Localized Diffusion Models for High Dimensional Distributions Generation

Georg A. Gottwald* Shuigen Liu† Youssef Marzouk‡ Sebastian Reich§
Xin T. Tong†

May 8, 2025

Abstract

Diffusion models are the state-of-the-art tools for various generative tasks. However, estimating high-dimensional score functions makes them potentially suffer from the curse of dimensionality (CoD). This underscores the importance of better understanding and exploiting low-dimensional structure in the target distribution. In this work, we consider *locality structure*, which describes sparse dependencies between model components. Under locality structure, the score function is effectively low-dimensional, so that it can be estimated by a localized neural network with significantly reduced sample complexity. This motivates the *localized diffusion model*, where a localized score matching loss is used to train the score function within a localized hypothesis space. We prove that such localization enables diffusion models to circumvent CoD, at the price of additional localization error. Under realistic sample size scaling, we show both theoretically and numerically that a moderate localization radius can balance the statistical and localization error, leading to a better overall performance. The localized structure also facilitates parallel training of diffusion models, making it potentially more efficient for large-scale applications.

1 Introduction

Over the past decade, numerous neural network (NN)-based sampling algorithms have emerged in the machine learning literature, demonstrating remarkable performance across various tasks. These methods, often referred to as generative models, include approaches such as normalizing flows [35], variational encoders [24], generative adversarial network [18], and diffusion models [37, 21, 38]. Among all generative models, diffusion model (DM), usually referring to the denoising diffusion probabilistic model (DDPM) [21], is a state-of-the-art and widely used approach. It has gained great popularity due to its capability in generating high quality samples, particularly in tasks such as image synthesis [21, 14]. A series of recent studies [28, 9, 4, 32, 7, 43] theoretically justify the approximation and generalization capabilities of DMs over a broad class of target distributions. However, there remains limited understanding of its effectiveness in handling high-dimensional distributions.

DMs are known to be expensive when it comes to training with high-dimensional data. The training sample size needs to grow exponentially with the problem dimension [39, 32], known as the *curse of dimensionality* (CoD) in the literature. Various attempts have been made to avoid it by leveraging low-dimensional structures within the target distribution. The *manifold hypothesis* [15] which assumes that the data lies on a low-dimensional manifold, is often envoked

*The University of Sydney (georg.gottwald@sydney.edu.au)
†National University of Singapore (shuigen@u.nus.edu, mattxin@nus.edu.sg).
‡Massachusetts Institute of Technology (ymarz@mit.edu)
§University of Potsdam (sereich@uni-potsdam.de)

to postulate such structures. For such data, [32, 7, 40, 1] show that the sample complexity of DMs depends on the dimension of the manifold rather than the ambient dimension, and DMs can avoid the CoD with appropriate NN structure. There are also studies considering Gaussian mixtures [36, 45, 17] to avoid the CoD. In both manifold hypothesis and the Gaussian mixture models, although the ambient space is high dimension, there is a low-dimensional latent structure that effectively characterizes the target distribution.

While these concepts of low effective dimension can cover many applications, there are still important cases left open. One large class of high-dimensional distributions are those with *locality structure* [5, 41, 16, 13]. We say a distribution has locality structure if each model component only has strong conditional dependence on a sparse selection of the other model components. An illustrative example is the Ginzburg-Landau model from statistical physics [27], where d particles in one-dimensional configurations follow the distribution

$$p(x_1, \ldots, x_d) = \frac{1}{Z} \exp\left(\sum_{j=1}^{d} V(x_j) + \sum_{j=1}^{d-1} W(x_j, x_{j+1}) \right).$$

In this model, each particle, denoted by x_j, interacts directly only with its neighbors $x_{j\pm1}$. Such sparse dependence structure arises naturally in spatial models, and has been successfully applied in various fields such as spatial statistics [5], data assimilation [34], quantum mechanics [25] and sampling [41]. We refer to [13] for a detailed review on the locality structure.

An important property of distributions with locality structure, or *localized distributions*, is that their score functions are effectively low-dimensional [41, 13]. Due to the conditional independence, the score component $s_j(x) = \nabla_j \log p(x)$ depends only on $x_{\mathcal{N}_j}$, where $x_{\mathcal{N}_j}$ is the component conditionally dependent on x_j. If the conditional dependencies are sparse, the dimension of $x_{\mathcal{N}_j}$ is much smaller than the ambient dimension d, so that the score function can be regarded as a collection of low-dimensional functions $\{s_j\}_{j\in[d]}$. This suggests that learning the score functions of localized distributions does not suffer from the CoD.

Motivated by this, we propose the *localized diffusion model* (LDM), which embeds the locality structure into the hypothesis space of the score function, reducing a high dimensional score matching problem to a low dimensional one. With small effective dimension, the statistical error of score estimation is significantly reduced. On the other hand, localizing the hypothesis space introduces additional localization error. By a complete approximation and generalization analysis, we show that by adjusting the localization radius, one can balance the tradeoff between the statistical error and the localization error to achieve smaller overall error. This can be interpreted as a tradeoff between variance and bias. Such tradeoff is validated by numerical experiments on high-dimensional time series data. Finally, we find that LDM can be interpreted as a collection of diffusion models on low-dimensional marginals. That is, we construct the samplers by combining local samplers for the marginals of the localized distributions. This allows LDM to be trained parallelly, which is practically important for large-scale applications.

The paper is organized as follows. In Section 2, we review diffusion models and the locality structure, and show that the locality structure is approximately preserved in the forward diffusion process. In Section 3, we introduce the localized diffusion model and analyze its approximation and statistical error. In Section 4, we present numerical experiments to validate our theoretical results.

1.1 Related Work

Analysis of Diffusion Models Since the introduction of DMs [37, 21, 38], there has been a surge of interest in understanding their theoretical properties. Our work is built on two main lines of research: the convergence of DMs and the statistical analysis of DMs. A comprehensive review of all related work is beyond the scope of this paper; we refer to [8, 17] for an in-depth overview.

The convergence of DMs considers error bounds of the sampled distribution given the learned score function. Early work [28] provides a TV guarantee by assuming a log-Sobolev inequality. Later, by using Girsanov theorem, this condition is relaxed to bounded moment conditions [9, 6]. A growing body of work is trying to further relax assumptions and improve error bounds. For instance, [4] proves a linear-in-dimension bound under the KL divergence, [10] uses a relative score approach and derives bounds without early stopping. [33] considers the manifold data, and improves the bound of the discretization error to scale linearly with the manifold dimension.

The statistical analysis of DMs essentially studies the sample complexity of estimating the score function. [32, 43] prove that the diffusion model reaches the minimax rate for distribution estimation. To avoid the CoD, [32, 7] considers linear subspace data, and later [40, 1] extends it to general manifold data. Recently, [44] relaxes the manifold assumption, and improves the ambient dimension dependence in the generalization bound. Other types of low-dimensional structures are also considered. [36] considers certain Gaussian mixtures, and shows that the sample complexity does not depend exponentially on the dimension. [17] further extends it to general Gaussian mixtures with edited diffusion models.

We mention that a recent work [30] considers similar settings as ours. They apply the diffusion models for high-dimensional graphical models. Inspired by variational inference denoising algorithms, they design a residual network to efficiently approximate the score function, and prove that its sample complexity does not suffer from CoD. However, their result depends on an explicit solution of the denoising algorithms, and only applies to Ising model-type distributions. The method we propose in this paper applies to general high-dimensional graphical models.

Localized Sampler In recent years, there has been a fast growing interest in sampling methods that leverage locality structures [46, 31, 41, 20]. These localized samplers follow the general strategy to build samplers by combining local samplers for the marginals. [31] propose to apply the localization technique in Markov chain Monte Carlo (MCMC) and introduces a localized Metropolis-within-Gibbs sampler. [41] extends this idea and develops the MALA-within-Gibbs sampler, which is proven to admit a dimension independent convergence rate. Beyond MCMC, [46] proposes Message Passing Stein Variational Gradient Descent. It finds the descent direction coordinate-wisely, and reduces the degeneracy issue of kernel methods in high dimensions. [20] proposes a localized version of the Schrödinger Bridge (SB) sampler [19], which replaces a single high-dimensional SB problem by d low-dimensional SB problems, avoiding the exponential dependence of the sample complexity on the dimension.

1.2 Notations

- Sets. Denote $[n] = \{1, 2, \dots, n\}$, and the cardinality of a set A as $|A|$. Given $x \in \mathbb{R}^n$ and $A \subset [n]$, denote x_A as the subvector of x with components' indices from A.
- Norms. For a vector $x \in \mathbb{R}^d$, denote $\|x\|$ as its ℓ_2-norm. For a matrix $A \in \mathbb{R}^{m\times n}$, denote $\|A\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|}$ as the 2-matrix norm. For a probability distribution p and a function f, denote $\|f\|_{L^2(p)} = \left(\int f^2(x)p(x)\mathrm{d}x\right)^{1/2}$ as the weighted L^2-norm.
- Probability. Denote $\mathsf{Law}(X)$ as the distribution of a random variable X. Denote the covaraince matrix of X, Y as $\mathsf{Cov}_p(X, Y) := \mathbb{E}_p[(X - \mathbb{E}_p[X])\,(Y - \mathbb{E}_p[Y])^{\mathrm{T}}]$. Denote $\mathsf{N}(\mu, \Sigma)$ as the Gaussian distribution with mean μ and covariance Σ. Denote $X \perp\!\!\!\perp Y \mid Z$ if X is independent of Y given Z; i.e. $\mathbb{P}(X, Y|Z) = \mathbb{P}(X|Z)\mathbb{P}(Y|Z)$.

2 Diffusion Models and Localized Distributions

2.1 Diffusion Models

Diffusion models operate by simulating a process that gradually transforms a simple initial distribution, often Gaussian noise, into a complex target distribution, which represents the data of interest. The core formulation involves two processes: a forward Ornstein–Uhlenbeck (OU) diffusion process which evolves data samples from the data distribution p_0 to noisy samples drawn from a Gaussian distribution, and a reverse diffusion process that learns to progressively denoise the samples and effectively reconstruct the original data distribution.

Consider a forward OU process $(X_t)_{t\in[0,T]}$ that is intialized with the target distribution p_0 and follows the process, i.e.,

$$\mathrm{d}X_t = -X_t\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \quad X_0 \sim p_0. \tag{2.1}$$

Denote its reverse process as $(Y_t)_{t\in[0,T]}$ s.t. $Y_t = X_{T-t}$. Under mild conditions, Y_t follows the reverse SDE [38]

$$\mathrm{d}Y_t = \left(Y_t + 2\nabla \log p_{T-t}(Y_t)\right)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \quad Y_0 \sim p_T, \tag{2.2}$$

where we denote $p_t = \mathsf{Law}(X_t)$. The target distribution p_0 can then be sampled by first sampling $Y_0 \sim p_T$ and then evolving Y_t according to (2.2) to obtain a sample $Y_T \sim p_0$.

To implement the above scheme, several approximations are needed:

1. Score estimation. The score function $s(x,t) := \nabla \log p_t(x)$ is not accessible, and needs to be estimated from the data via the denoising score matching scheme [42, 37, 21]

$$\widehat{s} = \arg\min_{s_\theta} \mathcal{L}(s_\theta),$$

$$\mathcal{L}(s_\theta) := \int_0^T \mathbb{E}_{x_0\sim p_0}\left[\mathbb{E}_{x_t\sim p_{t|0}(x_t|x_0)}\left[\left\|s_\theta(x_t,t) - \nabla_{x_t}\log p_{t|0}(x_t|x_0)\right\|^2\right]\right]\mathrm{d}t. \tag{2.3}$$

   In the sampling process, the true score $\nabla \log p_{T-t}(Y_t)$ in (2.2) is approximated by the estimated score $\widehat{s}(Y_t, T-t)$.

2. Approximation of p_T. The initial distribution p_T in the reverse process is intractable. But since the OU process converges exponentially to $p_\infty = \mathsf{N}(0,I)$, we can approximate p_T by $\mathsf{N}(0,I)$ in (2.2), i.e., Y_0 is drawn from $\mathsf{N}(0,I)$.

3. Early stopping. The reverse process is usually stopped at $t = T - \underline{t}$ for some small $\underline{t} > 0$ to avoid potential blow up of the score function s_t as $t \to 0$. The early stopped samples satisfy $Y_{T-\underline{t}} \sim p_{\underline{t}}$, which should be close to p_0 when $\underline{t}$ is small.

4. Time discretization. The Euler-Maruyama scheme is used to discretize (2.2). Pick time steps $0 = t_0 < t_1 < \cdots < t_N = T - \underline{t}$, and evolve $n = 0, 1, \ldots, N-1$ by

$$Y_{t_{n+1}} = Y_{t_n} + \left(Y_{t_n} + 2\widehat{s}(Y_{t_n}, T - t_n)\right)\Delta t_n + \sqrt{2\Delta t_n}\xi_n, \tag{2.4}$$

   where $\Delta t_n = t_{n+1} - t_n$ and $\xi_n \sim \mathsf{N}(0,I)$. Design of the time steps (the schedule) is crucial for the empirical performance of the sampling process.

Note the OU process admits an explicit transition kernel

$$p_{t|0}(x_t|x_0) = \mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I), \quad \alpha_t := \mathrm{e}^{-t}, \quad \sigma_t := \sqrt{1 - \mathrm{e}^{-2t}}. \tag{2.5}$$

So that $\nabla_{x_t}\log p_{t|0}(x_t|x_0) = -\sigma_t^{-2}(x_t - \alpha_t x_0)$, and $p_{t|0}(x_t|x_0)$ can be realized as

$$x_t = \alpha_t x_0 + \sigma_t \epsilon_t, \quad \epsilon_t \sim \mathsf{N}(0,I).$$

Therefore, the denoising score matching loss in (2.3) can be written as

$$\mathcal{L}(s_\theta) = \int_{\underline{t}}^{T} \mathbb{E}_{x_0\sim p_0}\mathbb{E}_{\epsilon_t\sim \mathsf{N}(0,I)} \left[\left\| s_\theta(\alpha_t x_0 + \sigma_t \epsilon_t, t) + \sigma_t^{-1}\epsilon_t \right\|^2\right] \mathrm{d}t, \tag{2.6}$$

where we involved the early stopping truncation. The above loss provides a convenient form for implementation [22].

2.2 Locality Structure

We will use the undirected graphical model [29, 26], also known as Markov random field, to describe the locality structure. In this model, the conditional dependencies of a collection of random variables are encoded in the underlying dependency graph. So that the sparsity of the graph can be used to characterize the locality structure in the joint distribution of these random variables. We will define the localized and approximately localized distributions based on the dependency graph.

