \frac{\beta_{T-u}}{\tau}\mathrm{d}u\right] \|q_0 P_T - q_\infty^\tau\| \\
&\leq \frac{\sqrt{m_T}\bar{D}}{\sigma_T} \exp\left[\frac{R^2}{2\sigma_{T-t_K}^2(i)} + \frac{(1-\eta^2)}{2}\int_0^{t_K} \frac{\beta_{T-u}}{\tau}\mathrm{d}u\right] .
\end{aligned}$$

For the last term, we use exactly the same process with Bortoli (2022) with bounded $\sigma_{T-t_K}^2$:

$$\begin{aligned}
W_1\left(Q_{t_K}^{q_0 P_T}, q_0\right) &\leq \mathbb{E}\left[\|X - m_{T-t_K}X + \sigma_{T-t_K} Z\|\right] \\
&\leq (\frac{R}{\tau} + \sqrt{d})\sigma_{T-t_K} \\
&\leq 2(\frac{R}{\tau} + \sqrt{d})\sqrt{\delta} \,,
\end{aligned}$$

where the second inequality follows that $\sigma_{T-t_K}^2 + \tau m_{T-t_K} = \tau$. ■

In the end of the section, we provide the proof of Corollary 3.

Corollary 3. *Assume* **Assumption 1**, **Assumption 2** *and* $\left\|\nabla^2 \log q_t\left(x_t\right)\right\| \leq \Gamma/\sigma_t^2$. *Let* $\eta = 0$ *(reverse PFODE),* $\epsilon \in (0, 1/32), \tau = T^2$, $\beta_t = t$, $\bar{\gamma}_K = \mathit{argmax}_{k\in\{0,\ldots,K-1\}}\gamma_k, \gamma_K = \delta$*, we have*

$$\begin{aligned}
W_1\left(R_K^{q_\infty^\tau}, q_0\right) \leq C_2(\tau) T \frac{\bar{\beta}^{\frac{\Gamma}{2}}}{\delta^\Gamma} \exp\left[\frac{\Gamma+2}{2}\right] [\kappa_2^2(\tau)(\frac{\bar{\beta}}{\delta^2} + \frac{1}{\tau})\bar{\gamma}_K^{1/2} + \kappa_2^2(\tau)]\bar{\gamma}_K^{1/2} \\
+ \frac{\bar{\beta}^{\frac{\Gamma}{2}}}{\delta^\Gamma} \exp\left[\frac{\Gamma+2}{2}\right] \frac{\bar{D}}{\sqrt{\tau}} + 2(\frac{R}{\tau} + \sqrt{d})\sqrt{\delta} \,,
\end{aligned}$$

where $C_2(\tau)$ *is linear in* τ^2, $\kappa_2(\tau) = \max\{\bar{\beta}, T\}\left(\frac{1}{\tau} + \frac{\bar{\beta}}{\delta^2}\right)$.

Proof. The proof of this corollary is almost identical to the proof of Theorem 2. We just need to replace the first bound for the tangent process in Lemma 2 by the second bound. ■

D LEMMAS FOR THE LOGARITHMIC DENSITY

In this section, we introduce auxiliary lemmas to control the gradient and Hessian of the logarithmic density under the manifold hypothesis. Lemma 13, Lemma 14 and Lemma 15 come from Lemma C.1, Lemma C.2, and Lemma C.5 of Bortoli (2022). Since these lemmas do not involve the relationship between m_t and σ_t, we can directly use the results from Bortoli (2022). Following Bortoli (2022), we also define a empirical version of q_0 with N datapoints, i.e. $q_0^N = (1/N)\sum_{k=1}^N X^k$, with $\left\{X^k\right\}_{k=1}^N \sim q_0^{\otimes N}$. We denote by $\left(q_t^N\right)_{t>0}$ such that for any $t>0$ the density w.r.t. the Lebesgue measure of the distribution of $\mathbf{X}_t^N$, and when $N \to +\infty$, $q_t^N = q_t$.

