LOCALIZATION IN SAMPLING: THEORY AND APPLICATIONS

SHUIGEN LIU

B.S., Peking University, China

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF MATHEMATICS

NATIONAL UNIVERSITY OF SINGAPORE

AUGUST 2025

Ph.D. Advisor:
Associate Professor TONG Xin

Ph.D. Coadvisor:
Professor BAO Weizhu

Examiners:
Professor REN Weiqing
Doctor LI Qianxiao

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis. This thesis has also not been submitted for any degree in any university previously.

Shuigen Liu
30 August 2025

ACKNOWLEDGEMENT

I would first like to express my sincerest gratitude to my advisor, Prof. Xin Tong, for his support and guidance throughout my doctoral studies. I have greatly benefied from his broad interest and insightful advice. None of my work here would be possible without his support and mentorship. I would also like to thank my coadvisor, Prof. Weizhu Bao, for his continuous support and high expectations that pushed me to do better. Special thanks go to my committee members, Prof. Weiqing Ren and Dr. Qianxiao Li, for their valuable comments and suggestions that have helped improve this thesis.

I would like to acknowledge Prof. Sebastian Reich and Prof. Youssef Marzouk; it has been a great pleasure to collaborate with them, and I have learned a lot from their insights, expertise and intellectual rigor. I would also like to extend my gratitude to Prof. Tiangang Cui, Prof. Yiqiu Dong, Prof. Georg Gottwald and Dr. Rafael Flock, for their valuable discussions and contributions to the work presented here.

I would also like to acknowledge Dr. Matthew Li, Dr. Ningxin Liu, Aimee Maurais for their enlightening discussions and collaborations. Matthew has also been very supportive and helpful during my visit at MIT. Many thanks goes to my colleagues and friends, Dr. Chushan Wang, Dr. Yifei Li, Dr. Tianqi Wang, Dr. Jinfeng Song, Dr. Tianyun Tang, Dr. Chaoyi Zhao, Qingyu Wu, Tao Zhou, Fei Peng, and many others for their support and friendship. Daily lunches and coffee breaks with my colleagues are among the most treasured memories of my time at NUS. Travels with Qingyu and Chushan also brightened my PhD years. I've been fortunate to have some many wonderful collaborators and friends during my PhD journey.

Finally, I would like to thank my parents for their unconditional love. I am also grateful to my sister for her long term support, as well as her hospitality during all the summer breaks.

Contents

Summary

Sampling from high-dimensional probability distributions is a fundamental challenge in computational mathematics and data science. It is often hindered by the curse of dimensionality (CoD), leading to prohibitive computational and statistical complexity. This thesis introduces and rigorously analyzes the localization method, which exploits sparse dependency structures to develop samplers whose complexity depends only mildly on the ambient dimension.

We begin by formalizing localized distributions as Markov random fields on graphs with polynomially growing neighborhood volumes. To quantify locality, we develop the marginal Stein's method, a novel analytical framework that (1) yields a dimension-independent marginal transport inequality, which bounds marginal 1-Wasserstein distances by local score differences; and (2) establishes exponential decay of correlations between distant components.

Building on this theory, we propose a general localization framework that constructs localized samplers by combining local samplers for the marginals. We study two examples: (1) We apply the MALA-within-Gibbs sampler to an image deblurring problem with smooth approximation, prove that its smoothing error is independent of total dimension, and demonstrate substantial speed-ups via local and parallel implementation. (2) We introduce localized diffusion models, where a localized score function is learned and used. We prove that localization can circumvent CoD with only an exponentially decaying error. We show both theoretically and numerically that a moderate localization radius can balance the statistical and localization error to achieve a better overall performance.

List of Figures

List of Tables

List of Symbols

Set

- $\mathbb{N} = \{0, 1, 2, \dots\}$, $\mathbb{Z}_+ = \{1, 2, \dots\}$.
- $\mathbb{R}_{\geq 0} = \{x \in \mathbb{R} : x \geq 0\}$, $\mathbb{R}_+ = \{x \in \mathbb{R} : x > 0\}$.
- $[n] = \{1, \dots, n\}$ for $n \in \mathbb{Z}_+$.
- $-I = I^{\mathrm{c}} = [n] \setminus I$ for $I \subseteq [n]$; $-k = [n] \setminus \{k\}$ for $k \in [n]$.
- $\#(I)$ or $|I|$: cardinality of a set I.
- $A \times B = \{(a, b) : a \in A, b \in B\}$: Cartesian product.
  $A^{\otimes k} = A \times A \cdots \times A$: k-fold Cartesian product.

Linear Algebra

- $x_I := (x_i)_{i \in I}$ for $x \in \mathbb{R}^d$ and $I \subseteq [d]$.
- $\mathrm{tr}(x) = \sum_{i=1}^{d} x_i$.
- $\|x\|_p$: vector ℓ_p-norm; for $p = 2$, briefly denote $\|x\| = \|x\|_2$.
- $\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}$: matrix ℓ_p-norm; for $p = 2$, briefly denote $\|A\| = \|A\|_2$.
- $\|A\|_{\mathrm{F}} = \sqrt{\sum_{i,j} A_{ij}^2}$: Frobenius norm.
  $\langle A, B \rangle_{\mathrm{F}} = \sum_{i,j} A_{ij} B_{ij}$: Frobenius inner product.
- $x^{\otimes k}$: tensor product of k copies of x.
- Without specified otherwise, the default block decomposition:

$$x = (x_1, \dots, x_{\mathsf{b}}) \in \mathbb{R}^d, \quad x_i \in \mathbb{R}^{d_i}, \quad \sum_{i=1}^{\mathsf{b}} d_i = d.$$