2.2.1 Sparse Graphical Models

Following [13], consider an undirected graph $G = (V, E)$ and an associated random variable

$$X = (X_i)_{i\in V} \in \mathbb{R}^d, \quad X_i \in \mathbb{R}^{d_i}, \quad d = \sum_{i\in V} d_i. \tag{2.7}$$

Here we assume that the dimension of each component d_i is small, but the total dimension d is large. We say X has *dependency graph* G, if for any nonadjacent vertices $i, j \in V$, X_i, X_j are conditionally independent given the rest of the components $(X_k)_{k\neq i,j}$, i.e.,

$$X_i \perp\!\!\!\perp X_j \mid (X_k)_{k\neq i,j}. \tag{2.8}$$

X is called a sparse graphical model if the dependency graph G is sparse, which essentially encodes the sparse local dependencies in X. The following equivalent characterization [41] of the sparse graphical models will be crucial. Let $p = \mathsf{Law}(X)$. If $p(x)$ is twice differentiable, then (2.8) equivalent to

$$\forall \text{ nonadjacent } i, j \in V \;\Rightarrow\; \nabla^2_{ij} \log p(x) = 0. \tag{2.9}$$

Let $b = |V|$, and attach each vertex in V with a unique index $j \in [b]$. Denote

$$\mathcal{N}_j := \{i \in V : (i, j) \in E\} \tag{2.10}$$

as the neighboring vertices of j. For simplicity, we require that E includes all the self-loops in G; i.e. $j \in \mathcal{N}_j$. We further denote the extended neighborhood of j as

$$\mathcal{N}^r_j = \{i \in V : \mathsf{d}_G(i, j) \le r\}, \tag{2.11}$$

where $\mathsf{d}_G(i, j)$ is the graph path distance between $i, j \in V$, i.e.,

$$\mathsf{d}_G(i, j) = \min\{n \ge 0 : \exists \text{ path of length } n \text{ from } i \text{ to } j\}. \tag{2.12}$$

2.2.2 Localized Distributions

Now we define localized and approximately localized distributions. The former is precisely the sparse graphical models, and the latter is a relaxation based on (2.9), which allows exponentially small long-range dependencies.

Definition 2.1. *A distribution p is called localized w.r.t. an undirected graph G if it satisfies (2.8). A distribution p is called approximately localized w.r.t. G, if there exists dimensional independent constants $c_p, C_p > 0$ such that*

$$\left\|\nabla_{ij}^2 \log p\right\|_\infty \le C_p \exp\left(-c_p \mathsf{d}_G(i,j)\right). \tag{2.13}$$

Here $\|\cdot\|_\infty$ denotes the L^∞-norm, and $\mathsf{d}_G(i,j)$ is the graph distance (2.12).

For localized distributions, consider the j-th component of its score function

$$s_j(x) = \nabla_j \log p(x). \tag{2.14}$$

Note that it is only a function of $x_{\mathcal{N}_j}$, since by (2.9), for any $i \notin \mathcal{N}_j$,

$$\nabla_i s_j(x) = \nabla_{ij}^2 \log p(x) = 0.$$

For sparse graph G, $|\mathcal{N}_j| \ll |V|$, so that s_j is essentially a low-dimensional function, which implies that estimation of s_j does not suffer from the curse of dimensionality. This motivates us to leverage the locality structure in the hypothesis space of the score function, and to localize the score matching procedure. The detailed methods will be discussed in Section 3.

However, the low-dimensionality in the score functions only holds for localized distributions. For approximately localized distributions, the score functions can only be approximated by low-dimensional functions. To improve the approximation accuracy, we can use the expanded neighborhood (2.11) for the approximate scores, i.e.,

$$s_j(x) \approx \widehat{s}_{\theta,j}(x_{\mathcal{N}_j^r}).$$

Here r is the radius of the neighborhood, and can be tuned to balance the approximation accuracy and the sample complexity. Note by (2.13), the approximation error decays exponentially with the radius r, while the dimension of $\widehat{s}_{\theta,j}$ only grows polynomially with r. Section 3 will provide a detailed analysis of the approximation error and the tradeoff in the choices of r.

2.3 Locality Structure in Diffusion Models

We show in this section that the locality structure is preserved in the forward OU process, which lays the foundation for the localized score matching in diffusion models.

The explicit transition kernel (2.5) of the OU process implies that p_t has an explicit density

$$p_t(x_t) = \int \mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I) p_0(x_0) \mathrm{d}x_0.$$

p_t can be viewed as an interpolation between p_0 and $p_\infty = \mathsf{N}(0, I)$. Suppose p_0 is a localized distribution w.r.t. an undirected graph G. It is obvious that p_∞ is localized, but their interpolation p_t may not remain strictly localized. However, p_t is still approximately localized, as proved in the following theorem.

Theorem 2.1. *Suppose p_0 is localized w.r.t. an undirected graph G. Assume additionally that p_0 is log-concave and smooth, i.e., $\exists 0 < m \le M < \infty$ s.t. $mI \preceq -\nabla^2 \log p_0(x) \preceq MI$. Then for any $t \in (0, T]$, p_t is approximately localized w.r.t. G. Specifically,*

$$\left\|\nabla_{ij}^2 \log p_t\right\|_\infty \le \frac{\alpha_t^2}{\sigma_t^2\left(m\sigma_t^2 + \alpha_t^2\right)} \left(1 - \frac{m\sigma_t^2 + \alpha_t^2}{M\sigma_t^2 + \alpha_t^2}\right)^{\mathsf{d}_G(i,j)}. \tag{2.15}$$

Here $\alpha_t = \mathrm{e}^{-t}$ and $\sigma_t = \sqrt{1 - \mathrm{e}^{-2t}}$ (cf. (2.5)), and $\mathsf{d}_G(i,j)$ is the graph distance (2.12).

The proof can be found in Appendix A.1. The first step is to show that

$$\nabla^2_{ij} \log p_t(x_t) = \alpha_t^2 \sigma_t^{-4} \mathsf{Cov}_{p_{0|t}(x_0|x_t)} (x_{0,i}, x_{0,j}) .$$

The bound then directly follows Proposition 2.2 below, which establishes the exponential decay of correlations between x_i, x_j w.r.t. their graph distance $\mathsf{d}_G(i,j)$ for localized distributions. This is a ubiquitous property for distributions with locality structure [25, 34, 13].

Remark 2.1. *(1) While Theorem 2.1 assumes log-concavity to apply Proposition 2.2, the exponential decay of correlations is ubiquitous and does not inherently depend on log-concavity. The assumption is adopted here for simplicity and to derive an explicit quantitative bound.*

(2) It is natural to ask if Theorem 2.1 can be extended to the case where p_0 is only approximately localized. The answer depends on the sparsity of the graph G and the decay rate of $\nabla^2_{ij} \log p_0$. The resulting bound will be complicated, and we do not pursue it here.

We now state the key proposition:

Proposition 2.2. *Suppose p is localized w.r.t. an undirected graph G and is log-concave and smooth, i.e., $\exists 0 < m \le M < \infty$ s.t. $mI \preceq -\nabla^2 \log p(x) \preceq MI$. Then for any i, j and Lipschitz functions $f : \mathbb{R}^{d_i} \to \mathbb{R}$ and $g : \mathbb{R}^{d_j} \to \mathbb{R}$, it holds*

$$\left| \mathsf{Cov}_{p(x)} (f(x_i), g(x_j)) \right| \le \frac{1}{m} \left(1 - \frac{m}{M}\right)^{\mathsf{d}_G(i,j)} |f|_{\mathrm{Lip}} \, |g|_{\mathrm{Lip}} . \tag{2.16}$$

The proof can be found in Appendix A.2.

Remark 2.2. *We note that the condition number $\kappa := \frac{M}{m}$ of typical localized distributions is independent of the dimension d. This is in contrast to the distributions for fixed-domain models with finer resolution. The key difference is the different nature of the high-dimensionality. An illustrative example is the 1d lattice model:*

$$p(x) \propto \exp\left(\frac{1}{2} x^{\mathrm{T}} A x - \frac{\gamma}{2} \|x\|^2\right),$$

where $x \in \mathbb{R}^d$, and $x^{\mathrm{T}} A x$ comes from a discretized Laplacian.

1. *Fixed-domain type. Fix the domain $[0,1]$ and take $x_k = kh$ and $h = (d+1)^{-1}$. Then*

$$-\nabla^2 \log p(x) = -A + \gamma I = \frac{1}{h^2} \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ 0 & -1 & 2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 2 \end{bmatrix} + \gamma I.$$

   The condition number is thus

$$\kappa = \frac{\gamma + 4h^{-2} \sin^2 \frac{d\pi}{2(d+1)}}{\gamma + 4h^{-2} \sin^2 \frac{\pi}{2(d+1)}} \approx \frac{\sin^2 \frac{d\pi}{2(d+1)}}{\sin^2 \frac{\pi}{2(d+1)}} \asymp d^2.$$

2. *Extended-domain (locality) type. Fix the mesh size $h = h_0$, and consider an extended domain $[0, (d+1)/h_0]$. Take $x_k = kh_0$, then $-\nabla^2 \log p(x)$ has the same form as above with $h = h_0$. Therefore,*

$$\kappa = \frac{\gamma + 4h_0^{-2} \sin^2 \frac{d\pi}{2(d+1)}}{\gamma + 4h_0^{-2} \sin^2 \frac{\pi}{2(d+1)}} \approx \frac{\gamma + 4h_0^{-2}}{\gamma} \asymp 1.$$

   In summary, the high-dimensionality in distributions of fixed-domain type comes from refined discretization; while for locality structure, it comes from an extended domain. Since interaction is still local in the extended system, the condition number should be dimension independent.

3 Localized Diffusion Models

3.1 Localized Denoising Score Matching

3.1.1 Localized Hypothesis Space

To exploit the locality structure in diffusion models, we introduce the localized hypothesis space for the score function,

$$\mathscr{H}_r = \left\{ s_\theta : \mathbb{R}^{d+1} \to \mathbb{R}^d \mid s_{\theta,j}(x,t) = u_{\theta,j}(x_{\mathcal{N}_j^r}, t), u_{\theta,j} \in \mathscr{U}_j, j \in [b] \right\}, \tag{3.1}$$

where r denotes the localization radius, $\mathcal{N}_j^r$ is the extended neighborhood (2.11), and $\mathscr{U}_j$ is certain hypothesis space for the j-th component of the score function to be specified later. Note here we use $s_{\theta,j}(\cdot,t)$ to approximate the score function of p_t in light of Theorem 2.1.

Define the *effective dimension* of s_θ as

$$d_{\text{eff}} := \max_j d_{j,r}, \quad d_{j,r} := \sum_{i \in \mathcal{N}_j^r} d_i. \tag{3.2}$$

Since $s_\theta(\cdot,t)$ can be viewed as a collection of functions $\{u_{\theta,j}(\cdot,t) : \mathbb{R}^{d_{j,r}} \to \mathbb{R}^{d_j}\}_{j\in[b]}$, it is essentially a function of d_{eff} variables. For sparse graph, $d_{\text{eff}} \ll d$, so that intuitively estimating s_θ in $\mathscr{H}_r$ does not suffer from the CoD.

3.1.2 ReLU Neural Network

In practice, $\mathscr{H}_r$ can be realized by a neural network (NN) with locality constraints. Following [32], we introduce the hyperparameters of a sparse NN as follows:

- $\mathsf{L} \in \mathbb{Z}_+$ denotes the depth of the NN.
- $\mathsf{W} = (\mathsf{w}_0, \dots, \mathsf{w}_\mathsf{L}) \in \mathbb{R}^{\mathsf{L}+1}$ denotes the width vector of the NN.
- S, B denote the sparsity and boundedness of the parameters.

Consider the ReLU NN class with hyperparameters $(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B})$:

$$\begin{aligned} \mathsf{NN}(\mathsf{L},\mathsf{W},\mathsf{S},\mathsf{B}) &= \{u_\theta : \mathbb{R}^{\mathsf{w}_0} \to \mathbb{R}^{\mathsf{w}_\mathsf{L}} \mid \theta \in \Theta(\mathsf{L},\mathsf{W},\mathsf{S},\mathsf{B})\}, \\ \Theta(\mathsf{L},\mathsf{W},\mathsf{S},\mathsf{B}) &= \left\{ \theta = \{W_l, b_l\}_{l=1}^{\mathsf{L}} \mid W_l \in \mathbb{R}^{\mathsf{w}_l \times \mathsf{w}_{l-1}}, b_l \in \mathbb{R}^{\mathsf{w}_l}, \|\theta\|_0 \le \mathsf{S}, \|\theta\|_\infty \le \mathsf{B} \right\}, \\ u_\theta(x) &= W_\mathsf{L}\sigma(W_{\mathsf{L}-1}\sigma(\cdots\sigma(W_1 x + b_1)\cdots) + b_{\mathsf{L}-1}) + b_\mathsf{L}, \end{aligned} \tag{3.3}$$

where $\sigma(x) = \max\{0, x\}$ is the ReLU activation function (operated element-wise for a vector) and $\|\theta\|_0, \|\theta\|_\infty$ are the vector ℓ_0 and ℓ_∞ norms of the parameter θ.

One can choose the hypothesis space $\mathscr{U}_j$ as consisting of such ReLU NNs:

$$\mathscr{U}_j = \mathsf{NN}(\mathsf{L}^j, \mathsf{W}^j, \mathsf{S}^j, \mathsf{B}^j), \quad \text{where } \mathsf{w}_0^j = d_{j,r} + 1,\ \mathsf{w}_\mathsf{L}^j = d_j. \tag{3.4}$$

Here the hyperparameters $\mathsf{L}^j, \mathsf{W}^j, \mathsf{S}^j, \mathsf{B}^j$ are to be determined later.

3.1.3 Localized Score Matching

Given the hypothesis space $\mathscr{H}_r$ (3.1) with localized NN score $\mathscr{U}_j$ (3.4), we can learn the localized score function by minimizing the denoising score matching loss (2.6). Given i.i.d. sample $\{X^{(i)}\}_{i=1}^N$ from p_0, the population loss (2.6) is approximated by the empirical loss, i.e.,

$$\widehat{s} = \arg\min_{s_\theta \in \mathscr{H}_r} \widehat{\mathcal{L}}_N(s_\theta), \tag{3.5}$$

with

$$\widehat{\mathcal{L}}_N(s_\theta) = \frac{1}{N}\sum_{i=1}^{N}\int_{\underline{t}}^{T} \mathbb{E}_{\epsilon_t\sim\mathsf{N}(0,I)}\left[\left\|s_\theta(\alpha_t X^{(i)} + \sigma_t\epsilon_t, t) + \sigma_t^{-1}\epsilon_t\right\|^2\right]\mathrm{d}t. \tag{3.6}$$

Notice $\widehat{\mathcal{L}}_N$ is decomposable: $\widehat{\mathcal{L}}_N(s_\theta) = \sum_{j=1}^{b}\widehat{\mathcal{L}}_{j,N}(u_{\theta,j})$, where

$$\widehat{\mathcal{L}}_{j,N}(u_{\theta,j}) = \frac{1}{N}\sum_{i=1}^{N}\int_{\underline{t}}^{T} \mathbb{E}_{\epsilon_t\sim\mathsf{N}(0,I)}\left[\left\|u_{\theta,j}(\alpha_t X^{(i)}_{\mathcal{N}_j^r} + \sigma_t\epsilon_{t,\mathcal{N}_j^r}, t) + \sigma_t^{-1}\epsilon_{t,j}\right\|^2\right]\mathrm{d}t. \tag{3.7}$$

The optimal $\widehat{u}_j$ then solves

$$\widehat{u}_j = \arg\min_{u_{\theta,j}\in\mathscr{U}_j}\widehat{\mathcal{L}}_{j,N}(u_{\theta,j}). \tag{3.8}$$

This allows for *parallel training* of the localized NNs, i.e., the components of the score function can be trained independently. Note the score function need not be a gradient field, which introduces great flexibility in designing hypothesis space.