Lemma 13. *Assume* **Assumption 2**. *Then for any* $t \in (0, T]$ *and* $x_t \in \mathbb{R}^d$ *we have that*

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

In addition, we have

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

Lemma 14. *Assume* **Assumption 2**. *Then for any* $t \in (0, T]$, $x_t \in \mathbb{R}^d$ *and* $M \in \mathcal{M}_d\left(\mathbb{R}^d\right)$

$$\left\langle M, \nabla^2 \log q_t\left(x_t\right) M\right\rangle \leq -\left(1-m_t^2 R^2 /\left(2 \sigma_t^2\right)\right) / \sigma_t^2\|M\|^2 .$$

In addition, we have

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

The following lemma shows that the derivatives up to the fourth order are uniformly bounded since $\tau \in [T, T^2]$. Thus we can use the stochastic extension of the Alekseev–Gröbner formula (Del Moral and Singh, 2022).

Lemma 15. *Assume* **Assumption 2**. *Then, there exists* $\bar{C} \geq 0$ *such that for any* $t \in (0, T]$ *we have*

$$\left\|\nabla^2 \log q_t(x)\right\|+\left\|\nabla^3 \log q_t(x)\right\|+\left\|\nabla^4 \log q_t(x)\right\| \leq \bar{C} / \sigma_t^8 .$$

The following lemma shows that $\|\partial_t \nabla \log q_t\left(x_t\right)\|$ is bounded. The proof before using the relationship between σ_t and m_t is identical compared to Lemma C.3 in Bortoli (2022). For the sake of completeness, we also give the proof process of this part.

Lemma 16. *Assume* **Assumption 2**. *Then for any* $t \in (0, T]$ *and* $x_t \in \mathbb{R}^d$ *we have*

$$\|\partial_t \nabla \log q_t\left(x_t\right)\| \leq\left(\beta_t / \sigma_t^6\right)\left(2+R^2\right)\left(R+\left\|x_t\right\|\right) .$$

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

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

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

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

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 $\nabla f_t^k(x)=-\left(x-m_t X^k\right) / \sigma_t^2$, and using $m_t \leq 1$ we get

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

and

$$\partial_t f_t^k(t)=\partial_t \sigma_t^2 /\left(2 \sigma_t^4\right)\left\|x-m_t X^k\right\|^2+\partial_t m_t / \sigma_t^2\left\langle X^k, x-m_t X^k\right\rangle .$$

Notice the fact that $\partial_t \sigma_t^2 = -2\tau m_t \partial_t m_t = 2\beta_t m_t^2$ and $\partial_t m_t = -\frac{\beta_t}{\tau} m_t$, combined with the above equality, we know that

$$\begin{aligned}
\partial_t f_t^k(t) &= -\beta_t m_t/\sigma_t^2 \left[-\left(m_t/\sigma_t^2\right) \|x - m_t X^k\|^2 + \frac{1}{\tau}\left\langle x - m_t X^k, X^k \right\rangle \right] \\
&= -\beta_t m_t/\sigma_t^2 \left\langle x - m_t X^k, -\left(m_t/\sigma_t^2\right)\left(x - m_t X^k\right) + \frac{1}{\tau} X^k \right\rangle \\
&= -\beta_t m_t/\sigma_t^4 \left\langle x - m_t X^k, -m_t x + \left(m_t^2 + \frac{\sigma_t^2}{\tau}\right) X^k \right\rangle \\
&= \beta_t m_t/\sigma_t^4 \left(m_t \|x\|^2 + m_t \left\|X^k\right\|^2 + \left(1 + m_t^2\right)\left\langle x, X^k \right\rangle \right),
\end{aligned}$$

where the last equality holds that $\tau m_t^2 + \sigma_t^2 = \tau$. The rest of the proof is identical to the Lemma C.3 in Bortoli (2022).