Analysis

- $A = \mathcal{O}(B) \Leftrightarrow \exists C > 0$ s.t. $|A| \leq CB$ for all sufficiently large $B > 0$.
- $\mathbf{1} : \mathcal{X} \to \mathbb{R}$: all 1 function, i.e. $\forall x \in \mathcal{X},\ \mathbf{1}(x) \equiv 1$.
- $\nabla_I f = \nabla_{x_I} f$; $\nabla^k_{I_1,\ldots,I_k} f = \nabla^k_{x_{I_1},\ldots,x_{I_k}} f$.
- $L^p(\pi) = \{u : \|u\|_{L^p(\pi)} < \infty\}$, where

$$\|u\|_{L^p(\pi)} = \left(\int_{\mathbb{R}^d} |u(x)|^p \, \pi(x) \mathrm{d}x\right)^{1/p}.$$

- $|f|_{\mathrm{Lip}} = \sup_{x \neq y} \frac{|f(x)-f(y)|}{\|x-y\|}$.
- $H^1(\pi) = \{u \in H^1_{\mathrm{loc}} : \|u\|_{H^1(\pi)} < \infty\}$, where

$$\|u\|_{H^1(\pi)} = \left(\int_{\mathbb{R}^d} |u(x)|^2 \, \pi(x) \mathrm{d}x + \int_{\mathbb{R}^d} \|\nabla u(x)\|^2 \, \pi(x) \mathrm{d}x\right)^{1/2}.$$

- $H^1_0(\pi) = \{u \in H^1(\pi) : \int u(x)\pi(x)\mathrm{d}x = 0\}$. Denote

$$\|u\|_{H^1_0(\pi)} = \left(\int_{\mathbb{R}^d} \|\nabla u(x)\|^2 \, \pi(x) \mathrm{d}x\right)^{1/2}.$$

Probability

- $\mathcal{P}(\mathcal{X})$: the space of probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$.
- $\mathcal{P}_p(\mathbb{R}^d) := \{\pi \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|x\|^p \, \mathrm{d}\pi(x) < \infty\}$.
- π_i: marginal distribution of π on the i-th component x_i.
- $T_{\#}\pi \in \mathcal{P}(\mathcal{Y})$: pushforward distribution of $\pi \in \mathcal{P}(\mathcal{X})$ under $T : \mathcal{X} \to \mathcal{Y}$.
- $X \perp\!\!\!\perp Y \mid Z$: X and Y are conditional independent given Z.
- $\mathsf{Law}(X)$: the law of a random variable X.
- $\mathsf{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X]) \, (Y - \mathbb{E}[Y])^{\mathrm{T}}]$.
  $\mathsf{Cov}(X) := \mathsf{Cov}(X, X)$.
- $\mathsf{N}(\mu, C)$: Gaussian distribution with mean μ and covariance matrix C.
  $\mathsf{N}(x; \mu, C)$: probability density function of $\mathsf{N}(\mu, C)$ evaluated at x.

- $\mathsf{KL}(\mu\|\nu) = \mathbb{E}_\mu[\log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}]$: Kullback-Leibler (KL) divergence between μ and ν.
- $\mathsf{TV}(\mu, \nu) = \frac{1}{2}\int |\mathrm{d}\mu - \mathrm{d}\nu|$: total variation distance between μ and ν.
- $\mathsf{W}_p(\mu, \nu)$: p-Wasserstein distance.

Graph

- $\mathsf{G} = (\mathsf{V}, \mathsf{E})$: graph G with vertex set V and edge set E.
- $i \sim j \;\Leftrightarrow\; (i, j) \in \mathsf{E}$ for $i, j \in \mathsf{V}$.
- $\mathcal{N}_i := \{j \in \mathsf{V} : i \sim j\}$.
- $\mathsf{d}_\mathsf{G}(i, j)$: graph distance.

Chapter 1

Introduction

This thesis studies the localization method for sampling from high-dimensional probability distributions. Sampling in high dimensions is a fundamental problem in computational mathematics, with a wide range of applications in data science, Bayesian statistics, and machine learning. However, it poses computational challenges due to the *curse of dimensionality* (CoD). This underscores the importance of better understanding and exploiting low-dimensional structures in the target distributions. In this thesis, we study the *locality structure*, which captures sparse dependencies between model components. We propose the *localization method* to exploit the locality structure for developing efficient sampling algorithms in high dimensions. This thesis aims to provide a comprehensive study on both the theoretical and algorithmic aspects of the localization method in sampling.

We introduce the *marginal Stein's method*, a novel analytical framework for quantifying the effects of locality in high-dimensional distributions. Leveraging this method, we derive a *marginal transport inequality* and prove that distributions with a locality structure exhibit *exponential decay of correlations*. We also discuss a Langevin semigroup interpretation of the method, which presents its own theoretical interest. In the algorithmic aspect, we discuss the general principle of *localized sampling*, which effectively reduces a high-dimensional sampling problem to a collection of low-dimensional subproblems. To illustrate the approach, we present in detail the *MALA-within-Gibbs* sampler and the *localized diffusion models*. The theoretical and empirical investigations validate the effectiveness of the localization method in high-dimensional sampling problems.

1.1 Motivation

Sampling from high-dimensional distributions is a fundamental challenge in computational mathematics and data science. Denote $\pi \in \mathcal{P}(\mathbb{R}^d)$ as the target distribution, where $d \gg 1$ is the dimension of the state space. The sampling task is to draw samples

$$X^{(1)},\ X^{(2)}, \ldots, X^{(n)} \sim \pi,$$

This task has many scientific and engineering applications, including uncertainty quantification [107, 50], data assimilation [74, 93], Bayesian inference [106, 13], statistical physics [10, 11], and machine learning [59, 67, 61].