Remark 3.1. *For general distributions, the components of the score function are correlated, so that $\{s_{\theta,j}(x)\}_{j=1}^b$ should be trained simultaneously. However, for approximately localized distributions, most components of s_θ are almost uncorrelated, which facilitates parallel training.*

3.2 Error Analysis

3.2.1 Error Decomposition

We do not consider time discretization here for simplicity. The sampling process is

$$\mathrm{d}\widehat{Y}_t = \left(\widehat{Y}_t + 2\widehat{s}(\widehat{Y}_t, T-t)\right)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \quad \widehat{Y}_0 \sim \mathsf{N}(0, I). \tag{3.9}$$

And we take the early stopped distribution $\widehat{q}_{T-\underline{t}} = \mathsf{Law}(\widehat{Y}_{T-\underline{t}})$ as the approximation of p_0. It suffices to consider the error between $\widehat{q}_{T-\underline{t}}$ and $p_{\underline{t}}$, as it is easier to control the early stopping error, i.e., the distance between $p_{\underline{t}}$ and p_0. The following error decomposition is standard [9].

Proposition 3.1. *Under Novikov's condition [9]:*

$$\mathbb{E}_{\mathsf{Q}}\left[\exp\left(\frac{1}{2}\int_0^{T-\underline{t}}\|\widehat{s}(Y_t, T-t) - s(Y_t, T-t)\|^2\,\mathrm{d}t\right)\right] < \infty,$$

where $\mathsf{Q} = \mathsf{Law}(Y_{[0,T-\underline{t}]})$ denotes the path measure of the reverse process (2.2). It holds that

$$\mathsf{KL}(p_{\underline{t}}\|\widehat{q}_{T-\underline{t}}) \le \mathrm{e}^{-2T}\mathsf{KL}(p_0\|\mathsf{N}(0,I)) + \int_{\underline{t}}^{T}\mathbb{E}_{x_t\sim p_t}\left[\|\widehat{s}(x_t,t) - s(x_t,t)\|^2\right]\mathrm{d}t. \tag{3.10}$$

The proof can be found in Appendix B.1. We note that the first term on the right hand side can be replaced by $\mathrm{e}^{-2(T-\underline{t})}\mathsf{KL}(p_{\underline{t}}\|\mathsf{N}(0,I))$ when p_0 is singular w.r.t. $\mathsf{N}(0,I)$, so that it always decays exponentially in T regardless of p_0. Thus it suffices to control the second term; i.e. the score approximation error.

3.2.2 Localized Score Function

As discussed in Section 2.3, strict locality is not preserved in the forward OU process, so that the true score $s \notin \mathscr{H}_r$ in general. It is therefore crucial to control the approximation error of the best possible approximation $s^* \in \mathscr{H}_r$.

Consider taking $\mathscr{U}_j = C^2(\mathbb{R}^{d_{j,r}+1})$ in the localized hypothesis space $\mathscr{H}_r$ (3.1), so that the only constraint in $\mathscr{H}_r$ is the locality structural constraint (note we always consider at least twice

differentiable functions). Then the best possible approximation error can be identified as the *localization error* of the score function. To avoid confusion, we denote $\mathscr{H}_r^*$ as the hypothesis space when we take $\mathscr{U}_j = C^2(\mathbb{R}^{d_{j,r}+1})$.

Motivated by (3.10), we consider the optimal approximation in the $L^2(p_t)$ sense, i.e.,

$$s^* = \mathop{\arg\min}_{s_\theta \in \mathscr{H}_r^*} \int_{\underline{t}}^{T} \int \|s_\theta(x,t) - s(x,t)\|^2 \, p_t(x) \mathrm{d}x \mathrm{d}t$$

$$\Leftrightarrow \ \forall j, \ s_j^*(x,t) = u_j^*(x_{\mathcal{N}_j^r}, t), \quad u_j^* = \mathop{\arg\min}_{u_{\theta,j} \in \mathscr{U}_j} \int_{\underline{t}}^{T} \int \|u_{\theta,j}(x_{\mathcal{N}_j^r}, t) - s_j(x,t)\|^2 p_t(x) \mathrm{d}x \mathrm{d}t.$$

Using the property of conditional expectation, it is straightforward to show that the optimizer is

$$\begin{aligned} u_j^*(x_{\mathcal{N}_j^r}, t) &= \mathbb{E}_{x' \sim p_t} \left[s_j(x', t) \middle| x'_{\mathcal{N}_j^r} = x_{\mathcal{N}_j^r} \right] \\ &= \frac{1}{p_t(x_{\mathcal{N}_j^r})} \int \nabla_j \log p_t(x_{\mathcal{N}_j^r}, x_{\mathcal{N}_j^{r\perp}}) p_t(x_{\mathcal{N}_j^r}, x_{\mathcal{N}_j^{r\perp}}) \mathrm{d}x_{\mathcal{N}_j^{r\perp}}. \end{aligned} \tag{3.11}$$

Here we denote $\mathcal{N}_j^{r\perp} := [b] \setminus \mathcal{N}_j^r$.

Due to the approximate locality (Theorem 2.1), one can expect that the approximation error decays exponentially with the radius r. Before presenting the approximation result, we introduce a quantitative condition [13] characterizing the sparsity of the graph G.

Definition 3.1. *An undirected graph G is called (S,ν)-local if*

$$\forall j \in V, \ r \in \mathbb{N}, \quad |\mathcal{N}_j^r| \le 1 + S r^\nu. \tag{3.12}$$

In the above definition, S denotes the maximal size of the immediate neighbor, and ν denotes the *ambient dimension* of the graph, which controls the growth rate of the neighborhood volume with the radius. Here we require it growing at most polynomially to ensure effective locality. Note the ambient dimension ν is typically a small number. A motivating example for Definition 3.1 is the lattice model $\mathbb{Z}^\nu$, where a naive bound of the neighborhood volume is

$$|\mathcal{N}_j^r| = |\{i \in \mathbb{Z}^\nu : \|i\|_1 \le r\}| \le (2r+1)^\nu < 1 + (3r)^\nu.$$

So that $\mathbb{Z}^\nu$ is $(3^\nu, \nu)$-local.

Now we state the approximation result.

Theorem 3.2. *Let p_0 satisfy the conditions in Theorem 2.1, and its dependency graph is (S,ν)-local. Consider the hypothesis space $\mathscr{H}_r^*$ (3.1) with $\mathscr{U}_j = C^2(\mathbb{R}^{d_{j,r}+1})$. Then there exists an optimal approximation $s^* \in \mathscr{H}_r^*$ such that*

$$\int_{\underline{t}}^{T} \left\| s_j^*(x,t) - s_j(x,t) \right\|_{L^2(p_t)}^2 \mathrm{d}t \le C d_j (r+1)^\nu \mathrm{e}^{-c(r+1)}, \tag{3.13}$$

where C and c are some dimensional independent constants depending on m, M, S, ν, i.e.,

$$C = 2S \max\{1, m^{-1}\} \nu! \kappa^{2\nu+1} \log \kappa, \quad c = -2 \log(1 - \kappa^{-1}).$$

Note (3.13) is independent of $\underline{t}, T$. Moreover, for any $s_\theta \in \mathscr{H}_r^*$, the Pythagorean equality holds

$$\|s_{\theta,j}(x,t) - s_j(x,t)\|_{L^2(p_t)}^2 = \left\| s_{\theta,j}(x,t) - s_j^*(x,t) \right\|_{L^2(p_t)}^2 + \left\| s_j^*(x,t) - s_j(x,t) \right\|_{L^2(p_t)}^2. \tag{3.14}$$

The proof can be found in Appendix B.2. (3.13) provides an upper bound for the hypothesis error of using a localized score function to approximate the true score function. Note the bound is *independent* of the ambient dimension d, although the true score $s_j(x,t)$ is a d-dimensional function. Secondly, the bound decays exponentially (up to a polynomial factor) w.r.t. the radius r, so that a small r is sufficient to achieve a good approximation. Finally, note taking summation over $j \in [b]$ in (3.13) gives the total approximation error

$$\int_0^T \|s_\theta(x,t) - s(x,t)\|_{L^2(p_t)}^2 \,\mathrm{d}t \le Cd(r+1)^\nu \mathrm{e}^{-c(r+1)},$$

which scales linearly with the dimension d.

3.3 Sample Complexity

In this section, we demonstrate the key advantage of the localized diffusion models, i.e., that the sample complexity is independent of the ambient dimension d. We will show that the denoising score matching with the localized hypothesis space $\mathscr{H}_r$ is equivalent to fitting the L^2-optimal localized score in (3.11). Since the localized scores are low-dimensional functions, the sample complexity should be independent of d.

3.3.1 Equivalent to Diffusion Models for Marginals

A key observation is that the localized denoising score matching loss (3.7) is *equivalent* to the j-th component loss of the score function when we use standard diffusion model to approximate the *marginal distribution* $p_0(x_{\mathcal{N}_j^r})$. To be precise, denote its population version as

$$\mathcal{L}_j(u_{\theta,j}) = \mathbb{E}_{x_0\sim p_0} \int_{\underline{t}}^T \mathbb{E}_{\epsilon_t\sim \mathsf{N}(0,I)} \left[\left\| u_{\theta,j}(\alpha_t x_{0,\mathcal{N}_j^r} + \sigma_t \epsilon_{t,\mathcal{N}_j^r}, t) + \sigma_t^{-1}\epsilon_{t,j} \right\|^2 \right] \mathrm{d}t. \tag{3.15}$$

The following proposition shows the equivalence.

Proposition 3.3. *The following equalities hold:*

$$\begin{aligned}
\mathcal{L}_j(u_{\theta,j}) &= \mathbb{E}_{x_{0,\mathcal{N}_j^r}\sim p_0} \int_{\underline{t}}^T \mathbb{E}_{\epsilon_t\sim \mathsf{N}(0,I)} \left[\left\| u_{\theta,j}(\alpha_t x_{0,\mathcal{N}_j^r} + \sigma_t \epsilon_{t,\mathcal{N}_j^r}, t) + \sigma_t^{-1}\epsilon_{t,j} \right\|^2 \right] \mathrm{d}t \\
&= \mathbb{E}_{x_{0,\mathcal{N}_j^r}\sim p_0} \int_{\underline{t}}^T \mathbb{E}_{x_{t,\mathcal{N}_j^r}\sim p_{t|0}(x_{t,\mathcal{N}_j^r}|x_{0,\mathcal{N}_j^r})} \left[\left\| u_{\theta,j}(x_{t,\mathcal{N}_j^r}, t) - \nabla_j \log p_{t|0}(x_{t,\mathcal{N}_j^r}|x_{0,\mathcal{N}_j^r}) \right\|^2 \right] \mathrm{d}t \\
&= \int_{\underline{t}}^T \mathbb{E}_{x_{t,\mathcal{N}_j^r}\sim p_t} \left[\left\| u_{\theta,j}(x_{t,\mathcal{N}_j^r}, t) - u_j^*(x_{t,\mathcal{N}_j^r}, t) \right\|^2 \right] \mathrm{d}t + \text{const.}
\end{aligned}$$

Here u_j^* is the optimal localized approximation of the score function (3.11), and the constant depends only on p_0.

The proof can be found in Appendix B.3. Proposition 3.3 implies that the localized score matching can be regarded as b diffusion models, each of which aims to fit (one component of) the score function of a low-dimensional marginal distribution. Using the minimax results of diffusion models, e.g. [32], one immediately obtains that the sample complexity of the localized score matching is essentially independent of the ambient dimension d.

3.3.2 A Complete Error Analysis

We provide a concrete result below. Following [32], we assume a further boundedness constraint on the hypothesis space $\mathscr{H}_r$ (3.1):

$$\mathscr{H}_r^N = \left\{ s \in \mathscr{H}_r \;\middle|\; \forall j,\ \|s_j(\cdot,t)\|_\infty \lesssim \frac{\log^2 N}{\sigma_t} \right\}. \tag{3.16}$$

The constraint is natural as the score function scales with σ_t^{-1}; see [32] for more discussions. We also assume the following technical regularity conditions on the target distribution.

Assumption 3.1. The target distribution p_0 satisfies the following conditions:

1. (Boundedness) p_0 is supported on $[-M, M]^d$, and its density is upper and lower bounded by some constants C_p, C_p^{-1} respectively.

2. (γ-smoothness) For any $j \in [b]$, its marginal density $p_0(x_{\mathcal{N}_j^r}) \in \mathcal{B}_R(B_{a,b}^\gamma([-M, M]^{d_{j,r}}))$. Here $B_{a,b}^\gamma$ denotes the Besov space with $0 < a, b \le \infty$ and $\gamma > (1/a - 1/2)_+$, and $\mathcal{B}_R$ denotes the ball of radius R in the Besov space.

3. (Boundary smoothness) $p_0(x_{\mathcal{N}_j^r})|_\Omega \in \mathcal{B}_1(C^\infty(\Omega))$, where $\Omega = [-M, M]^{d_{j,r}} \setminus [-M + a_0, M - a_0]^{d_{j,r}}$ is the boundary region for some sufficiently small width $a_0 > 0$. Given sample size N, one can take $a_0 \approx N^{-\frac{1}{d_{\text{eff}}}}$, where d_{eff} is the effective dimension (3.2).

Remark 3.2. *[32] only considers the standard domain $[-1, 1]^d$. It can be simply extended to $[-M, M]^d$ by scaling argument. Denote $p^M := M^d p_0(M\cdot)$, then p^M is supported on $[-1, 1]^d$ and satisfies the same regularity conditions. Note the scaling only affects the radius R of the Besov space, and does not change the scaling of the sample complexity.*

See [32] for more discussions on the regularity conditions. The following theorem provides an overall error analysis by combining Proposition 3.1, Theorem 3.2 and Theorem 4.3 in [32]. We comment that [44] points out a flaw in the proof in [32], but the issue is fixed in [44].