So using $m_t \leq 1$ we have

$$\begin{aligned}
\left|\partial_t f_t^k(x) - \partial_t f_t^j(x)\right| &\leq 2\beta_t m_t^2 R^2/\sigma_t^4 + \beta_t m_t \left(1 + m_t^2\right) R\|x\|/\sigma_t^4 \\
&\leq 2\left(\beta_t/\sigma_t^4\right) R(R + \|x\|)
\end{aligned}$$

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

$$\nabla \partial_t f_t^k(x) = 2\beta_t m_t^2/\sigma_t^4 x + \left(\beta_t m_t/\sigma_t^4\right)\left(1 + m_t^2\right) X^k.$$

So we can bound the norm of it by

$$\|\partial_t \nabla f_t^k(x)\| \leq 2\left(\beta_t/\sigma_t^4\right)(R + \|x\|).$$

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

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

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 following lemma, similar to Chen et al. (2023d), we obtain a better control on the time discretization error instead of controlling $\|\partial_t \nabla \log q_t(x_t)\|$ for $\forall x_t \in \mathbb{R}^d$.

Lemma 17. *Assume* **Assumption 2** *and X_t satisfies the forward process Eq. (1). Define* $L = \max_{t \in [0, T-\delta]} \left\|\nabla^2 \log q_{T-t}(\mathbf{Y}_t)\right\| \leq \left(1 + R^2\right)/\sigma_\delta^4$, *then we have that*

$$\begin{aligned}
&\mathbb{E}_{Q_{t_K}^{q_T^\tau}}\left[\left\|\nabla \ln \frac{q_{T-t_k}}{q_{T-t}}(\mathbf{Y}_{t_k})\right\|^2\right] \\
&\lesssim \tau L^2 d\bar{\gamma}_K + \tau L^2 \bar{\gamma}_K^2 (d\tau + R^2) + \tau L^3 \bar{\gamma}_K^2 + \tau L^4 \bar{\gamma}_K^2 (\beta_T d\bar{\gamma}_K + R^2 \bar{\gamma}_K^2).
\end{aligned}$$

Proof. Due to the property of the forward process, we know that if $S : \mathbb{R}^d \to \mathbb{R}^d$ is the mapping $S(x) := \exp(-(t - t_k))x$, then $q_{T-t_k} = S_\# q_{T-t} * \text{normal}\left(0, \tau\left(1 - \exp(-2\int_{t_k}^{t_k+1} \beta_s/\tau \mathrm{d}s)\right)\right)$ Similar to Chen et al. (2023d), we define $\alpha = \exp\left[\int_{t_k}^{t_{k+1}} \frac{\beta_s}{\tau}\mathrm{d}s\right] = 1 + O(\bar{\gamma}_K)$ and $\sigma^2 = \tau\left(1 - \exp(-2\int_{t_k}^{t_k+1} \beta_s/\tau \mathrm{d}s)\right) = O(\tau\bar{\gamma}_K)$. Then we can use Lemma C.12 of Lee et al. (2022) to obtain

$$\begin{aligned}
&\mathbb{E}_{Q_{t_K}^{q_T^\tau}}\left[\left\|\nabla \ln \frac{q_{T-t_k}}{q_{T-t}}(\mathbf{Y}_{t_k})\right\|^2\right] \\
&\lesssim \tau L^2 d\bar{\gamma}_K + \tau L^2 \bar{\gamma}_K^2 \left\|\mathbf{Y}_{t_k}\right\|^2 + \tau L^2 \bar{\gamma}_K^2 \left\|\nabla \ln q_{T-t}(\mathbf{Y}_{t_k})\right\|^2 \\
&\lesssim \tau L^2 d\bar{\gamma}_K + \tau L^2 \bar{\gamma}_K^2 (d\tau + R^2) + \tau L^3 \bar{\gamma}_K^2 + \tau L^4 \bar{\gamma}_K^2 (\beta_T d\bar{\gamma}_K + R^2 \bar{\gamma}_K^2).
\end{aligned}$$