One of the central challenges in sampling is high dimensionality, which may arise from several sources: (i) the large number of model parameters (e.g., the neural networks); (ii) the large data size; (iii) discretization of continuous models; (iv) large and complex systems (e.g., the climate models). Sampling in high dimensions is often hindered by the curse of dimensionality (CoD) [59], which refers to the general phenomenon where the computational or sample complexity of an algorithm grows exponentially with the dimension of the problem. This makes many standard sampling methods difficult to scale to high dimensions.

To mitigate the CoD, it is crucial to understand and exploit low-dimensional structures in the target distributions. *Manifold hypothesis* [45] is one of the widely studied structures, which assumes that the data lies on a low-dimensional manifold. Many dimension reduction methods in sampling are based on this structure, e.g. stochastic spectral methods [79], likelihood-informed subspace (LIS) method [34], Manifold Metropolis-adjusted Langevin Algorithm (MMALA) [51], and variational autoencoders (VAEs) [67].

While manifold hypothesis can cover many applications, there are still important cases left open. A large class of high-dimensional distributions are those with locality structure, which captures sparse dependencies between model components [49, 58, 37, 110]. Such sparse dependencies can be modelled by an *Markov random field*, also known as *undirected graphical model* [24, 69]. It introduces an undirected graph $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ to represent the dependencies between model components. Each edge in G encodes a direct dependency between two components. Or equivalently, two components that are not connected by an edge in G are *conditionally independent*,

i.e. let $(X_u)_{u\in\mathsf{V}}$ be the Markov random field labelled by the vertices in V, then

$$X_u \perp\!\!\!\perp X_v \mid X_{\mathsf{V}\setminus\{u,v\}}, \quad \forall (u,v) \notin \mathsf{E}. \tag{1.1.1}$$

Consequently, a sparse graph G indicates that most components only depend directly on a small subset of others. This sparsity reflects a certain low-dimensional structure, which can be leveraged to mitigate the CoD in sampling. In recent years, this structure has attracted a growing interest, with many new methods developed for efficient sampling [121, 81, 110, 57]. In this thesis, we refer to these methods as *localization methods*, which essentially turn a high-dimensional sampling problem into a series of lower-dimensional problems in an exact or approximate manner by exploiting the locality structure.

However, despite recent advances, the theoretical understanding of these methods remains limited. Most existing approaches are ad hoc, and often developed without rigorous guarantees. This thesis aims to fill this gap by developing a theoretical framework to analyze the locality structure and the localization method in sampling, which provides the foundation for the design and analysis of localized samplers in high dimensions. We will focus on two main questions:

Question 1. How to relate the locality structure to quantitative properties of the target distribution?

Question 2. How to exploit the locality structure to develop efficient sampling algorithms in high dimensions?

1.2 Literature review

1.2.1 Dimension reduction in sampling

Dimension reduction is the common strategy to mitigate the CoD in sampling. One class of methods seeks a low-dimensional subspace that captures the most important directions along which the distribution or the likelihood changes most significantly. These methods include likelihood-informed subspace (LIS) method [34], active subspace method [28], or certified dimension reduction [120]. For instance, in LIS, the subspace can be identified using the leading eigenvectors of a Gram matrix of the gradient of the log-likelihood function, i.e.

$$H = \mathbb{E}_{x\sim\pi}\left[\nabla \log l(x)(\nabla \log l(x))^{\mathrm{T}}\right].$$

These methods are among the most widely applied dimension reduction methods in sampling, due to their low computational cost, simple implementation, and certified approximation error [32, 35, 120].

There are also other methods based on the low-dimensional subspace or manifold. [79] proposes to use spectral methods to find reduced representation of the parameters using polynomial chaos expansion. [51] proposes the Riemann-manifold Hamiltonian Monte Carlo (RMHMC) and Manifold Metropolis-Adjusted Langevin Algorithm (MMALA), which exploit the Riemannian geometry to improve the sampling efficiency. [29] proposes the preconditioned Crank-Nicolson (pCN) sampler, which avoids the curse of dimensionality by operating directly in the infinite-dimensional setting. Modern generative models, such as variational autoencoders (VAEs) [67], generative adversarial networks (GANs) [54], and latent diffusion models [97] all leverage certain low-dimensional latent structures.

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. In recent years, there has been a fast growing interest in sampling methods that leverage locality structures [121, 81, 110, 57]. [81] propose to apply the localization technique in Markov chain Monte Carlo (MCMC) and introduces a localized Metropolis-within-Gibbs sampler. [110] extends this idea and develops the MALA-within-Gibbs sampler, which is proven to admit a dimension independent convergence rate. Beyond MCMC, [121] 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. [57] proposes a localized version of the Schrödinger Bridge (SB) sampler [55], 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. Detailed discussions on these methods are presented in Chapters 4 and 5.

1.2.2 Markov random fields

In probability and physics, locality structure is often modelled by Markov random fields (MRFs), also known as Markov network or undirected graphical models [24, 118, 69]. The concept of MRFs came from attempts to generalize the seminal Ising model [65] to more general settings. Mathematical foundation of MRFs is established in the 1970s [25, 1, 91]. A remarkable result is the *Hammersley-Clifford*

theorem proved by Hammersley and Clifford in an unpublished manuscript [25], which states the equivalence between MRFs and Gibbs random fields, whose density can be factorized over the cliques of a graph (i.e. , c.f. Section 1.5.2):

$$\pi(x) \propto \prod_{C \in \mathcal{C}} \phi_C(x_C),$$

where $\mathcal{C}$ denotes the collection of cliques in the graph, x_C is the restriction of x to the vertices in C, and ϕ_C is the *clique potential*. This factorization is known as *clique factorization*. Also remarkable is the earlier work of Dobrushin [39], Lanford and Ruelle [73] on the existence and uniqueness of Gibbs measures. In particular, it is guaranteed by the renowned *Dobrushin condition* on the conditional distributions [39]. More interesting results on MRFs are documented in [66, 118, 69].