Theorem 3.4. *Let p_0 satisfy Assumption 3.1 and the conditions in Theorem 3.2. Given sample size N, let $\mathscr{H}_r^N$ be the bounded hypothesis space (3.16) with $\mathscr{U}_j = \mathsf{NN}(\mathsf{L}^j, \mathsf{W}^j, \mathsf{S}^j, \mathsf{B}^j)$ (3.4). Denote $n_j = N^{-d_j/(2\gamma + d_j)}$, and choose the hyperparameters*

$$\mathsf{L}^j = \mathcal{O}(\log^4 n_j), \quad \left\|\mathsf{W}^j\right\|_\infty = \mathcal{O}(n_j \log^6 n_j), \quad \mathsf{S}^j = \mathcal{O}(n_j \log^8 n_j), \quad \mathsf{B}^j = n_j^{\mathcal{O}(\log\log n_j)}.$$

choose $\underline{t} = \mathcal{O}(N^{-k})$ for some $k > 0$ and $T \asymp \log N$. Let $\widehat{s}$ be the minimizer of the empirical loss (3.6) in $\mathscr{H}_r^N$. Denote $\widehat{q}_{T-\underline{t}}$ as the sampled distribution using learned score $\widehat{s}$. Then it holds that

$$\mathbb{E}_{\{X^{(i)}\}_{i=1}^N}[\mathsf{KL}(p_{\underline{t}} \| \widehat{q}_{T-\underline{t}})] \le \mathrm{e}^{-2T}\mathsf{KL}(p_0 \| \mathsf{N}(0, I)) + Cd(r+1)^\nu \mathrm{e}^{-c(r+1)} + C' b N^{-\frac{2\gamma}{d_{\text{eff}}+2\gamma}} \log^{16} N. \tag{3.17}$$

Here d_{eff} is the effective dimension (3.2), C, c are dimensional independent constants in Theorem 3.2, and C' is a dimensional independent constant.

The proof can be found in Appendix B.4. There are three sources of error in (3.17):

(1) Approximation error of p_T, which decays exponentially in terminal time T;

(2) Localization error of the score function, which decays exponentially in localization radius r;

(3) Statistical error, which decays polynomially in N, with statistical rate $\frac{2\gamma}{d_{\text{eff}}+2\gamma}$.

Remark 3.3. *(1) Compared to the vanilla method, the localized diffusion models achieve a much faster statistical rate $\frac{2\gamma}{d_{\text{eff}}+2\gamma} \gg \frac{2\gamma}{d+2\gamma}$, and thus potentially mitigate the curse of dimensionality.*

(2) (3.17) indicates a trade off in the choice of localization radius r. A smaller r leads to smaller statistical error but induces larger localization error. Note $d_{\text{eff}} \asymp r^\nu$ (see Definition 3.1), so that the optimal choice is $r^* = \mathcal{O}((\log N)^{\frac{1}{\nu+1}})$. When $\log N \ll d^{\frac{\nu+1}{\nu}}$, one can show that the overall error is greatly reduced compared to the usual statistical error:

$$\mathrm{e}^{-cr^*} + N^{-\frac{2\gamma}{d_{\text{eff}}^*+2\gamma}} \ll N^{-\frac{2\gamma}{d+2\gamma}}.$$

This is usually the case in high-dimensional problems, as one cannot obtain a large sample size N exponentially in d.

(3) We compare the sampled distribution to the early-stopped distribution $p_{\underline{t}}$ by convention. In fact, the early-stopping error can be controlled straightforwardly in Wasserstein distance. For instance, by Lemma 3 in [9], it holds that $\mathsf{W}_2^2(p, p_{\underline{t}}) \lesssim d\underline{t}$. So that the overall error

$$\mathbb{E}_{\{X^{(i)}\}_{i=1}^N}[\mathsf{W}_2^2(p, \widehat{q}_{T-\underline{t}})] \lesssim \mathsf{W}_2^2(p, p_{\underline{t}}) + \mathbb{E}_{\{X^{(i)}\}_{i=1}^N}[\mathsf{W}_2^2(p_{\underline{t}}, \widehat{q}_{T-\underline{t}})] \lesssim dN^{-k} + \mathbb{E}_{\{X^{(i)}\}_{i=1}^N}[\mathsf{KL}(p_{\underline{t}} \| \widehat{q}_{T-\underline{t}})].$$

Here the second inequality uses Talagand's inequality. The early-stopping error does not deteriorate the order of convergence if one take $k \geq \frac{1}{2}$.

4 Numerical Experiments

4.1 Gaussian model

In this section, we verify the quantitative results obtained before using Gaussian models. First, we use randomly generated Gaussian distributions to show that the locality is approximately preserved in OU process. Second, we consider sampling a discretized OU process, and show that a suitable localization radius is important to balance the localization and statistical error.

4.1.1 Approximate locality

Consider localized Gaussian distribution

$$p_0 = \mathsf{N}(0, C_0),$$

where the precision matrix $P_0 := C_0^{-1}$ is a banded matrix s.t.

$$P_0(i, j) = 0, \quad \forall |i - j| > r_0.$$

We will generate random localized precision matrices P_0 with different dimensions and bandwidths, by taking $P_0 = LL^{\mathrm{T}}$, where L is a randomly generated banded lower triangular matrix. As the condition number plays an important role in the locality, we will also record the condition number of the precision matrices.

We consider diffusion models to sample the distribution. The score function admits an explicit form $s(x, t) = \nabla \log p_t(x) = -P_t x$, where

$$P_t := -\nabla^2 \log p_t = (\alpha_t^2 C_0 + \sigma_t^2 I)^{-1}.$$

We will focus on P_t, as the locality of the score function $s(\cdot, t)$ is *equivalent* to the locality of the precision matrix P_t for Gaussians.

First, we show in the top-left plot in Figure 1 that the $|P_t(i, j)|$ is indeed exponentially decaying with $|i - j|$. Here we take a snapshot of the precision matrix at $t = 0.1039$, which is the time with maximal effective localization radius (see bottom-left plot in Figure 1). We note that the precise exponential decay is not chosen artificially, and any snapshot will yield similar results.

We then compute the effective localization radius of P_t, which is defined as the largest r such that the average of the r-th off-diagonal elements is larger than a threshold. More precisely,

$$r_{\mathrm{loc}}(t) := \max\Big\{1 \leq r < d \ : \ \frac{1}{d-r} \sum_{1 \leq i \leq d-r} |P_t(i, i+r)| \geq \epsilon \cdot \frac{1}{d}\mathrm{tr}(P_t)\Big\}. \tag{4.1}$$

We take the threshold rate $\epsilon = 0.001$. We plot the function $r_{\mathrm{loc}}(t)$ for different dimensions d, bandwidths r_0 and condition numbers κ in Figure 1.

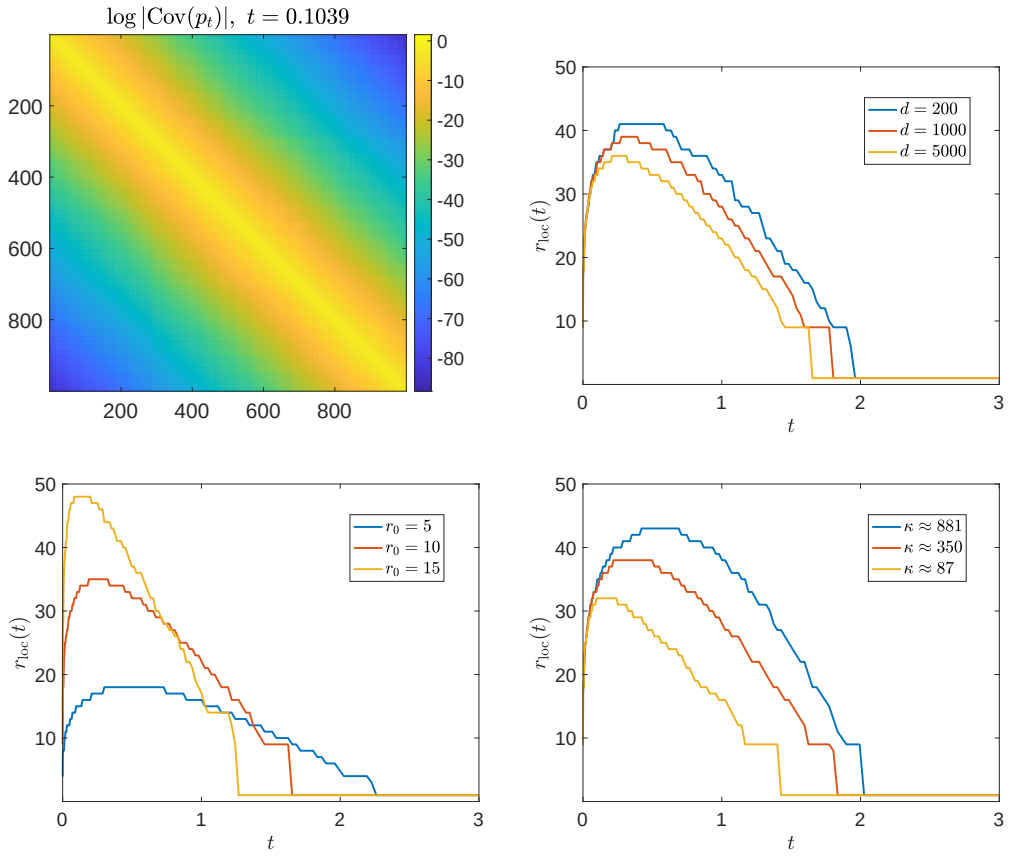

Figure 1: Top-left: The precision matrix P_t at $t = 0.1039$, plotted in $\log |P_t|$ scale. We can see precise exponential decay of $P_t(i, j)$ in $|i - j|$. The rest plots are the localization radius $r_{\text{loc}}(t)$ (4.1) under different problem dimension d, precision matrix bandwidth r_0 and condition number κ. Top-right: $r_{\text{loc}}(t)$ with different dimensions. Here $r_0 = 10$ and the condition numbers are similar ($\kappa \approx 193, 191, 197$). Bottom-left: $r_{\text{loc}}(t)$ with different bandwidths. Here $d = 1,000$ and condition numbers $\kappa \approx 163, 146, 132$. Bottom-right: $r_{\text{loc}}(t)$ with different condition numbers. Here $d = 1,000$ and $r_0 = 10$.

From Figure 1, we can see that the effective localization radius $r_{\text{loc}}(t)$ first increases with t, and then decreases to 1 when t is large. This is due to the fact that P_t can be regarded as an interpolation between P_0 and $P_\infty = I$. Note this is consistent with the theoretical prediction in Theorem 2.1, where the bound of $\|\nabla^2_{ij} \log p_t\|$ first increases with t and then decreases to 0. Next, we can see that the effective localization radius $r_{\text{loc}}(t)$ is almost independent of the dimension d, consistent with our motivation that the locality structure is approximately preserved with dimension independent radius. We can also see that the effective localization radius $r_{\text{loc}}(t)$ is almost linear in the bandwidth r_0, and increases with the condition number κ.

4.1.2 Balance of localization error and statistical error

Consider a discretized OU process $X \in \mathbb{R}^d$ ($d = 101$), where X_n follows the dynamics

$$X_1 \sim \mathsf{N}(0, 1), \quad X_{n+1} = \alpha_h X_n + \sigma_h \xi_n, \quad \xi_n \sim \mathsf{N}(0, 1),$$

where $\alpha_h = \mathrm{e}^{-h}, \sigma_h^2 = 1 - \alpha_h^2$ ($h = 0.2$), and $X_1, \xi_1, \ldots, \xi_{100}$ are independent. Notice X follows a Gaussian distribution

$$p_0(x) = \mathsf{N}(x_1; 0, 1) \prod_{n=1}^{d-1} \mathsf{N}(x_{n+1}; \alpha_h x_n, \sigma_h^2). \tag{4.2}$$

Consider using diffusion model to sample the above distribution. Since the marginals of the forward process are all Gaussians, the score function is a linear function in x. Given data sample $\{X^{(i)}\}_{i=1}^N$, we estimate the score of the linear form $\widehat{s}(t,x) = -\widehat{P}_t x$ by the loss (2.6), which admits an explicit solution

$$\widehat{P}_t = (\alpha_t^2 \widehat{C}_0 + \sigma_t^2 I)^{-1}, \tag{4.3}$$

where $\widehat{C}_0$ is the empirical covariance of $\{X^{(i)}\}_{i=1}$. The non-localized backward process is

$$Y_{t_{n+1}} = Y_{t_n} + \Delta t_n \left(I - 2\widehat{P}_{T-t_n}\right) Y_{t_n} + \sqrt{2\Delta t_n}\xi_n. \tag{4.4}$$

Here $\widehat{P}_t$ is the estimated optimal precision matrix (4.3), $\xi_n \sim \mathsf{N}(0, I), Y_0 \sim \mathsf{N}(0, I)$, and $\Delta t_n = t_{n+1} - t_n$ is the time step. We use the linear variance schedule $\beta_n = (\beta_N - \beta_1)\frac{n-1}{N-1} + \beta_1$ $(1 \le n \le N)$ [21], which corresponds to $\Delta t_n = -\frac{1}{2}\log(1 - \beta_{N-n})$ $(0 \le n \le N-1)$. We take $N = 1,000, \beta_1 = 10^{-4}$ and $\beta_N = 0.05$.

A straightforward localization of (4.4) is

$$\begin{aligned} Y^{\mathrm{loc},r}_{t_{n+1}} &= Y^{\mathrm{loc},r}_{t_n} + \Delta t_n \left(I - 2\widehat{P}^{\mathrm{loc},r}_{T-t_n}\right) Y^{\mathrm{loc},r}_{t_n} + \sqrt{2\Delta t_n}\xi_n, \\ \widehat{P}^{\mathrm{loc},r}_{T-t_n}(i,j) &:= \widehat{P}_{T-t_n}(i,j)\mathbf{1}_{|i-j|\le r}. \end{aligned} \tag{4.5}$$

We will use (4.5) to sample the target distribution with different localization radii r, and compare it to the reference sampling process (4.4). Although the localized score $\widehat{s}^{\mathrm{loc},r}(t,x) = -\widehat{P}^{\mathrm{loc},r}_t x$ in (4.5) is not the minimzer of $\widehat{\mathcal{L}}_N(s_\theta)$ (3.6), it is very close to the minimizer, and it still yields a good approximation.

As all the distributions involved are Gaussian, we can use the sample covariance to measure the localization error. We take data sample size $N = 10^3$ and generated sample size $N_{\mathrm{gen}} = 10^4$. The results are shown in Figure 2. In the top-right plot in Figure 2, we measure the relative ℓ^2-error of the sample covariance

$$\mathrm{err} := \frac{\|\widehat{C} - C\|_2}{\|C\|_2}, \tag{4.6}$$

where $C = P_0^{-1}$ is the true covariance, $\widehat{C}$ is the sample covariance of samples from (4.4) or (4.5), and $\|\cdot\|_2$ is the matrix 2-norm. The reference error is computed using the sample covariance of the non-localized backward process (4.4). For each localization radius, we run 30 independent experiments (with new data sample) and compute the mean and standard deviation of the relative error. The plot shows that as the localization radius increases, the overall error first decays quickly, and then gradually increases. This is due to the balance between the localization error and the statistical error, as shown more clearly in the bottom plots.