The last inequality follows Lemma 21 and the fact that

$$\begin{aligned}\|\nabla \ln q_{T-t}(\mathbf{Y}_{t_k})\|^2 &\lesssim \|\nabla \ln q_{T-t}(\mathbf{Y}_t)\|^2 + \|\nabla \ln q_{T-t}(\mathbf{Y}_{t_k}) - \nabla \ln q_{T-t}(\mathbf{Y}_t)\|^2 \\ &\lesssim \|\nabla \ln q_{T-t}(\mathbf{Y}_{t_k})\|^2 + L^2(\beta_T d\bar{\gamma}_K + R^2\bar{\gamma}_K^2) \\ &\lesssim L + L^2(\beta_T d\bar{\gamma}_K + R^2\bar{\gamma}_K^2)\,.\end{aligned}$$

■

E AUXILIARY LEMMAS

Lemma 18. *For any $s, t \in [0, T]$ we have*

$$\int_s^t \beta_{T-u}/\sigma_{T-u}^2 \mathrm{d}u = \left[-\frac{1}{2}\log\left(\exp\left[2\int_0^{T-u}\frac{\beta_v}{\tau}\mathrm{d}v\right]-1\right)\right]_s^t,$$

$$\int_s^t \beta_{T-u}m_{T-u}^2/\sigma_{T-u}^4 \mathrm{d}u = \left[(1/2\tau)\,/\left(1-\exp\left[-2\int_0^{T-u}\frac{\beta_v}{\tau}\mathrm{d}v\right]\right)\right]_s^t.$$

Proof. We directly compute

$$\begin{aligned}\int_s^t \beta_{T-u}/\sigma_{T-u}^2 \mathrm{d}u &= \frac{1}{\tau}\int_s^t \beta_{T-u}/\left(1-\exp\left[-2\int_0^{T-u}\frac{\beta_v}{\tau}\mathrm{d}v\right]\right)\mathrm{d}u \\ &= \frac{1}{\tau}\int_s^t \beta_{T-u}\exp\left[2\int_0^{T-u}\frac{\beta_v}{\tau}\mathrm{d}v\right]/\left(\exp\left[2\int_0^{T_u}\frac{\beta_v}{\tau}\mathrm{d}v\right]-1\right)\mathrm{d}u \\ &= -\frac{1}{2}\int_s^t \partial_u \log\left(\exp\left[2\int_0^{T-u}\frac{\beta_v}{\tau}\mathrm{d}v\right]-1\right)\mathrm{d}u\,.\end{aligned}$$

Similarly

$$\begin{aligned}&\int_s^t \beta_{T-u}m_{T-u}^2/\sigma_{T-u}^4 \\ &= \frac{1}{\tau^2}\int_s^t \beta_{T-u}\exp\left[-2\int_0^{T-u}\frac{\beta_v}{\tau}\mathrm{d}v\right]/\left(1-\exp\left[-2\int_0^{T-u}\frac{\beta_v}{\tau}\mathrm{d}v\right]\right)^2\mathrm{d}u \\ &= (1/2\tau)\int_t^s \partial_u\left(1-\exp\left[-2\int_0^{T-u}\frac{\beta_v}{\tau}\mathrm{d}v\right]\right)^{-1}\mathrm{d}u.\end{aligned}$$

■

Lemma 19. *Assume* **Assumption 1**. *For $i \in \{1, 2\}$, we have $\sigma_{T-t_K}^2(i) \le 2\delta$ and $\sigma_u^{-2}(i) \le \sigma_{T-t_K}^{-2}(i) \le \frac{1}{\tau} + \frac{\bar{\beta}}{\delta^{4-i}}, \forall u \in [T - t_K, T]$.*

Proof.