Exponential decay of correlations is an important property of MRFs in the high-temperature or weakly-coupled regime, which states that the correlations between different components decay exponentially with their distance. A vast literature in statistical physics and probability theory has been devoted to establishing this property under different conditions [71, 40, 70, 41]. It has also been studied in different contexts due to its broad range of applications.

A basic formulation of exponential decay of correlations can be cast in a linear algebraic form [38, 6]. Let $A \in \mathbb{R}^{d \times d}$ be a banded and positive matrix, i.e. $A(i,j) = 0$ if $|i-j| > B$ and $mI \preceq A \preceq MI$ for some $0 < m \leq M < \infty$. Then

$$|A^{-1}(i,j)| \lesssim \lambda^{-|i-j|},$$

for some $\lambda \in (0,1)$. In probability language, this means if the precision matrix A of a Gaussian distribution is banded, then the covariance matrix A^{-1} has exponentially decaying properties.

In quantum mechanics, the exponential decay of correlations are abstracted as the *nearsightedness principle* [68], which states the properties of a quantum system are primarily determined by local interactions and are insensitive to distant perturbations. The principle can be formalized as the exponential decay of the density operator $\rho(\boldsymbol{r}, \boldsymbol{r}')$ w.r.t. the physical distance $\|\boldsymbol{r} - \boldsymbol{r}'\|$ [68, 53, 7], i.e.

$$|\rho(\boldsymbol{r}, \boldsymbol{r}')| \lesssim \exp\left(-\alpha \|\boldsymbol{r} - \boldsymbol{r}'\|\right).$$

This enables efficient sparse approximation of ρ, which can be regarded as the

quantum version of localization of distributions.

1.2.3 Localization method

Due to the ubiquitous presence of locality structures, various localization methods has been developed in different fields, including numerical linear algebra [6, 99], spatial statistics [8, 112, 37], data assimilation [63, 58, 92] etc. See [60] for a comprehensive review on more applications in physics, biology, and data science.

In numerical linear algebra, the *incomplete Cholesky preconditioners* [26] are developed for solving large block tridiagonal linear systems. The preconditioner is taken as a banded matrix approximation of the inverse of target tridiagonal matrix. Similar to it is the *sparse Cholesky factorization* [99], where it uses a KL loss to approximate the Cholesky factorization subject to a sparsity pattern $S \subseteq [d] \times [d]$:

$$L = \arg\min_{\hat{L} \in \mathcal{S}} \mathsf{KL}\left(\mathsf{N}(0, C) \| \mathsf{N}(0, (\hat{L}\hat{L}^{\mathrm{T}})^{-1})\right).$$

Here $\mathcal{S} = \{L \in \mathbb{R}^{d\times d} : \forall (i,j) \notin S \ \Rightarrow \ L_{ij} = 0\}$. The computational cost is significantly reduced, achieving nearly linear scaling in space complexity. Localization trick is also used in computation of matrix functions $f(A)$ [6]. A popular approach is Chebyshev polynomial approximation (suppose $\mathrm{spec}(A) \subseteq [-1, 1]$)

$$f(A) \approx P_N(A) := \sum_{k=0}^{N} c_k(f) T_k(A),$$

where T_k is the k-th Chebyshev polynomial. When $A \in \mathbb{R}^{d\times d}$ is banded, the computational cost of $P_N(A)$ can be reduced to $\mathcal{O}(d)$. This is a typical *linear scaling method* in electronic structure computation [53]. For more discussions on localization methods in numerical linear algebra, we refer to [6].

In spatial statistics, the *Vecchia approximation* [112] proposes to approximate Gaussian processes by removing conditioners at large distances, i.e.

$$\pi(X_1)\prod_{i=1}^{d-1}\pi(X_{i+1}|X_{1:i}) \approx \pi(X_1)\prod_{i=1}^{d-1}\pi(X_{i+1}|X_{\mathcal{N}_{i+1}\cap[i]}),$$

where $\mathcal{N}_{i+1} \cap [i]$ only contains a small subset of preceding points, thus largely reduces the complexity of sampling. Various extensions have been proposed based on the nearest-neighbor approximation idea, see a brief review in [37]. It is also

pointed out in [99] that the Vecchia approximation can be interpreted as a localized Cholesky decomposition.

In data assimilation, localization methods [63, 58] are introduced to mitigate spurious long-range correlations arising from small ensemble sizes in the ensemble Kalman filter. For instance, the covariance localization artificially removes or dampens long-range correlations in the ensemble covariance $\widehat{C}$, i.e.

$$\widehat{C}_{\text{loc}} = \Psi \circ \widehat{C}, \quad \Psi_{ij} = \psi(|i-j|),$$

where $\circ$ denotes the Hadamard product and $\psi : \mathbb{N} \to \mathbb{R}_+$ is a rapidly decaying function. Such localization methods have been shown to effectively reduce sampling errors and improve filter accuracy [63, 58, 92].

1.2.4 Other related works

Stein's method

Stein's method is a useful approach for quantifying distances between probability distributions. First developed in [104] for Gaussian approximation, it has been extended to various distributions, including Poisson [17], binomial [105], and high dimensional settings [96, 23]. We refer to Stein's monograph [105] for a comprehensive review.

Analysis of diffusion models

Since the introduction of diffusion models (DMs) [101, 61, 102], there has been a surge of interest in understanding their theoretical properties. Analysis of localized diffusion models in Chapter 6 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 thesis; we refer to [19, 47] for an in-depth overview.