In the bottom row of Figure 2, we plot the entrywise error of the sample covariance (normalized by $\|C\|_2$) for different localization radii r. The localization error dominates when the localization radius is small, and we can see that the off-diagonal covariance is not accurately estimated when $r = 4$. The off-diagonal part is approximately recovered when $r = 12$, and the overall error decreases to minimal. As the localization radius r further increases, the statistical error begins to dominate, leading to spurious long-range correlations as observed in the case $r = 35$. This is a well-known phenomenon caused by insufficient sample size [23]. This suggests a suitable localization radius is important to balance the localization and statistical error to reduce the overall error, validating the result in Theorem 3.4.

4.2 Cox-Ingersoll-Ross model

We consider the Cox-Ingersoll-Ross (CIR) model [11, 12]

$$\mathrm{d}X = 2a(b - X)\,\mathrm{d}t + \sigma\sqrt{X}\,\mathrm{d}W_t, \tag{4.7}$$

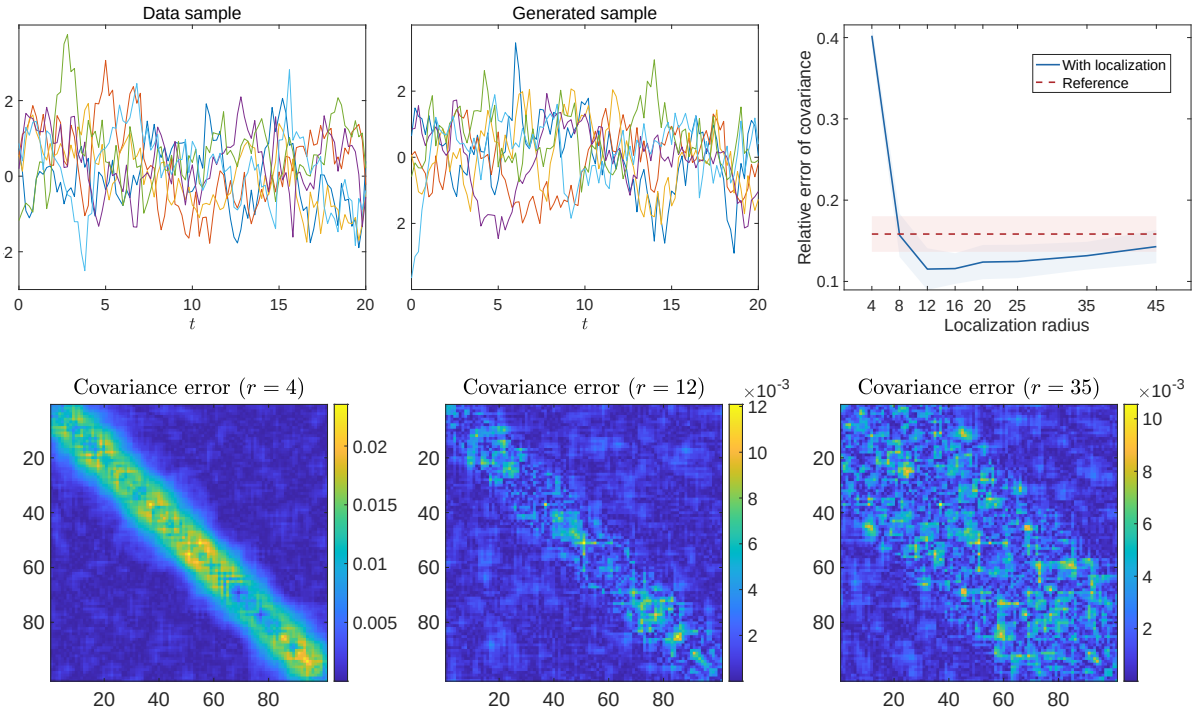

Figure 2: Top left: Trajectories directly sampled from OU process. Top middle: Sampled trajectories using the localized sampling process (4.5) with localization radius $r = 12$. Top right: Relative ℓ^2-error (4.6) of the sample covariance for different localization radii r; the reference error is from the non-localized sampling process (4.4). The shaded area denotes the 1σ region. Bottom: Entrywise error of the sample covariance with different localization radius $r \in \{4, 12, 35\}$.

where W_t is standard one-dimensional Brownian motion. The CIR model (4.7) possesses a closed form solution

$$\frac{X(t)}{c(t)} \sim H(t), \quad c(t) = \frac{\sigma^2}{8a}(1 - e^{-2at}), \tag{4.8}$$

where $H(t)$ is a noncentral χ-squared distribution with $8ab/\sigma^2$ degrees of freedom and noncentrality parameter $c(t)^{-1}e^{-2at}X(0)$.

We generate artificial data by integrating the CIR model (4.7) with an Euler–Maruyama discretization and a time step of $h = 0.01$, sampling at every $\Delta t = 1$ time unit. We determine the score from $M = 50$ independent sample trajectories, each of length $N = 50$, i.e., each trajectory covers 50 time units. We choose $a = 1.136$, $b = 1.1$ and $\sigma = 0.4205$.

For the diffusion model we choose a linear variance schedule with $\beta(t) = (\beta_T - \beta_0)t/T + \beta_0$ with $T = 0.05$, $\beta_T = 0.5$ and $\beta_0 = 0.0001$, and where we sample the diffusion time $t \in [0, T]$ in steps of 0.001 diffusion time units. The discount factor is given by $\alpha(t) = 1 - \beta(t)$. The score is estimated from $5,000$ randomly selected training points, differing in their uniformly sampled diffusion times and initial training sample. To learn the score function we employ a neural network with 3 hidden layers of sizes $2r + 2$, 6 and 3, respectively, with an input dimension of $2r + 2$ coming from the localized states of dimension $2r + 1$ and the diffusion time. The weights of the neural network are determined by minimizing the MSE error using an Adams optimizer with a learning rate $\eta = 0.00005$.

We show in Figure 3 a comparison of the empirical histograms and the auto-correlation

functions of the training data and the data generated by the diffusion model. The histograms are produced from $5,000$ training and generated time series. The auto-correlation function $\langle C(\tau)\rangle$ is computed as an ensemble average over the samples. It is seen that if the localization radius is chosen too small with $r = 0$, i.e., assuming a δ-correlated process, the auto-correlation function rapidly decays as the localized diffusion models have no information about the correlations present in the data. Interestingly, the empirical histogram is relatively well approximated even with $r = 0$. On the other extreme, for large localization radius $r = 20$ the number of independent training samples with $M = 50$ is not sufficiently large to generate $N = 50$-dimensional samples, and the auto-correlation function exhibits an increased variance. We found that a localization radius of $r = 2$ can be employed to yield excellent agreement of the histogram and the auto-correlation function. We checked that varying the localization radius from $r = 2$ to $r = 8$ yields similar results.

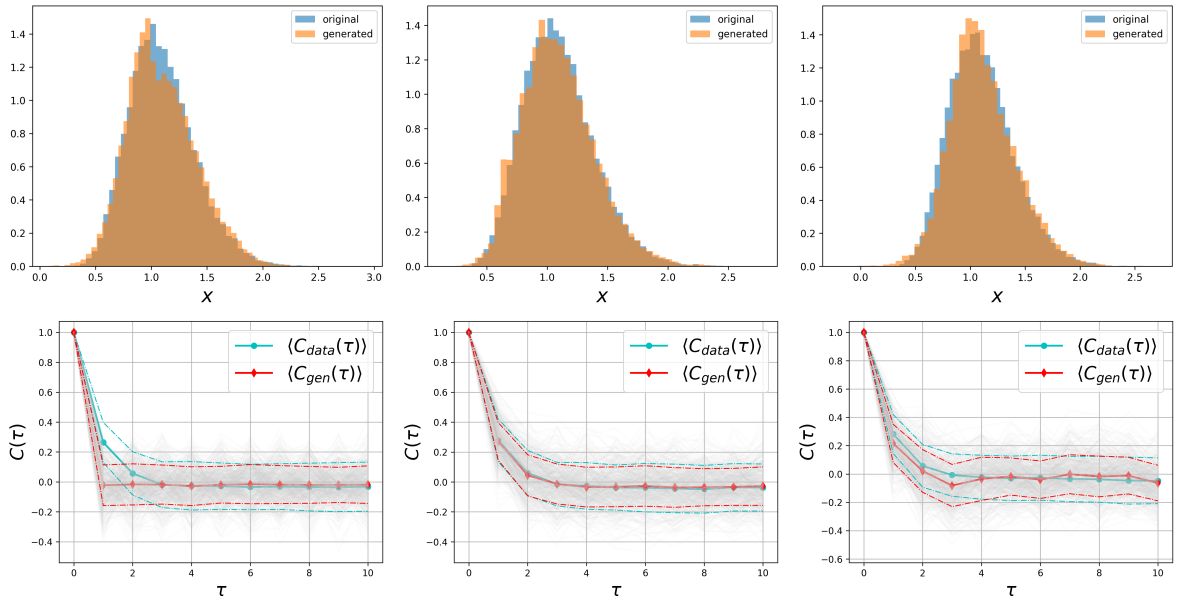

Figure 3: Comparison of the data obtained from the original CIR model (4.7) and from the diffusion model for localization radii $r = 0$ (left), $r = 2$ (middle) and $r = 20$ (right). Top: Empirical histogram. Bottom: Auto-correlation function, averaged over all $2,500$ samples. The dashed lines mark deviations of the sample mean that are 1 standard deviation away. The light grey lines show the individual auto-correlation functions of the generated data.

For the training we estimate the score function at entry i for $i = 2r+1, \dots, N-2r-1$ from the localized state $(x_r)_i = [x_{i-r}, \dots, x_i, \dots, x_{i+r}] \in \mathbb{R}^{2r+1}$. Due to stationarity of the process, each component of the score function $s_i((x_r)_i)$ will be the same except the boundaries, i.e. $i \le r$ or $i \ge d - r$. This allows us to train a single score function which takes a $(2r + 2)$-dimensional input ($2r + 1$ for the localized state and 1 for the diffusion time) to generate a 1-dimensional output of the score function at location $r < i < d - r$. To deal with the boundaries of the time series for $i = 1, \dots, r$ and $i = N - r, \dots, N$, we pad with the time series x, reflected around i. During the training process we have employed independent noise for each localized region. We have checked that the results do not change if the noise in the diffusion model is kept constant for each local input or if varied when cycling through the localized regions.

5 Conclusions

In this work, we study how locality structure can be exploited in diffusion models to sample high-dimensional distributions. We show that the locality structure is approximately preserved in the forward diffusion process, which guarantees that localization error decays exponentially in the localization radius. We propose the localized diffusion model, where we learn the score function within a localized hypothesis space by optimizing a localized score matching loss. We show that the localized diffusion model avoids the curse of dimensionality, and the rate of the statistical error depends on the effective dimension rather than the ambient dimension. Through both theoretical analysis and numerical experiments, we demonstrate that a suitable localization radius can balance the localization and statistical error to reduce the overall error. This validates the effectiveness of localization method in diffusion models for localized distributions.

However, several interesting questions remain open. First, the locality structure should not rely on the log-concavity of the distributions, and it would be interesting to extend the theoretical results to non-log-concave distributions. Second, designing of localized hypothesis space requires prior knowledge of the locality structure. Although it can be learned by many existing methods, it would be interesting to investigate how to combine them, or even learn the locality structure adaptively in the diffusion model. We leave these questions for future work.

Acknowledgements The work of SR has been funded by Deutsche Forschungsgemeinschaft (DFG) - Project-ID 318763901 - SFB1294. GAG acknowledges funding from the Australian Research Council, grant DP220100931. GAG thanks Yuguang Hu and Xiyu Wang for valuable discussions on the implementation of diffusion models. The work of SL is partially supported by NUS Overseas Research Immersion Award (ORIA). The work of XTT is supported by Singapore MOE grant A-8002956-00-00 .

A Proofs in Section 2

A.1 Proof of Theorem 2.1

Proof. Recall

$$p_t(x_t) = \int \mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I) p_0(x_0) \mathrm{d}x_0.$$

We first compute the Hessian of the log density of p_t:

$$\begin{aligned}
\nabla^2 \log p_t(x_t) &= \frac{\nabla^2 p_t(x_t)}{p_t(x_t)} - \frac{\nabla p_t(x_t)}{p_t(x_t)} \frac{\nabla p_t(x_t)^{\mathrm{T}}}{p_t(x_t)} \\
&= \frac{1}{p_t(x_t)} \int \left(-\frac{x_t - \alpha_t x_0}{\sigma_t^2} \right) \left(-\frac{x_t - \alpha_t x_0}{\sigma_t^2} \right)^{\mathrm{T}} \mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I) p_0(x_0) \mathrm{d}x_0 \\
&\quad - \frac{1}{p_t(x_t)} \int \left(-\frac{x_t - \alpha_t x_0}{\sigma_t^2} \right) \mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I) p_0(x_0) \mathrm{d}x_0 \\
&\quad \cdot \frac{1}{p_t(x_t)} \int \left(-\frac{x_t - \alpha_t x_0}{\sigma_t^2} \right)^{\mathrm{T}} \mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I) p_0(x_0) \mathrm{d}x_0 \\
&= \sigma_t^{-4} \mathbb{E}_{p_{0|t}(x_0|x_t)} \left(x_t - \alpha_t x_0 \right) \left(x_t - \alpha_t x_0 \right)^{\mathrm{T}} \\
&\quad - \sigma_t^{-4} \mathbb{E}_{p_{0|t}(x_0|x_t)} \left(x_t - \alpha_t x_0 \right) \mathbb{E}_{p_{0|t}(x_0|x_t)} \left(x_t - \alpha_t x_0 \right)^{\mathrm{T}} \\
&= \sigma_t^{-4} \mathsf{Cov}_{p_{0|t}(x_0|x_t)} \left(x_t - \alpha_t x_0, x_t - \alpha_t x_0 \right) \\
&= \alpha_t^2 \sigma_t^{-4} \mathsf{Cov}_{p_{0|t}(x_0|x_t)} \left(x_0, x_0 \right),
\end{aligned}$$

where $p_{0|t}(x_0|x_t)$ is the distribution of x_0 conditioned on the value of x_t. As a consequence

$$\nabla^2_{ij} \log p_t(x_t) = \alpha_t^2 \sigma_t^{-4} \mathsf{Cov}_{p_{0|t}(x_0|x_t)} (x_{0,i}, x_{0,j}) . \tag{A.1}$$

Consider the conditional distribution $p_{0|t}(x_0|x_t)$, whose log density is

$$\log p_{0|t}(x_0|x_t) = -\log p_t(x_t) + \log p_0(x_0) - \frac{1}{2\sigma_t^2} \|x_t - \alpha_t x_0\|^2 - \frac{d}{2} \log(2\pi\sigma_t^2).$$