$$\begin{aligned}\sigma_{T-t_K}^2(i) &= \tau\left(1-\exp\left[-2\int_0^{T-t_K}\frac{\beta_s}{\tau}\,\mathrm{d}s\right]\right) \\ &\le 2\int_0^{T-t_K}\beta_s\,\mathrm{d}s \le 2\delta\,,\end{aligned}$$

where the first inequality follows from for any $a \ge 0, \exp[-a] \ge 1 - a$; the second inequlity follows from **Assumption 1** and $\delta \le 1$.

$$\sigma_{T-t_K}^{-2}(i) = \frac{1}{\tau}\left(1-\exp\left[-2\int_0^{T-t_K}\frac{\beta_s}{\tau}\,\mathrm{d}s\right]\right)^{-1} \leq \frac{1}{\tau}\left(1+\left(2\int_0^{T-t_K}\frac{\beta_s}{\tau}\,\mathrm{d}s\right)^{-1}\right)$$
$$\leq \frac{1}{\tau}+\frac{\bar{\beta}}{\delta^{4-i}}\,,$$

where the first inequality follows from for any $a \geq 0, 1/(1+\exp[-a]) \leq 1+1/a$, the second inequality follows from **Assumption 1**. It is easy to check that $\sigma_u^{-2}(i) \leq \sigma_{T-t_K}^{-2}(i), \forall u \in [T-t_K, T]$.

■

Using the bound on $\sigma_{T-t_K}^{-2}(i)$ immediately yields the following control of $\beta_u/\sigma_u^2(i)$.

Lemma 20. *Assume* **Assumption 1**. *Then, we have for any $u \in [T-t_K, T]$: (1) if $i=1$, then*

$$\frac{\beta_u}{\sigma_u^2(i=1)} \leq \kappa_1(\tau) = \max\{\bar{\beta}, T^2\}\left(\frac{1}{\tau}+\frac{\bar{\beta}}{\delta^3}\right)$$

(2) if $i=2$, then

$$\frac{\beta_u}{\sigma_u^2(i=2)} \leq \kappa_2(\tau) = \max\{\bar{\beta}, T\}\left(\frac{1}{\tau}+\frac{\bar{\beta}}{\delta^2}\right)$$

In the rest of this section, we provide the useful lemma to achieve polynomial sample complexity for VE-based models with reverse SDE. As shown in Lemma 13, we also need to control $\mathbb{E}[\|\mathbf{X}_t\|^2]$ in the forward process. The following lemmas shows that this term is bounded by the R^2 and exploding variance.

Lemma 21. *Suppose that* **Assumption 2***hold. Let $(\mathbf{X}_t)_{t\in[0,T]}$ denote the forward process Eq.* (1). *Then, for all $t \geq 0$,*

$$\mathbb{E}\left[\|\mathbf{X}_t\|^2\right] \leq d\sigma_t^2 \vee R^2\,.$$

Proof. As shown in Eq. (2),

$$\mathbb{E}\left[\|\mathbf{X}_t\|^2\right] \leq \mathbb{E}\left[\|\mathbf{X}_0\|^2\right] + \sigma_t^2 d \leq d\sigma_t^2 \vee R^2\,.$$

■

Lemma 22 (movement bound for VESDE). *Let $(\mathbf{X}_t)_{t\in[0,T]}$ denote the forward process Eq.* (1). *For $0 \leq s < t$ with $\delta := t-s$, if $\delta \leq 1$, then*

$$\mathbb{E}\left[\|\mathbf{X}_t-\mathbf{X}_s\|^2\right] \lesssim 2\beta_t\delta d + \delta^2 R^2\,.$$

Proof.

$$\mathbb{E}\left[\|\mathbf{X}_t-\mathbf{X}_s\|^2\right] \lesssim \mathbb{E}\left[\left\|\sqrt{2\beta_t}\,(B_t-B_s)\right\|^2\right] + \delta\int_s^t \mathbb{E}\left[\|\mathbf{X}_r\|^2\right]\mathrm{d}r \lesssim 2\beta_t\delta d+\delta^2R^2\,.$$

■

Similar to Chen et al. (2023d), 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.