The convergence of DMs considers error bounds of the sampled distribution given the learned score function. Early work [75] provides a TV guarantee by assuming a log-Sobolev inequality. Later, by using Girsanov theorem, this condition is relaxed to bounded moment conditions [22, 16]. A growing body of work is trying to further relax assumptions and improve error bounds. For instance, [5] proves a linear-in-dimension bound under the KL divergence, [27] uses a relative score approach and derives bounds without early stopping. [90] 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. [84, 117] prove that the diffusion model reaches the minimax rate for distribution estimation. To avoid the CoD, [84, 18] considers linear subspace data, and later [109, 2] extends it to general manifold data. Recently, [119] relaxes the manifold assumption, and improves the ambient dimension dependency in the generalization bound. Other types of low-dimensional structures are also considered. [100] considers certain Gaussian mixtures, and shows that the sample complexity does not depend exponentially on the dimension. [47] further extends it to general Gaussian mixtures with edited diffusion models.

A recent work [80] 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. But their result depends on an explicit solution of the denoising algorithms, and only applies to Ising-type distributions. Our localized diffusion model, however, applies to general high-dimensional graphical models.

1.3 Contributions

The main contributions of this thesis are twofold: (i) theoretical development of the *marginal Stein's method*, which provides a quantitative analysis method for the locality structure and localization method in sampling; (ii) algorithmic development of localized samplers, which reduces a high-dimensional sampling problem to a collection of low-dimensional subproblems; this includes two case studies: an application of the MALA-within-Gibbs sampler adapted to an image deblurring problem, and the design and analysis of localized diffusion models.

For the theoretical aspect, we develop the marginal Stein's method, a novel analysis method for quantifying the effects of locality in high-dimensional distributions. The method provides an approach to translate the structral assumptions of the target distribution into quantitative properties.

Specifically, we define (s, ν)-*local graph* $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ that has a controlled growth rate of neighborhood sizes: for any $i \in \mathsf{V}$, denote $\mathcal{N}_i^r$ as the set of vertices that are within distance r of i, then $|\mathcal{N}_i^r| \leq 1 + \mathsf{s}r^\nu$. We prove that such structural

assumption leads to many interesting properties of Markov random fields that are associated with such local graphs. These include

- *marginal transport inequality* (Theorem 3.3):

$$\max_{i\in[\mathsf{b}]} \mathsf{W}_1(\pi_i, \pi_i') \le c_{\pi'} \cdot \max_{j\in[\mathsf{b}]} \|\nabla_j \log \pi' - \nabla_j \log \pi\|_{L^1(\pi)}, \tag{1.3.1}$$

  which provides a refined control of the 1-Wasserstein distance between marginal distributions π_i, π_i', and here $c_{\pi'}$ is a *dimension-independent* constant;

- *exponential decay of correlations* (Theorem 3.5):

$$|\mathsf{Cov}_\pi (f(x_i), g(x_j))| \lesssim \exp\left(-c_\pi \mathsf{d}_{\mathsf{G}}(i,j)\right), \tag{1.3.2}$$

  which states that the correlations between different components x_i, x_j decay exponentially with their distance $\mathsf{d}_{\mathsf{G}}(i,j)$.

Some generalizations of the above results are discussed for application purposes. We also discuss technical aspects of the method. We interpret the marginal Stein's method from a Langevin semigroup perspective (Theorem 3.8), which presents its own theoretical interest.

For the numerical aspect, we propose a framework to localize existing samplers, that is, to build samplers by combining local samplers for the marginals (see Chapter 4). This effectively reduces a high-dimensional sampling problem to a collection of low-dimensional subproblems. The resulting localized sampler is essentially a Gibbs-type sampler that samples each component X_i conditionally on its neighbors $X_{\mathcal{N}_i^r}$ in the locality structure, i.e.

$$\mathsf{P}^{\mathrm{loc}}(x,y) = \prod_{i\in[\mathsf{b}]} \mathsf{P}_i(x_i, y_i \mid x_{\mathcal{N}_i^r}). \tag{1.3.3}$$

Here P_i are local transition kernels that update the component x_i to y_i conditioned on its r-neighbors $x_{\mathcal{N}_i^r} = \{x_j : \mathsf{d}_{\mathsf{G}}(i,j) \le r\}$ (see (2.1.2)). The validity of such localization is supported by exponential decay of correlations (1.3.2). It ensures that the localization error introduced by ignoring distant components is typically exponentially small in r, which is in general much smaller than the statistical error when learning or sampling with a high-dimensional sampler. The advantanges of the localized sampler include:

- *Localized and parallelable.* Localized implementation reduces the computational cost, and parallelization allows for faster sampling.
- *Reduced statistical complexity.* In many generative tasks, the sampler should be learned or partially learned from data. Learning localized samplers can circumvent the CoD due to its intrinsic low-dimensional nature.
- *Controlled localization error.* The localization error can be controlled using the marginal Stein's method.

We study in detail two concrete examples of localized samplers:

- *MALA-within-Gibbs* (Chapter 5), which localizes the classical Metropolis-adjusted Langevin algorithm (MALA). We apply this method to an image deblurring problem with smooth approximation. We prove that the approximation error is dimension independent by the marginal tranport inequality (1.3.1). We also show how to implement MALA-within-Gibbs in a localized and parallelized manner, which significantly accelerates the sampling process.
- *localized diffusion models* (Chapter 6), which localizes the modern score-based diffusion generative models. We propose to use localized score matching to train the score function in diffusion models within a localized hypothesis space. We prove that such localization enables diffusion models to circumvent CoD, at the price of additional localization error. 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.