Fix x_t, and denote for simplicity $q(x) = p_{0|t}(x|x_t)$. Then

$$\nabla^2 \log q(x) = \nabla^2 \log p_0(x) - \frac{\alpha_t^2}{\sigma_t^2} I.$$

Note by assumption, $\nabla^2_{ij} \log p_0 = 0$ if $i \notin \mathcal{N}_j$, and $mI \preceq -\nabla^2 \log p_0 \preceq MI$. So that

$$\forall i \notin \mathcal{N}_j, \quad \nabla^2_{ij} \log q(x) = 0.$$

$$\left(m + \frac{\alpha_t^2}{\sigma_t^2}\right) I \preceq -\nabla^2 \log p_0 \preceq \left(M + \frac{\alpha_t^2}{\sigma_t^2}\right) I. \tag{A.2}$$

By Proposition 2.2, for any Lipschitz functions f, g, we have

$$\left|\mathsf{Cov}_{q(x)} (f(x_i), g(x_j))\right| \le |f|_{\mathrm{Lip}} |g|_{\mathrm{Lip}} \left(m + \frac{\alpha_t^2}{\sigma_t^2}\right)^{-1} \left(1 - \frac{m\sigma_t^2 + \alpha_t^2}{M\sigma_t^2 + \alpha_t^2}\right)^{\mathsf{d}_G(i,j)} .$$

Recall (A.1), and by definition of the matrix norm,

$$\left\|\nabla^2_{ij} \log p_t(x_t)\right\| = \sup_{\|t_i\|=\|t_j\|=1} t_i^{\mathrm{T}} \nabla^2_{ij} \log p_t(x_t) t_j = \sup_{\|t_i\|=\|t_j\|=1} \alpha_t^2 \sigma_t^{-4} \mathsf{Cov}_{q(x)} \left(t_i^{\mathrm{T}} x_i, t_j^{\mathrm{T}} x_j\right) .$$

Take $f(x_i) = t_i^{\mathrm{T}} x_i$ and $g(x_j) = t_j^{\mathrm{T}} x_j$, and note $|f|_{\mathrm{Lip}} = |g|_{\mathrm{Lip}} = 1$, we obtain

$$\left\|\nabla^2_{ij} \log p_t(x_t)\right\| \le \alpha_t^2 \sigma_t^{-4} \left(m + \frac{\alpha_t^2}{\sigma_t^2}\right)^{-1} \left(1 - \frac{m\sigma_t^2 + \alpha_t^2}{M\sigma_t^2 + \alpha_t^2}\right)^{\mathsf{d}_G(i,j)} .$$

The conclusion follows by noting the above bound holds for all x. □

A.2 Proof of Proposition 2.2

Proof. By subtracting the mean, we assume w.l.o.g. that $\mathbb{E}_{p(x)}[f(x_i)] = \mathbb{E}_{p(x)}[g(x_j)] = 0$. Then

$$\mathsf{Cov}_{p(x)} (f(x_i), g(x_j)) = \int f(x_i) g(x_j) p(x) \mathrm{d}x.$$

Consider the marginal Stein equation [13]

$$-\Delta u_f(x) - \nabla \log p(x) \cdot \nabla u_f(x) = f(x_i).$$

By Lemma A.1, the following gradient estimate of u_f holds:

$$\|\nabla_j u_f\|_\infty \le \frac{1}{m} \left(1 - \frac{m}{M}\right)^{\mathsf{d}_G(i,j)} |f|_{\mathrm{Lip}} .$$

By integration by parts, it holds that

$$\begin{aligned}\int f(x_i)g(x_j)p(x)\mathrm{d}x &= \int \left(-\Delta u_f(x) - \nabla \log p(x)\cdot \nabla u_f(x)\right) g(x_j)p(x)\mathrm{d}x \\ &= \int \nabla u_f(x)\cdot \nabla_x g(x_j)p(x)\mathrm{d}x \\ &\quad + \int \nabla u_f(x)\cdot \nabla p(x) g(x_j)\mathrm{d}x - \int \nabla u_f(x)\cdot \nabla \log p(x) g(x_j)p(x)\mathrm{d}x \\ &= \int \nabla_j u_f(x)\cdot \nabla g(x_j)p(x)\mathrm{d}x.\end{aligned}$$

Here we use $\nabla_{x_i} g(x_j) = 0$ if $i \neq j$. Combined, we obtain

$$\begin{aligned}\left|\mathsf{Cov}_{p(x)}\left(f(x_i), g(x_j)\right)\right| &= \left|\int \nabla_j u_f(x)\cdot \nabla g(x_j)p(x)\mathrm{d}x\right| \\ &\leq \int \|\nabla_j u_f(x)\|\,\|\nabla g(x_j)\|\,p(x)\mathrm{d}x \leq \frac{1}{m}\left(1-\frac{m}{M}\right)^{\mathsf{d}_G(i,j)} |f|_{\mathrm{Lip}}\,|g|_{\mathrm{Lip}}\,.\end{aligned}$$

This completes the proof. □

Lemma A.1. *Suppose p is localized w.r.t. an undirected graph G and is log-concave and smooth, i.e., $\exists 0 < m \leq M < \infty$ s.t. $mI \preceq -\nabla^2 \log p(x) \preceq MI$. For any i and Lipschitz function $f : \mathbb{R}^{d_i} \to \mathbb{R}$, consider the marginal Stein equation*

$$-\Delta_p u_f(x) := -\Delta u_f(x) - \nabla \log p(x)\cdot \nabla u_f(x) = f(x_i) - \mathbb{E}_{p(x)}[f(x_i)]. \tag{A.3}$$

The following gradient estimate holds:

$$\|\nabla_j u_f\|_\infty \leq \frac{1}{m}\left(1-\frac{m}{M}\right)^{\mathsf{d}_G(i,j)} |f|_{\mathrm{Lip}}\,. \tag{A.4}$$

Proof. The proof is based on a refined analysis of that in [13]. Note $\Delta_p = \Delta + \nabla \log p \cdot \nabla$ is the generator of the Langevin dynamics

$$\mathrm{d}X_t^x = \nabla \log p(X_t^x)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \quad X_0^x = x. \tag{A.5}$$

The Stein equation with such generator type operators is known to admit explicit solutions [3]:

$$u_f(x) = -\int_0^\infty \mathbb{E}\left(f(X_{t,i}^x) - \mathbb{E}_\pi[f(x_i)]\right)\mathrm{d}t.$$

See also [13] for a detailed proof. Differentiating w.r.t x_j gives

$$\nabla_j u_f(x) = -\int_0^\infty \mathbb{E}\left[\nabla_j X_{t,i}^x \cdot \nabla f(X_{t,i}^x)\right]\mathrm{d}t.$$

Here $\nabla_j X_{t,i}^x$ is the partial derivative w.r.t x_j of the sample path. Note taking derivative on both sides is valid due to the exponential decay of $\nabla_j X_{t,i}^x$. Since f is Lipschitz, we obtain

$$\|\nabla_j u_f(x)\| \leq \int_0^\infty \mathbb{E}\left[\|\nabla_j X_{t,i}^x\|\left\|\nabla f(X_{t,i}^x)\right\|\right]\mathrm{d}t \leq |f|_{\mathrm{Lip}} \int_0^\infty \mathbb{E}\|\nabla_j X_{t,i}^x\|\mathrm{d}t. \tag{A.6}$$

So that it remains to control $\nabla_j X_{t,i}^x$. Differentiating w.r.t. x in (A.5), we obtain

$$\mathrm{d}\nabla X_t^x = -H_t \cdot \nabla X_t^x \mathrm{d}t, \quad H_t := -\nabla^2 \log p(X_t^x).$$

Denote $\mathsf{G}_t = \mathrm{e}^{mt}\nabla X_t^x$ and $\mathsf{H}_t = H_t - mI$, then it holds that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{G}_t = \mathrm{e}^{mt}\left(m\nabla X_t^x - H_t\nabla X_t^x\right) = -\mathsf{H}_t\mathsf{G}_t, \quad \mathsf{G}_0 = \nabla X_0^x = I.$$

By assumption, $0 \preceq \mathsf{H}_t \preceq (M-m)\mathsf{I}$, and H_t has dependency graph G. By Lemma 6.2 in [13],

$$\|\nabla_j X_t^x\| = \mathrm{e}^{-mt}\|\mathsf{G}_t(i,j)\| \le \mathrm{e}^{-Mt}\sum_{k=\mathsf{d}_G(i,j)}^{\infty}\frac{t^k(M-m)^k}{k!}.$$

Recall (A.6), this implies

$$\begin{aligned}
\|\nabla_j u_f(x)\| &\le |f|_{\mathrm{Lip}}\int_0^\infty \mathbb{E}\|\nabla_j X_{t,i}^x\|\mathrm{d}t \\
&\le |f|_{\mathrm{Lip}}\int_0^\infty \mathrm{e}^{-Mt}\sum_{k=\mathsf{d}_G(i,j)}^{\infty}\frac{t^k(M-m)^k}{k!}\mathrm{d}t \\
&= |f|_{\mathrm{Lip}}\frac{1}{M}\sum_{k=\mathsf{d}_G(i,j)}^{\infty}\left(1-\frac{m}{M}\right)^k = \frac{1}{m}\left(1-\frac{m}{M}\right)^{\mathsf{d}_G(i,j)}|f|_{\mathrm{Lip}}.
\end{aligned}$$

The conclusion follows by noting the above bound holds for all x. □

B Proofs in Section 3

B.1 Proof of Proposition 3.1

Proof. Denote the path measures for the reverse process (2.2) and the sampling process (3.9) as Q and $\widehat{\mathsf{Q}}$ respectively, i.e., $\mathsf{Q}_t = \mathsf{Law}(Y_t), \widehat{\mathsf{Q}}_t = \mathsf{Law}(\widehat{Y}_t)$. By the data-processing inequality, we have

$$\mathsf{KL}(p_{\underline{t}}\|\widehat{q}_{T-\underline{t}}) = \mathsf{KL}(\mathsf{Q}_{T-\underline{t}}\|\widehat{\mathsf{Q}}_{T-\underline{t}}) \le \mathsf{KL}(\mathsf{Q}_{[0,T-\underline{t}]}\|\widehat{\mathsf{Q}}_{[0,T-\underline{t}]}).$$

By the Girsanov theorem [2], we have

$$\begin{aligned}
\mathsf{KL}(\mathsf{Q}_{[0,T-\underline{t}]}\|\widehat{\mathsf{Q}}_{[0,T-\underline{t}]}) &= \mathsf{KL}(\mathsf{Q}_0\|\widehat{\mathsf{Q}}_0) + \int_0^{T-\underline{t}}\mathbb{E}_{y_t\sim\mathsf{Q}_t}\left[\|\widehat{s}(y_t,T-t)-s(y_t,T-t)\|^2\right]\mathrm{d}t \\
&= \mathsf{KL}(p_T\|\mathsf{N}(0,I)) + \int_{\underline{t}}^{T}\mathbb{E}_{x_t\sim p_t}\left[\|\widehat{s}(x_t,t)-s(x_t,t)\|^2\right]\mathrm{d}t.
\end{aligned}$$

By the convergence of the OU process [2], we have

$$\mathsf{KL}(p_T\|\mathsf{N}(0,I)) \le \mathrm{e}^{-2T}\mathsf{KL}(p_0\|\mathsf{N}(0,I)).$$

The conclusion follows by combining the above relations. □

B.2 Proof of Theorem 3.2

Proof. Note the optimal solution is given by (3.11), i.e.,

$$s_j^*(x,t) = \mathbb{E}_{x'\sim p_t}\left[\nabla_j \log p_t(x')\middle| x'_{\mathcal{N}_j^r} = x_{\mathcal{N}_j^r}\right].$$

By (A.2), p_t is $\left(m+\frac{\alpha_t^2}{\sigma_t^2}\right)$-strongly log-concave, so that the conditional distribution $p_t(x_{\mathcal{N}_j^{r\perp}}|x_{\mathcal{N}_j^r})$ is also $\left(m+\frac{\alpha_t^2}{\sigma_t^2}\right)$-strongly log-concave. By the Poincaré inequality [2],

$$\begin{aligned}
\left\|s_j^*(x,t)-s_j(x,t)\right\|_{L^2(p_t)}^2 &= \mathbb{E}_{x_{\mathcal{N}_j^r}\sim p_t}\left[\mathbb{E}_{x'\sim p_t}\left[\left\|s_j^*(x',t)-\nabla_j\log p_t(x')\right\|^2\middle| x'_{\mathcal{N}_j^r} = x_{\mathcal{N}_j^r}\right]\right] \\
&\le \mathbb{E}_{x_{\mathcal{N}_j^r}\sim p_t}\left[\left(m+\frac{\alpha_t^2}{\sigma_t^2}\right)^{-1}\mathbb{E}_{x'\sim p_t}\left[\|\nabla_{\mathcal{N}_j^{r\perp}}\nabla_j\log p_t(x')\|_{\mathrm{F}}^2\middle| x'_{\mathcal{N}_j^r} = x_{\mathcal{N}_j^r}\right]\right].
\end{aligned}$$

Here $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. By Theorem 2.1, it holds that

$$\left\|\nabla_{ij}^2 \log p_t(x)\right\|_\infty \le \frac{\alpha_t^2}{\sigma_t^2\left(m\sigma_t^2+\alpha_t^2\right)}\left(1-\frac{m\sigma_t^2+\alpha_t^2}{M\sigma_t^2+\alpha_t^2}\right)^{\mathsf{d}_G(i,j)}.$$

Since $\|\nabla_{ij}^2 \log p_t(x)\|_{\mathrm{F}}^2 \le d_j \|\nabla_{ij}^2 \log p_t(x)\|_\infty^2$, we obtain that

$$\begin{aligned}
&\mathbb{E}_{x'\sim p_t}\left[\|\nabla_{\mathcal{N}_j^r{}^\perp}\nabla_j \log p_t(x')\|_{\mathrm{F}}^2 \middle| x'_{\mathcal{N}_j^r} = x_{\mathcal{N}_j^r}\right] \\
=& \sum_{i:\mathsf{d}_G(i,j)>r} \mathbb{E}_{x'\sim p_t}\left[\|\nabla_{ij}^2 \log p_t(x')\|_{\mathrm{F}}^2 \middle| x'_{\mathcal{N}_j^r} = x_{\mathcal{N}_j^r}\right] \\
\le& d_j \sum_{i:\mathsf{d}_G(i,j)>r} \frac{\alpha_t^4}{\sigma_t^4\left(m\sigma_t^2+\alpha_t^2\right)^2}\left(1-\frac{m\sigma_t^2+\alpha_t^2}{M\sigma_t^2+\alpha_t^2}\right)^{2\mathsf{d}_G(i,j)}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\int_0^T \left\|s_j^*(x,t)-s_j(x,t)\right\|_{L^2(p_t)}^2 \mathrm{d}t \\
\le& \int_0^T \left[d_j \sum_{i:\mathsf{d}_G(i,j)>r}\left(m+\frac{\alpha_t^2}{\sigma_t^2}\right)^{-1}\frac{\alpha_t^4}{\sigma_t^4\left(m\sigma_t^2+\alpha_t^2\right)^2}\left(1-\frac{m\sigma_t^2+\alpha_t^2}{M\sigma_t^2+\alpha_t^2}\right)^{2\mathsf{d}_G(i,j)}\right]\mathrm{d}t \\
\le& d_j \sum_{k=r+1}^\infty |\{i:\mathsf{d}_G(i,j)=k\}| \int_0^\infty \frac{\alpha_t^4}{\sigma_t^2\left(m\sigma_t^2+\alpha_t^2\right)^3}\left(1-\frac{m\sigma_t^2+\alpha_t^2}{M\sigma_t^2+\alpha_t^2}\right)^{2k}\mathrm{d}t \\
\le& d_j \max\{1,m^{-1}\}\log\kappa \sum_{k=r+1}^\infty |\{i:\mathsf{d}_G(i,j)=k\}|(1-\kappa^{-1})^{2k}.
\end{aligned}$$