Lemma 23. *Suppose* **Assumption 2** *holds. Let $\epsilon_{W_2} > 0$. If $\beta_t^2 = t^2$ and $\tau = T^2$, we choose the early stopping parameter $\delta \leq \frac{\epsilon_{W_2}^{2/3}}{(d+R\sqrt{d})^{1/3}}$. If $\beta_t = t$ and $\tau = T$, we choose $\delta \leq \frac{\epsilon_{W_2}}{(d+R\sqrt{d})^{1/2}}$. If consider pure VESDE (SMLD) (Eq. (3)) with $\sigma_t^2 = t$, we choose $\delta \leq \frac{\epsilon_{W_2}^2}{d}$. Then we have $W_2(q_\delta, q_0) \leq \epsilon_{W_2}$.*

Proof. For the forward process Eq. (1), we know that $\mathbf{X}_t := m_t\mathbf{X}_0+\sigma_t Z$, where $Z \sim \text{normal}(0, I_d)$ is independent of X_0 and $m_t \leq 1$. Hence, for $\delta \lesssim 1$,

$$W_2^2(q_0, q_\delta) \leq (1-m_t)^2 \mathbb{E}\left[\|\mathbf{X}_0\|^2\right] + \mathbb{E}\left[\|\sigma_\delta Z\|^2\right].$$

For $\beta_t = t^2$ and $\tau = T^2$, we have that

$$W_2^2(q_0, q_\delta) \leq \delta^3 d + \frac{R^2\delta^6}{T^2}$$

Hence, we can take $\delta \leq \frac{\epsilon_{W_2}^{2/3}}{(d+R\sqrt{d})^{1/3}}$. For $\beta_t = t$ and $\tau = T$, we have that

$$W_2^2(q_0, q_\delta) \leq \delta^2 d + \frac{R^2\delta^4}{T^2}$$

Hence, we can take $\delta \leq \frac{\epsilon_{W_2}}{(d+R\sqrt{d})^{1/2}}$. For pure VESDE (Eq. (3)) with $\sigma_t = t$, we have

$$W_2^2(q_0, q_\delta) \leq \delta d.$$

■

F Additional Synthetic experiments

In this section, we do synthetic experiments to show the power of our new forward process with small drift term in different setting.

F.1 The Synthetic experiments with accurate score function

In this section, we do numerical experiments on 2-dimension Gaussian distribution to show the power of our new VESDE forward process in balancing different error sources.

Experiment Setting. We set the mean of target distribution $\mathbb{E}[q_0] = [6, 8]$, the covariance matrix $\text{Cov}[q_0] = \begin{bmatrix} 25 & 5 \\ 5 & 4 \end{bmatrix}$, the diffusion time $T = 2$, $\tau = T^2$ and the reverse beginning distribution is $\mathcal{N}(0, T^2\mathbf{I})$. We choose uniform stepsize $\gamma_k = h, \forall k \in [K]$ where $h \in \{0.005, 0.01, 0.02, 0.04\}$. For score functions, we directly calculate the ground truth score function instead of learning it by the score matching objective. We calculate the KL divergence between the generation distribution and target distribution q_0 as the experiments.

The implementable algorithm. We choose three different VESDE forward processes in the experiments: (1) aggressive $\beta_t = t^2$ with $\tau = T^2$; (2) conservative $\beta_t = t$ with $\tau = T^2$ and (3) VESDE without drift term Eq. (3) with $\sigma_t^2 = t^2$. After determining the forward process, we run the reverse SDE with the above $\gamma_k, k \in [K]$. For the discretization scheme, we choose two common method: exponential integrator (EI) (Zhang and Chen, 2022) and Euler-Maruyama (EM) discretization (Ho et al., 2020).