1.4 Thesis outline

The structure of the thesis is outlined in Figure 1.1. In Chapter 2, we begin by introducing the notion of **locality structure**. We then develop the **marginal Stein's method** in Chapter 3, which provides a framework to analyze the locality structure and localization method in sampling. Specifically, we derive in Section 3.1 a **marginal transport inequality** and establish in Section 3.2 the **exponential decay of correlations** in localized distributions. Some technical aspects of the marginal Stein's method are then discussed in Sections 3.3 and 3.4. In Chapter 4, we

summarize the general principle of **localization method in sampling**. Then we introduce two concrete examples: the **MALA-within-Gibbs** sampler in Chapter 5, and the **localized diffusion models** in Chapter 6. Finally, we conclude in Chapter 7 with a summary and a discussion of future research directions.

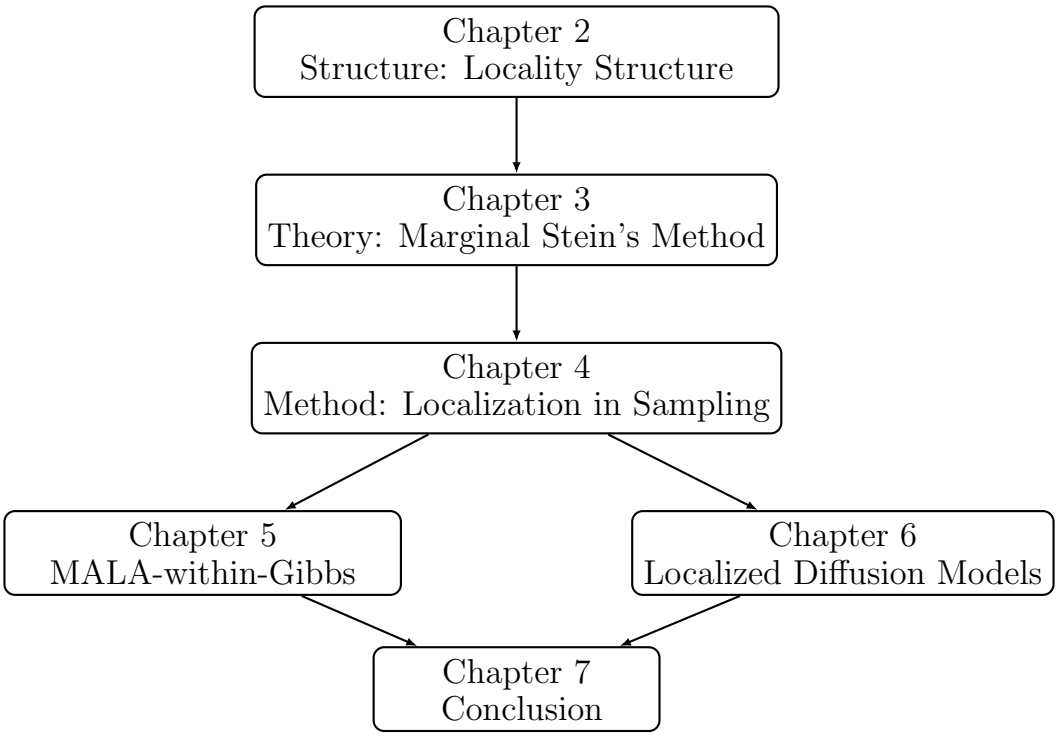

Figure 1.1: Structure of the thesis

1.5 Preliminaries

In this section, we introduce some basic concepts in probability theory and graph theory that are frequently used throughout the thesis. The materials presented here are from standard textbooks, e.g. [42, 3, 12].

1.5.1 Probability theory

A probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the sample space, $\mathcal{F}$ is a σ-algebra on Ω, and $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$. Denote $\mathcal{B}(\mathcal{X})$ as the Borel σ-algebra on a topological space $\mathcal{X}$.

Independence and conditioning

Two random variables X, Y are independent if for any events C, D,

$$\mathbb{P}(X \in C, Y \in D) = \mathbb{P}(X \in C) \cdot \mathbb{P}(Y \in D).$$

Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. The conditional expectation of a random variable X given $\mathcal{G}$ is a random variable $\mathbb{E}[X \mid \mathcal{G}]$ that is $\mathcal{G}$-measurable and satisfies

$$\forall A \in \mathcal{G}, \quad \mathbb{E}\left[\mathbb{E}[X \mid \mathcal{G}] \cdot \mathbf{1}_A\right] = \mathbb{E}[X\mathbf{1}_A].$$

The conditional probability of an event A is defined as $\mathbb{P}(A \mid \mathcal{G}) = \mathbb{E}[\mathbf{1}_A \mid \mathcal{G}]$. The conditional distribution $\pi_{X|\mathcal{G}} = \text{`}\mathsf{Law}\,(X \mid \mathcal{G})\text{'}$ is defined as

$$\pi_{X|\mathcal{G}}(\cdot \mid \mathcal{G}) = \mathbb{P}(X \in \cdot \mid \mathcal{G}).$$

Note it is a $\mathbb{R}$-valued $\mathcal{G}$-measurable random variable. It is called regular if for any $\omega \in \Omega$, $\pi_{X|\mathcal{G}}(\cdot \mid \mathcal{G})(\omega)$ is a probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$. Let Y be a random variable, the conditional expectation, conditional probability, and conditional probability distribution of X given Y are defined similarly by taking $\mathcal{G} = \sigma(Y)$, the σ-algebra generated by Y. We say that two random variables X and Y are conditionally independent given Z if for any events A, B,

$$\mathbb{P}(X \in A, Y \in B \mid Z) = \mathbb{P}(X \in A \mid Z) \cdot \mathbb{P}(Y \in B \mid Z).$$