The last step uses Lemma B.1. By the Abel transformation and the sparsity assumption (3.12),

$$\begin{aligned}
\sum_{k=r+1}^\infty |\{i:\mathsf{d}_G(i,j)=k\}|(1-\kappa^{-1})^{2k} &= \sum_{k=r+1}^\infty \left[|\mathcal{N}_j^k|-|\mathcal{N}_j^{k-1}|\right](1-\kappa^{-1})^{2k} \\
&= \sum_{k=r+1}^\infty |\mathcal{N}_j^k|\left[(1-\kappa^{-1})^{2k}-(1-\kappa^{-1})^{2(k+1)}\right] - |\mathcal{N}_j^r|(1-\kappa^{-1})^{2(r+1)} \\
&\le S\kappa^{-1}(2-\kappa^{-1})\sum_{k=r+1}^\infty k^\nu(1-\kappa^{-1})^{2k} \le 2S\kappa^{-1}(1-\kappa^{-1})^{2r}\sum_{k=1}^\infty (k+r)^\nu(1-\kappa^{-1})^{2k}.
\end{aligned}$$

One can show that $\sum_{k\in\mathbb{Z}_+} k^n x^k \le n! x(1-x)^{-n-1}$ (see Lemma A.2 in [13]), so that

$$\begin{aligned}
\sum_{k=1}^\infty (k+r)^\nu(1-\kappa^{-1})^{2k} &= \sum_{k=1}^\infty \left(1+\frac{r}{k}\right)^\nu k^\nu(1-\kappa^{-1})^{2k} \le (r+1)^\nu \sum_{k=1}^\infty k^\nu(1-\kappa^{-1})^{2k} \\
&\le (r+1)^\nu \nu!(1-\kappa^{-1})^2[1-(1-\kappa^{-1})^2]^{-\nu-1} \le (r+1)^\nu\nu!(1-\kappa^{-1})^2\kappa^{2(\nu+1)}.
\end{aligned}$$

Combining the above inequalities, we obtain

$$\begin{aligned}
&\int_{\underline{t}}^T \left\|s_j^*(x,t)-s_j(x,t)\right\|_{L^2(p_t)}^2 \mathrm{d}t \le \int_0^T \left\|s_j^*(x,t)-s_j(x,t)\right\|_{L^2(p_t)}^2 \mathrm{d}t \\
&\le d_j \max\{1,m^{-1}\}\log\kappa \cdot 2S\kappa^{-1}(1-\kappa^{-1})^{2r}\cdot(r+1)^\nu\nu!(1-\kappa^{-1})^2\kappa^{2(\nu+1)} \\
&= Cd_j(r+1)^\nu(1-\kappa^{-1})^{2(r+1)}.
\end{aligned}$$

where we denote $C = 2S\max\{1, m^{-1}\}\nu!\kappa^{2\nu+1}\log\kappa$.

The second claim follows from the property of conditional expectation:

$$\begin{aligned}
\|s_{\theta,j}(x,t) - s_j(x,t)\|^2_{L^2(p_t)} &= \|u_{\theta,j}(x_{\mathcal{N}_j^r},t) - s_j(x,t)\|^2_{L^2(p_t)} \\
&= \mathbb{E}_{x_{\mathcal{N}_j^r}\sim p_t}\left[\mathbb{E}_{x'\sim p_t}\left[\|u_{\theta,j}(x_{\mathcal{N}_j^r},t) - u_j^*(x_{\mathcal{N}_j^r},t) + u_j^*(x_{\mathcal{N}_j^r},t) - s_j(x',t)\|^2 \middle| x'_{\mathcal{N}_j^r} = x_{\mathcal{N}_j^r}\right]\right] \\
&= \mathbb{E}_{x_{\mathcal{N}_j^r}\sim p_t}\left[\|u_{\theta,j}(x_{\mathcal{N}_j^r},t) - u_j^*(x_{\mathcal{N}_j^r},t)\|^2\right] \\
&\quad + \mathbb{E}_{x_{\mathcal{N}_j^r}\sim p_t}\left[\mathbb{E}_{x'\sim p_t}\left[\|u_j^*(x_{\mathcal{N}_j^r},t) - s_j(x',t)\|^2 \middle| x'_{\mathcal{N}_j^r} = x_{\mathcal{N}_j^r}\right]\right] \\
&= \|s_{\theta,j}(x,t) - s_j^*(x,t)\|^2_{L^2(p_t)} + \|s_j^*(x,t) - s_j(x,t)\|^2_{L^2(p_t)}.
\end{aligned}$$

This completes the proof. □

Lemma B.1. *Let $\kappa = M/m \geq 1$ and $k \geq 1$. It holds that*

$$\int_0^\infty \frac{\alpha_t^4}{\sigma_t^2\left(m\sigma_t^2 + \alpha_t^2\right)^3}\left(1 - \frac{m\sigma_t^2 + \alpha_t^2}{M\sigma_t^2 + \alpha_t^2}\right)^{2k} \mathrm{d}t \leq \max\{1, m^{-1}\}\log\kappa(1-\kappa^{-1})^{2k}.$$

Proof. Denote $\lambda = \dfrac{\alpha_t^2}{\sigma_t^2} = \dfrac{\mathrm{e}^{-2t}}{1-\mathrm{e}^{-2t}}$, then $\sigma_t^2 = \dfrac{1}{1+\lambda}$ and $\dfrac{\mathrm{d}\lambda}{\mathrm{d}t} = -2\lambda(1+\lambda)$. The integral is

$$\int_0^\infty \frac{\lambda^2(1+\lambda)^2}{(m+\lambda)^3}\left(1 - \frac{m+\lambda}{M+\lambda}\right)^{2k} \frac{\mathrm{d}\lambda}{2\lambda(1+\lambda)} = \int_0^\infty \frac{\lambda(1+\lambda)}{2\,(m+\lambda)^3}\left(1 - \frac{m+\lambda}{M+\lambda}\right)^{2k} \mathrm{d}\lambda.$$

Let $x = \lambda/m$, and the integral can be bounded by

$$\int_0^\infty \frac{mx(1+mx)}{2\,(m+mx)^3}\left(1 - \frac{m+mx}{M+mx}\right)^{2k} m\mathrm{d}x \leq \frac{\max\{1,m\}}{2m}\int_0^\infty \frac{x}{(1+x)^2}\left(1 - \frac{1+x}{\kappa+x}\right)^{2k} \mathrm{d}x.$$

Notice

$$\begin{aligned}
\frac{1}{(1-\kappa^{-1})^{2k}}\int_0^\infty \frac{x}{(1+x)^2}\left(1 - \frac{1+x}{\kappa+x}\right)^{2k} \mathrm{d}x &= \int_0^\infty \frac{x}{(1+x)^2}\left(\frac{\kappa}{\kappa+x}\right)^{2k} \mathrm{d}x \\
&= \int_0^\infty \frac{y}{(\kappa^{-1}+y)^2}\left(\frac{1}{1+y}\right)^{2k} \mathrm{d}y \leq \int_0^\infty \frac{y}{(\kappa^{-1}+y)^2}\left(\frac{1}{1+y}\right)^{2} \mathrm{d}y \\
&< \int_0^{\kappa^{-1}} \kappa^2 y\mathrm{d}y + \int_{\kappa^{-1}}^1 \frac{\mathrm{d}y}{y} + \int_1^\infty \frac{\mathrm{d}y}{y^3} = 1 + \log\kappa \leq 2\log\kappa.
\end{aligned}$$

The conclusion follows by combining the above inequalities. □

B.3 Proof of Proposition 3.3

Proof. The first equality directly follows from the definition (3.15). Since only $x_{0,\mathcal{N}_j^r}$ is involved, it suffices to take expectation w.r.t. the marginal distribution $p(x_{\mathcal{N}_j^r})$.

For the second inequality, notice

$$p_{t|0}(x_{t,\mathcal{N}_j^r}|x_{0,\mathcal{N}_j^r}) = \mathsf{N}(x_{t,\mathcal{N}_j^r}; \alpha_t x_{0,\mathcal{N}_j^r}, \sigma_t^2 I).$$

It holds that

$$\nabla_j \log p_{t|0}(x_{t,\mathcal{N}_j^r}|x_{0,\mathcal{N}_j^r}) = -\sigma_t^{-2}(x_{t,j} - \alpha_t x_{0,j}).$$

Note $x_{t,\mathcal{N}_j^r} = \alpha_t x_{0,\mathcal{N}_j^r} + \sigma_t \epsilon_t \sim p_{t|0}(x_{t,\mathcal{N}_j^r}|x_{0,\mathcal{N}_j^r})$ if $\epsilon_t \sim \mathsf{N}(0, I_r)$, so that

$$\begin{aligned}
&\mathbb{E}_{x_{t,\mathcal{N}_j^r} \sim p_{t|0}(x_{t,\mathcal{N}_j^r}|x_{0,\mathcal{N}_j^r})} \left[\left\| u_{\theta,j}(x_{t,\mathcal{N}_j^r}, t) - \nabla_j \log p_{t|0}(x_{t,\mathcal{N}_j^r}|x_{0,\mathcal{N}_j^r}) \right\|^2 \right] \\
&\qquad = \mathbb{E}_{\epsilon_t \sim \mathsf{N}(0,I)} \left[\left\| u_{\theta,j}(\alpha_t x_{0,\mathcal{N}_j^r} + \sigma_t \epsilon_{t,\mathcal{N}_j^r}, t) + \sigma_t^{-1} \epsilon_{t,j} \right\|^2 \right].
\end{aligned}$$

This verifies the second inequality.

For the third inequality, we first claim that

$$u_j^*(x_{t,\mathcal{N}_j^r}, t) = \nabla_j \log p_t(x_{t,\mathcal{N}_j^r}). \tag{B.1}$$

Given this, the third inequality follows from the basic trick in denoising score matching: take $y = x_{t,\mathcal{N}_j^r}, z = x_{0,\mathcal{N}_j^r}$ and $\pi(y,z) = p_{t,0}(x_{t,\mathcal{N}_j^r}, x_{0,\mathcal{N}_j^r})$ in the following identity:

$$\begin{aligned}
&\mathbb{E}_{z\sim\pi(z)} \mathbb{E}_{y\sim\pi(y|z)} \left\| s_\theta(y) - \nabla_y \log \pi(y|z) \right\|^2 \\
&= \mathbb{E}_{z\sim\pi(z)} \mathbb{E}_{y\sim\pi(y|z)} \left[\left\| s_\theta(y) \right\|^2 - 2(s_\theta(y))^{\mathrm{T}} \nabla_y \log \pi(y|z) + \left\| \nabla_y \log \pi(y|z) \right\|^2 \right] \\
&= \mathbb{E}_{z\sim\pi(z)} \mathbb{E}_{y\sim\pi(y|z)} \left[\left\| s_\theta(y) \right\|^2 + 2\mathrm{tr}\left(\nabla s_\theta(y) \right) + \left\| \nabla_y \log \pi(y|z) \right\|^2 \right] \\
&= \mathbb{E}_{y\sim\pi(y)} \left[\left\| s_\theta(y) \right\|^2 + 2\mathrm{tr}\left(\nabla s_\theta(y) \right) + \left\| \nabla_y \log \pi(y) \right\|^2 \right] + \mathrm{const} \\
&= \mathbb{E}_{y\sim\pi(y)} \left\| s_\theta(y) - \nabla_y \log \pi(y) \right\|^2 + \mathrm{const}.
\end{aligned}$$

Here the second inequality follows from integration by parts; in the third inequality, we take

$$\mathrm{const} = \mathbb{E}_{z\sim\pi(z)} \mathbb{E}_{y\sim\pi(y|z)} \left\| \nabla_y \log \pi(y|z) \right\|^2 - \mathbb{E}_{y\sim\pi(y)} \left\| \nabla_y \log \pi(y) \right\|^2,$$

which is independent of θ; the last equality follows from the same integration by parts trick.

It then suffices to prove (B.1). Note that

$$\begin{aligned}
u_j^*(x_{t,\mathcal{N}_j^r}, t) &= \mathbb{E}_{x'_t \sim p_t} \left[s_j(x'_t, t) \middle| x'_{t,\mathcal{N}_j^r} = x_{t,\mathcal{N}_j^r} \right] \\
&= \frac{1}{p_t(x_{\mathcal{N}_j^r})} \int \nabla_j \log p_t(x_{t,\mathcal{N}_j^r}, x_{t,\mathcal{N}_j^{r\perp}}) p_t(x_{t,\mathcal{N}_j^r}, x_{t,\mathcal{N}_j^{r\perp}}) \mathrm{d}x_{t,\mathcal{N}_j^{r\perp}} \\
&= \frac{\int \nabla_j p_t(x_{t,\mathcal{N}_j^r}, x_{t,\mathcal{N}_j^{r\perp}}) \mathrm{d}x_{t,\mathcal{N}_j^{r\perp}}}{\int p_t(x_{t,\mathcal{N}_j^r}, x_{t,\mathcal{N}_j^{r\perp}}) \mathrm{d}x_{t,\mathcal{N}_j^{r\perp}}}.
\end{aligned}$$

Since

$$p_t(x_t) = \int \mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I) p_0(x_0) \mathrm{d}x_0.$$

$$\Rightarrow \ \nabla_j p_t(x_t) = \int \left(-\sigma_t^{-2}(x_{t,j} - \alpha_t x_{0,j}) \right) \mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I) p_0(x_0) \mathrm{d}x_0.$$

So that

$$\begin{aligned}
u_j^*(x_{t,\mathcal{N}_j^r}, t) &= \frac{\int \left(-\sigma_t^{-2}(x_{t,j} - \alpha_t x_{0,j}) \right) \mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I) p_0(x_0) \mathrm{d}x_0 \mathrm{d}x_{t,\mathcal{N}_j^{r\perp}}}{\int \mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I) p_0(x_0) \mathrm{d}x_0 \mathrm{d}x_{t,\mathcal{N}_j^{r\perp}}} \\
&= \frac{\int \left(-\sigma_t^{-2}(x_{t,j} - \alpha_t x_{0,j}) \right) \mathsf{N}(x_{t,\mathcal{N}_j^r}; \alpha_t x_{0,\mathcal{N}_j^r}, \sigma_t^2 I) p_0(x_{0,\mathcal{N}_j^r}) \mathrm{d}x_{0,\mathcal{N}_j^r}}{\int \mathsf{N}(x_{t,\mathcal{N}_j^r}; \alpha_t x_{0,\mathcal{N}_j^r}, \sigma_t^2 I) p_0(x_{0,\mathcal{N}_j^r}) \mathrm{d}x_{0,\mathcal{N}_j^r}}.
\end{aligned}$$