Observations. The experimental results are shown in Fig. 1. We note that the red line (EI, VESDE without drift, $\sigma_t^2 = t^2$) and orange line (conservative drift VESDE, $\beta_t = t$ and $\tau = T^2$) has a similar trend. Furthermore, the conservative drift VESDE has better performance compared to pure VESDE without drift term. Hence, our new forward process is representative enough to represent current VESDE, as discussed in Section 3.1.

The experimental results also support our theoretical results and show the power of the new forward process in balancing different error terms. As shown in Fig. 1, the process with aggressive $\beta_t = t^2$ with small drift term achieves the best and second performance in EI and EM discretization since it can balance the reverse beginning and discretization. The third best process is conservative $\beta_t = t$ with the small drift term. The reason is that though it can not achieve a $\exp(-T)$ forward process guarantee, it also has a constant decay on prior information, as shown in Section 3.1. This decay slightly reduces the effect of the reverse beginning error. The worse process is VESDE without drift term since it is hard to balance different error sources. Our experimental results also show that EI discretization is better than EM discretization.

F.2 The Synthetic experiments with approximated score function

In this section, instead of using an accurate score function, we train an approximated score function on the pure VESDE (Eq. (3)) without drift term on two synthetic datasets: multiple Swiss rolls and 1-D GMM. Then, for the drift VESDE, we do not train the approximated score corresponding to Eq. (1); we directly use the approximated score learned by pure VESDE and show that the drift VESDE can improve the generated distribution without the training process.

Datasets. The 1-D GMM distribution contains three modes:

$$\frac{3}{10}\mathcal{N}\left(-8, 0.01\right) + \frac{3}{10}\mathcal{N}\left(-4, 0.01\right) + \frac{4}{10}\mathcal{N}\left(3, 1\right) .$$

For multiple Swiss rolls, we use a similar code compared to Listing 2 of Lai et al. (2023), except Line 6. We change Line 6. to data /=10. to obtain a larger variance dataset. Each dataset contains 50000 datapoints.

The implementable algorithm. In this subsection, we adapt the code of a particular repository[1], which corresponds to Eq. (3) with $\sigma_t^2 = t$, as mentioned in the Appendix C.2 of Karras et al. (2022). Hence, we choose two forward processes: (1) conservative $\beta_t = 1$ with $\tau = T$; (2) pure VESDE without drift term (Eq. (3)) with $\sigma_t^2 = t$. To match our analysis, we choose two sampling methods for the reverse process: Euler-Maruyama method for reverse SDE and RK45 ODE solver for the reverse PFODE method.

We note that although aggressive setting $\beta_t = t$ and $\tau = T$ has shown its power in theory (Lemma 5) and the experiments with accurate score (Fig. 1), other sampling issues may arise in practice. We leave the experimental exploration for drift VESDE with aggressive β_t as a future work.

The training detail. For each dataset, we train a score function with pure VESDE (Eq. (3), $\sigma_t^2 = t$) by using exactly the same network compared to the above repository. We train for 200 epochs with batch size 200 and learning rate 10^{-4}. For both training and inference, the start time is $\delta = 10^{-5}$. For the conservative VESDE, we directly adapt the checkpoint learned by the pure VESDE since the conservative drift VESDE has a similar trend compared to pure VESDE, as shown in Fig. 1.

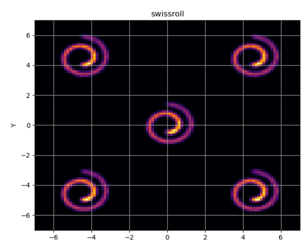
Observation. We do experiments with $T = 100$ and lager $T = 625$ and these two choice show similar phenomenon. In this paragraph, we first use $T = 100$ as an example to discuss the results. As shown in Table 1, the conservative drift VESDE has smaller KL divergence compared to pure VESDE under all sampling methods and datasets. From Fig. 2 and Fig. 3, it is clear that pure VESDE has low density on the Swiss roll except the center one, which means that though pure VESDE can deal with small $\mathbb{E}[q_0]$, it is hard to deal with large dataset variance $\mathrm{Cov}[q_0]$, as we discuss in Section 4. For conservative drift VESDE ($\beta_t = 1$ and $\tau = T$), as we discuss in Section 3.1, there is a constant decay on the prior information $\mathbb{E}[q_0]$ and $\mathrm{Cov}[q_0]$, which is helpful in deal with large dataset mean and variance. The experimental results support our augmentation. Fig. 2c, Fig. 3c and Fig. 4c show that the density of the generated distribution is more uniform compared to pure VESDE, which means that the drift VESDE can deal with large dataset mean and variance.