Markov kernel

A **transition kernel** K from $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ to $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ is a function

$$K : \mathcal{X} \times \mathcal{B}(\mathcal{Y}) \to \mathbb{R}_{\geq 0}.$$

For function $f : \mathcal{Y} \to \mathbb{R}$, measure $\pi \in \mathcal{P}(\mathcal{X})$, define function $Kf : \mathcal{X} \to \mathbb{R}$ and measure $\pi K \in \mathcal{P}(\mathcal{Y})$ as

$$Kf(x) := \int K(x, \mathrm{d}y) f(y), \quad \pi K = \int \pi(\mathrm{d}x) K(x, \cdot).$$

For two kernels $K_1 : \mathcal{X} \times \mathcal{B}(\mathcal{Y}) \to \mathbb{R}$, $K_2 : \mathcal{Y} \times \mathcal{B}(\mathcal{Z}) \to \mathbb{R}$, define kernel $K_1 K_2$ as

$$K_1 K_2(x, \cdot) = K_1 \circ K_2(x, \cdot) = \int K_1(x, \mathrm{d}y) K_2(y, \cdot).$$

A Markov kernel K is a transition kernel such that $K\mathbf{1} = \mathbf{1}$, or equivalently,

$$\forall x \in \mathcal{X}, \quad K(x, \mathcal{Y}) \equiv 1.$$

1.5.2 Graph theory

An **undirected graph** G consists of a set of **vertices** V and **edges** $\mathsf{E} \subseteq \mathsf{V} \times \mathsf{V}$. In this thesis, we allow self-loops, but do not allow multiple edges between the same pair of vertices. For finite graph, denote $\mathsf{b} = \#\mathsf{V}$, and attach each vertex with a unique index $i \in [\mathsf{b}]$. We denote i as the vertex for simplicity. An (induced) subgraph $\mathsf{G}' = (\mathsf{V}', \mathsf{E}')$ of G with vertex set $\mathsf{V}' \subseteq \mathsf{V}$ includes all edges in G that connect two vertices in V', i.e. $\mathsf{E}' = \{(i, j) \in \mathsf{E} : i, j \in \mathsf{V}'\}$.

Cliques

A **complete graph** is a graph in which every pair of distinct vertices is connected by an edge, i.e. $\forall i, j \in \mathsf{V}, i \neq j \Rightarrow (i, j) \in \mathsf{E}$. A subset of vertices $C \subseteq \mathsf{V}$ is called a **clique** if C is a complete graph as a subgraph of G, i.e. every pair of vertices in C is connected by an edge. Denote $\mathcal{C} = \mathcal{C}(\mathsf{G})$ as the collection of all the cliques in G. A **maximal clique** is a clique $C \in \mathcal{C}$ that is not strictly contained in any larger clique, i.e. $\forall A \supsetneq C \Rightarrow A \notin \mathcal{C}$.

Path and graph distance

A **path** in G is a sequence of distinct vertices $(i_0, i_1, \ldots, i_l)$ s.t. $(i_{k-1}, i_k) \in \mathsf{E}$ for $k \in [l]$. We say it is a path from i_0 to i_l (or connecting i_0 and i_l) with length l. The **graph distance** $\mathsf{d}_{\mathsf{G}}(i, j)$ between two vertices $i, j \in \mathsf{V}$ is the length of the shortest path connecting them, i.e.

$$\mathsf{d}_{\mathsf{G}}(i, j) = \inf\{l \in \mathbb{Z}_+ : \exists \text{ path from } i \text{ to } j \text{ with length } l\}. \tag{1.5.1}$$

For simplicity, we let

- $\mathsf{d}_{\mathsf{G}}(i, i) = 0$ for all i.
- $\mathsf{d}_{\mathsf{G}}(i, j) = \infty$ if no path from i to j exists.

The **average path length** is defined as

$$\ell_{\mathsf{G}} = \frac{1}{\mathsf{b}(\mathsf{b}-1)} \sum_{i \neq j} \mathsf{d}_{\mathsf{G}}(i, j). \tag{1.5.2}$$

Note the average path length measures how a graph is connected.

Chapter 2

Locality Structure and Localized Distribution

Locality is an important structure of many physical systems. *The principle of locality* in physics states that an object is influenced directly *only* by its immediate surroundings. In other words, physical interactions are inherently local, and any influence from a distant event must propagate through the intermediate space. Consequently, many spatial or temporal models exhibit the locality structure. A most celebrated example is the *Ising model* in statistical mechanics, where the interaction between spins is limited to nearest neighbors. Understanding how the locality structure affects the properties of the system is an important problem in statistical mechanics and many other fields. In applications, the locality structure provides a promising strategy for designing scalable algorithms that can potentially mitigate the curse of dimensionality. The idea has been studied across various fields, including data assimilation, spatial statistics, image processing, and quantum mechanics. In recent years, there has been a fast growing interest in sampling methods that leverage the locality structure. We aim to provide a clear and quantitative characterization of the locality structure in this chapter.

In this chapter, we first introduce the *Markov random field* to model the locality structure. Next we define the central concept of this thesis, *localized distributions*, as Markov random fields on *localized graphs*. We will discuss the key properties of localized distribution and its relaxations.

2.1 Locality structure

2.1.1 Markov random field

Markov random field (MRF), also known as the *Markov network* or the *undirected graphical model*, is a widely adopted mathematical tool to model the locality structure in distributions. In this section, we introduce the mathematical foundation and important properties of MRF. We will refer to MRF on a *localized graph* as **localized distribution**, a central concept in this thesis that will be used frequently in the following chapters.