On the other hand,

$$\nabla_j \log p_t(x_{t,\mathcal{N}_j^r}) = \frac{\nabla_j p_t(x_{t,\mathcal{N}_j^r})}{p_t(x_{t,\mathcal{N}_j^r})} = \frac{\int \nabla_j \mathsf{N}(x_{t,\mathcal{N}_j^r}; \alpha_t x_{0,\mathcal{N}_j^r}, \sigma_t^2 I) p_0(x_{0,\mathcal{N}_j^r}) \mathrm{d}x_{0,\mathcal{N}_j^r}}{\int \mathsf{N}(x_{t,\mathcal{N}_j^r}; \alpha_t x_{0,\mathcal{N}_j^r}, \sigma_t^2 I) p_0(x_{0,\mathcal{N}_j^r}) \mathrm{d}x_{0,\mathcal{N}_j^r}}$$
$$= \frac{\int \left(-\sigma_t^{-2}(x_{t,j} - \alpha_t x_{0,j})\right) \mathsf{N}(x_{t,\mathcal{N}_j^r}; \alpha_t x_{0,\mathcal{N}_j^r}, \sigma_t^2 I) p_0(x_{0,\mathcal{N}_j^r}) \mathrm{d}x_{0,\mathcal{N}_j^r}}{\int \mathsf{N}(x_{t,\mathcal{N}_j^r}; \alpha_t x_{0,\mathcal{N}_j^r}, \sigma_t^2 I) p_0(x_{0,\mathcal{N}_j^r}) \mathrm{d}x_{0,\mathcal{N}_j^r}} = u_j^*(x_{t,\mathcal{N}_j^r}, t).$$

This completes the proof. □

B.4 Proof of Theorem 3.4

Proof. By the Pythagorean equality (3.14),

$$\mathbb{E}_{x_t \sim p_t}\left[\|\widehat{s}(x_t,t) - s(x_t,t)\|^2\right] = \sum_{j=1}^b \mathbb{E}_{x_t \sim p_t}\left[\|\widehat{s}_j(x_t,t) - s_j(x_t,t)\|^2\right]$$
$$= \sum_{j=1}^b \mathbb{E}_{x_t \sim p_t}\left[\left\|\widehat{s}_j(x_t,t) - s_j^*(x_t,t)\right\|^2\right] + \sum_{j=1}^b \mathbb{E}_{x_t \sim p_t}\left[\left\|s_j^*(x_t,t) - s_j(x_t,t)\right\|^2\right].$$

Combining Proposition 3.1 and Theorem 3.2, we obtain

$$\begin{aligned}
\mathsf{KL}(p_{\underline{t}} \| \widehat{q}_{T-\underline{t}}) &\le \mathrm{e}^{-2T}\mathsf{KL}(p_0 \| \mathsf{N}(0,I)) + \int_{\underline{t}}^T \mathbb{E}_{x_t \sim p_t}\left[\|\widehat{s}(x_t,t) - s(x_t,t)\|^2\right] \mathrm{d}t \\
&= \mathrm{e}^{-2T}\mathsf{KL}(p_0 \| \mathsf{N}(0,I)) + \int_{\underline{t}}^T \mathbb{E}_{x_t \sim p_t}\left[\|s^*(x_t,t) - s(x_t,t)\|^2\right] \mathrm{d}t + \mathcal{R} \\
&\le \mathrm{e}^{-2T}\mathsf{KL}(p_0 \| \mathsf{N}(0,I)) + Cd(r+1)^\nu \mathrm{e}^{-c(r+1)} + \mathcal{R},
\end{aligned}$$

where we denote

$$\mathcal{R} = \sum_{j=1}^b \mathcal{R}_j, \quad \mathcal{R}_j = \int_{\underline{t}}^T \mathbb{E}_{x_t \sim p_t}\left[\left\|\widehat{s}_j(x_t,t) - s_j^*(x_t,t)\right\|^2\right] \mathrm{d}t.$$

By Proposition 3.3, $\mathcal{R}_j$ is the j-th component loss of the score function when we use a standard diffusion model to approximate the marginal distribution $p_0(x_{N_j^r})$. Note one can use the same constructive solution as in [32] for the marginal target $p_0(x_{\mathcal{N}_j^r})$ with only the j-th component output as the constructive solution for $\widehat{s}_j$, and the statistic error analysis similarly applies.

Therefore, we can take the same hyperparameters as in [32]:

$$\mathsf{L}^j = \mathcal{O}(\log^4 n_j), \quad \left\|\mathsf{W}^j\right\|_\infty = \mathcal{O}(n_j \log^6 n_j), \quad \mathsf{S}^j = \mathcal{O}(n_j \log^8 n_j), \quad \mathsf{B}^j = n_j^{\mathcal{O}(\log\log n_j)},$$

where $n_j = N^{-d_j/(2\gamma+d_j)}$. Note n, N in our paper correspond to N, n in [32] respectively. Similarly for the time interval choices: $\underline{t} = \mathcal{O}(N^{-k})$ for some $k > 0$ and $T \asymp \log N$. The j-th component loss $\mathcal{R}_j$ is smaller than the overall score matching loss, which is further bounded in Theorem 4.3 in [32]:

$$\mathbb{E}_{\{X^{(i)}\}_{i=1}^N}[\mathcal{R}_j] \le C' N^{-\frac{2\gamma}{d_j+2\gamma}} \log^{16} N.$$

Therefore,

$$\mathbb{E}_{\{X^{(i)}\}_{i=1}^N}[\mathcal{R}] = \sum_{j=1}^b \mathbb{E}_{\{X^{(i)}\}_{i=1}^N}[\mathcal{R}_j] \le C' b N^{-\frac{2\gamma}{d_{\mathrm{eff}}+2\gamma}} \log^{16} N.$$

This completes the proof. □

References

[1] I. Azangulov, G. Deligiannidis, and J. Rousseau, *Convergence of diffusion models under the manifold hypothesis in high-dimensions*, arXiv preprint arXiv:2409.18804, (2024).

[2] D. Bakry, I. Gentil, and M. Ledoux, *Analysis and geometry of Markov diffusion operators*, vol. 348 of Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer, Cham, 2014.

[3] A. D. Barbour, *Stein's method for diffusion approximations*, Probab. Theory Related Fields, 84 (1990), pp. 297–322.

[4] J. Benton, V. D. Bortoli, A. Doucet, and G. Deligiannidis, *Nearly d-linear convergence bounds for diffusion models via stochastic localization*, in The Twelfth International Conference on Learning Representations, 2024.

[5] J. Besag, *Spatial interaction and the statistical analysis of lattice systems*, J. Roy. Statist. Soc. Ser. B, 36 (1974), pp. 192–236.

[6] H. Chen, H. Lee, and J. Lu, *Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions*, in Proceedings of the 40th International Conference on Machine Learning, vol. 202 of Proceedings of Machine Learning Research, PMLR, 23–29 Jul 2023, pp. 4735–4763.

[7] M. Chen, K. Huang, T. Zhao, and M. Wang, *Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data*, in Proceedings of the 40th International Conference on Machine Learning, vol. 202 of Proceedings of Machine Learning Research, PMLR, 23–29 Jul 2023, pp. 4672–4712.

[8] M. Chen, S. Mei, J. Fan, and M. Wang, *An overview of diffusion models: Applications, guided generation, statistical rates and optimization*, arXiv preprint arXiv:2404.07771, (2024).

[9] S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. 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, 2023.

[10] G. Conforti, A. Durmus, and M. G. Silveri, *KL convergence guarantees for score diffusion models under minimal data assumptions*, SIAM Journal on Mathematics of Data Science, 7 (2025), pp. 86–109.

[11] J. C. Cox, J. E. Ingersoll, Jr., and S. A. Ross, *An intertemporal general equilibrium model of asset prices*, Econometrica, 53 (1985), pp. 363–384.

[12] ——, *A theory of the term structure of interest rates*, Econometrica, 53 (1985), pp. 385–407.

[13] T. Cui, S. Liu, and X. Tong, *l_inf-approximation of localized distributions*, arXiv preprint arXiv:2410.11771, (2024).

[14] P. Dhariwal and A. Nichol, *Diffusion models beat GANs on image synthesis*, in Advances in Neural Information Processing Systems, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds., vol. 34, Curran Associates, Inc., 2021, pp. 8780–8794.

[15] C. Fefferman, S. Mitter, and H. Narayanan, *Testing the manifold hypothesis*, J. Amer. Math. Soc., 29 (2016), pp. 983–1049.

[16] R. Flock, S. Liu, Y. Dong, and X. T. Tong, *Local MALA-within-Gibbs for Bayesian image deblurring with total variation prior*, to appear in SIAM J. Sci. Comput., (2025).

[17] K. Gatmiry, J. Kelner, and H. Lee, *Learning mixtures of Gaussians using diffusion models*, arXiv preprint arXiv:2404.18869, (2024).

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, in Advances in Neural Information Processing Systems, vol. 27, Curran Associates, Inc., 2014.

[19] G. A. Gottwald, F. Li, S. Reich, and Y. Marzouk, *Stable generative modeling using Schrödinger bridges*, tech. rep., arXiv:2401.04372, 2024. to appear in Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences.

[20] G. A. Gottwald and S. Reich, *Localized Schrödinger bridge sampler*, arXiv preprint arXiv:2409.07968, (2024).

[21] J. Ho, A. Jain, and P. Abbeel, *Denoising diffusion probabilistic models*, in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds., vol. 33, Curran Associates, Inc., 2020, pp. 6840–6851.

[22] ——, *Denoising diffusion probabilistic models*, Advances in neural information processing systems, 33 (2020), pp. 6840–6851.

[23] P. L. Houtekamer and H. L. Mitchell, *A Sequential Ensemble Kalman Filter for Atmospheric Data Assimilation*, Monthly Weather Review, 129 (2001), p. 123.

[24] D. P. Kingma and M. Welling, *Auto-encoding variational Bayes*, arXiv preprint arXiv:1312.6114, (2013).

[25] W. Kohn, *Density functional and density matrix method scaling linearly with the number of atoms*, Phys. Rev. Lett., 76 (1996), pp. 3168–3171.

[26] D. Koller and N. Friedman, *Probabilistic graphical models*, Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, 2009. Principles and techniques.

[27] L. D. Landau, E. M. Lifshitz, E. M. Lifshitz, and L. Pitaevskii, *Statistical physics: Theory of the condensed state*, vol. 9, Butterworth-Heinemann, 1980.

[28] H. Lee, J. Lu, and Y. Tan, *Convergence for score-based generative modeling with polynomial complexity*, in Advances in Neural Information Processing Systems, vol. 35, Curran Associates, Inc., 2022, pp. 22870–22882.

[29] S. Z. Li, *Markov random field modeling in image analysis*, Advances in Pattern Recognition, Springer-Verlag London, Ltd., London, third ed., 2009. With forewords by Anil K. Jain and Rama Chellappa.

[30] S. Mei and Y. Wu, *Deep networks as denoising algorithms: Sample-efficient learning of diffusion models in high-dimensional graphical models*, IEEE Transactions on Information Theory, 71 (2025), pp. 2930–2954.

[31] M. Morzfeld, X. T. Tong, and Y. M. Marzouk, *Localization for MCMC: Sampling high-dimensional posterior distributions with local structure*, J. Comput. Phys., 380 (2019), pp. 1–28.

[32] K. Oko, S. Akiyama, and T. Suzuki, *Diffusion models are minimax optimal distribution estimators*, in Proceedings of the 40th International Conference on Machine Learning, vol. 202 of Proceedings of Machine Learning Research, PMLR, 23–29 Jul 2023, pp. 26517–26582.

[33] P. Potaptchik, I. Azangulov, and G. Deligiannidis, *Linear convergence of diffusion models under the manifold hypothesis*, arXiv preprint arXiv:2410.09046, (2024).

[34] P. Rebeschini and R. van Handel, *Can local particle filters beat the curse of dimensionality?*, Ann. Appl. Probab., 25 (2015), pp. 2809–2866.

[35] D. Rezende and S. Mohamed, *Variational inference with normalizing flows*, in Proceedings of the 32nd International Conference on Machine Learning, vol. 37 of Proceedings of Machine Learning Research, Lille, France, 07–09 Jul 2015, PMLR, pp. 1530–1538.

[36] K. Shah, S. Chen, and A. Klivans, *Learning mixtures of Gaussians using the DDPM objective*, in Advances in Neural Information Processing Systems, vol. 36, Curran Associates, Inc., 2023, pp. 19636–19649.

[37] Y. Song and S. Ermon, *Generative modeling by estimating gradients of the data distribution*, in Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., 2019.

[38] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, *Score-based generative modeling through stochastic differential equations*, in International Conference on Learning Representations, 2021.

[39] C. J. Stone, *Optimal global rates of convergence for nonparametric regression*, The Annals of Statistics, 10 (1982), pp. 1040 – 1053.

[40] R. Tang and Y. Yang, *Adaptivity of diffusion models to manifold structures*, in Proceedings of The 27th International Conference on Artificial Intelligence and Statistics, vol. 238 of Proceedings of Machine Learning Research, PMLR, 02–04 May 2024, pp. 1648–1656.

[41] X. T. Tong, M. Morzfeld, and Y. M. Marzouk, *MALA-within-Gibbs samplers for high-dimensional distributions with sparse conditional structure*, SIAM J. Sci. Comput., 42 (2020), pp. A1765–A1788.

[42] P. Vincent, *A connection between score matching and denoising autoencoders*, Neural Computation, 23 (2011), pp. 1661–1674.

[43] A. Wibisono, Y. Wu, and K. Y. Yang, *Optimal score estimation via empirical Bayes smoothing*, in Proceedings of Thirty Seventh Conference on Learning Theory, vol. 247 of Proceedings of Machine Learning Research, PMLR, 30 Jun–03 Jul 2024, pp. 4958–4991.

[44] K. Yakovlev and N. Puchkin, *Generalization error bound for denoising score matching under relaxed manifold assumption*, arXiv preprint arXiv:2502.13662, (2025).

[45] H. Zhang, J. Zhou, Y. Lu, M. Guo, P. Wang, L. Shen, and Q. Qu, *The emergence of reproducibility and consistency in diffusion models*, in Proceedings of the 41st International Conference on Machine Learning, vol. 235 of Proceedings of Machine Learning Research, PMLR, 21–27 Jul 2024, pp. 60558–60590.

[46] J. Zhuo, C. Liu, J. Shi, J. Zhu, N. Chen, and B. Zhang, *Message passing Stein variational gradient descent*, in Proceedings of the 35th International Conference on Machine Learning, vol. 80 of Proceedings of Machine Learning Research, PMLR, 10–15 Jul 2018, pp. 6018–6027.