We also do experiments with larger $T = 625$. As we discuss in Section 4, larger T will reduce the influence of the prior data information and have greater generated distribution, as shown in Fig. 4b and Fig. 3c. The experiments of 1D-GMM (Fig. 4) show a similar phenomenon compared to the multi Swiss rolls.

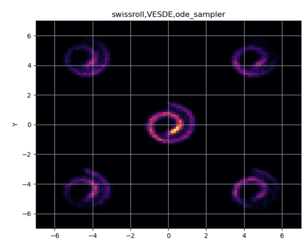
[1]https://colab.research.google.com/drive/120kYYBOVa1i0TD85RjlEkFjaWDxSFUx3?usp=sharing

Table 1: The KL divergence for pure VESDE (Eq. (3)) and conservative drift VESDE with different sampling method.

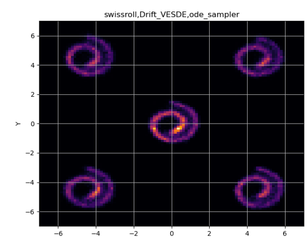
Forward Process	1-D GMM		Swiss roll	
	Reverse SDE	PFODE	Reverse SDE	PFODE
Pure VESDE ($T = 100$)	0.082	0.434	9.58	21.05
Drift VESDE ($T = 100$)	0.043	0.249	8.71	7.77
Pure VESDE ($T = 625$)	0.027	0.057	8.00	8.20
Drift VESDE ($T = 625$)	**0.025**	0.031	7.95	**7.21**



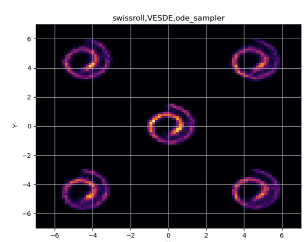
(a) Original Figure



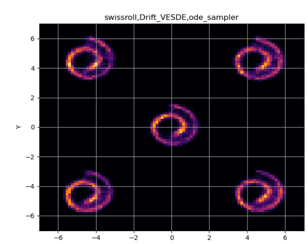
(b) Pure VESDE ($T = 100$)



(c) Drift VESDE ($T = 100$)

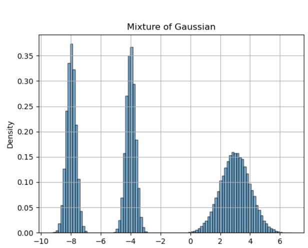


(d) Pure VESDE ($T = 625$)

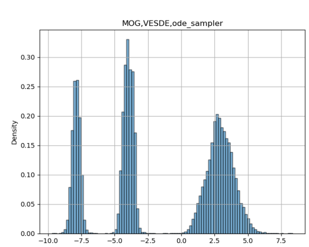


(e) Drift VESDE ($T = 625$)

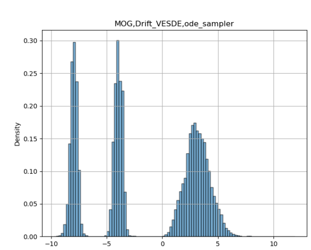
Figure 3: Experiment results of Swiss roll with reverse PFODE



(a) Original Figure



(b) Pure VESDE ($T = 625$)



(c) Drift VESDE ($T = 625$)

Figure 4: Experiment results of 1D-GMM with reverse PFODE