In short, a MRF $X = (X_i)_{i\in\mathsf{V}}$ is a collection of random variables defined on an undirected graph $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ and satisfies the *Markov property*

$$X_i \perp\!\!\!\perp X_j \mid X_{\mathsf{V}\setminus\{i,j\}}, \quad \text{if} \quad (i,j) \notin \mathsf{E}.$$

Here $X \perp\!\!\!\perp Y \mid Z$ denotes the conditional independence of X and Y given Z. The Markov property, or the conditional independence, precisely characterizes how the random variables are locally dependent. And the dependence structure is encoded in the *dependency graph* G.

In the following, we will introduce the locality of a graph to quantify the locality structure of MRF. We will also introduce several equivalent characterizations of MRF that are useful both theoretically and practically. For more detailed discussions on MRF, we refer to [66, 118, 69].

2.1.2 Localized graph

We first introduce some basic notations and definitions of graphs, which will be used to characterize the underlying dependency graph of MRF. For a more detailed introduction to graph theory, we refer to Section 1.5.2.

Let $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ be an undirected graph. Let $\mathsf{b} = \#(\mathsf{V})$, and we identify vertices in G with indices $i \in [\mathsf{b}]$. For convenience, we assume E contains all self-loops, i.e.,

$$\forall i \in [\mathsf{b}], \quad (i,i) \in \mathsf{E}.$$

For immediate neighbors, we denote

$$i \sim j \quad \text{iff} \quad (i,j) \in \mathsf{E}.$$

Note $\sim$ is a symmetric and self-reflexive, i.e.

- Symmetric: $i \sim j \Leftrightarrow j \sim i$.
- Self-reflexive: $\forall i \in [\mathsf{b}],\ i \sim i$.

We denote the *neighborhood* of vertex i as

$$\mathcal{N}_i = \{j \in [\mathsf{b}] : i \sim j\}. \tag{2.1.1}$$

To quantify the sparsity of the graph, we define the r-*neighborhood* of i as

$$\mathcal{N}_i^r = \{j \in [\mathsf{b}] : \mathsf{d}_\mathsf{G}(i,j) \leq r\}, \tag{2.1.2}$$

where $\mathsf{d}_\mathsf{G}(i,j)$ is the *graph distance* (1.5.1) between i and j, i.e. the minimum number of edges that must be traversed to go from i to j.

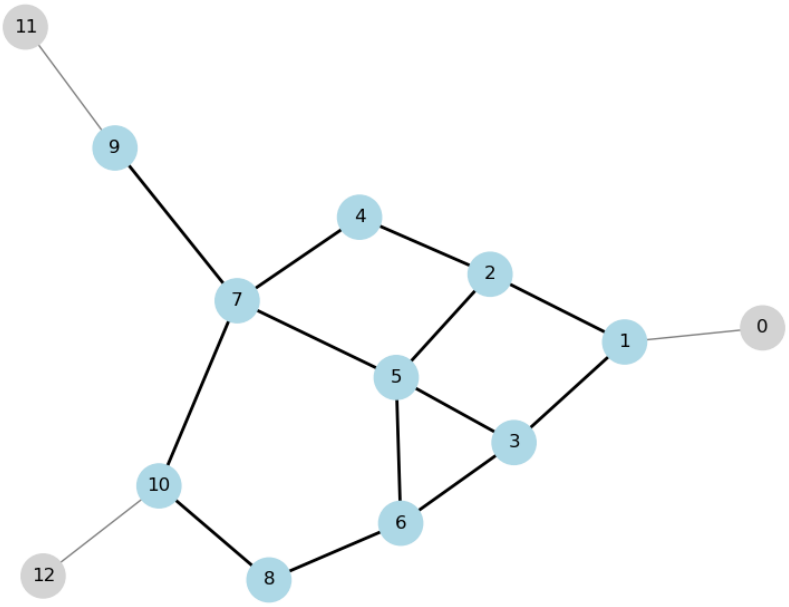

Figure 2.1: Blue vertices: 2-neighborhood of vertex 5 ($\mathcal{N}_5^2$)

For a graph that represents the locality structure, the growth of the neighborhood volume $\#(\mathcal{N}_i^r)$ with the radius r should not grow too fast. In many spatial models, it grows at most polynomially, with an order determined by the dimension of the ambient space in which the graph is embedded. This motivates the following definition of *localized graph*.

Definition 2.1 (Localized graph)**.** An undirected graph G is called $(\mathsf{s}, \boldsymbol{\nu})$**-local** for some $\mathsf{s}, \nu \in \mathbb{Z}_+$, if it satisfies:

$$\forall i \in [\mathsf{b}],\ r \in \mathbb{Z}_+, \quad |\mathcal{N}_i^r| \leq 1 + \mathsf{s} r^\nu. \tag{2.1.3}$$

In the above definition, s denotes the size of the immediate neighbors, and ν is the *ambient dimension* of graph, which controls the growth rate of the neighborhood volume with the radius. An important quantitative feature of localized graph is that s and ν are $\mathcal{O}(1)$ constants compared to the problem dimension, which ensures effective locality of the graph. The polynomial growth of the neighborhood volume with the radius r is a key aspect of this locality.

Localized graph arises naturally in discretization of spatial models. A typical example is the mesh grid in numerical PDEs. Due to the locality of the differential operators, most PDEs are local, and their spatial discretization leads to a localized graph. This is explicitly represented by the sparsity of the discretized difference operators in finite difference methods, or the sparsity of the stiffness matrix in finite element methods.

Example 2.1. A motivating example for Definition 2.1 is the lattice model $\mathbb{Z}^\nu$, where the neighborhood of a vertex $i \in \mathbb{Z}^\nu$ is defined as

$$\mathcal{N}_i = \{j \in \mathbb{Z}^\nu : \|i - j\|_1 \leq 1\}.$$

In this model, a naive bound of the r-neighborhood volume is

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

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

Remark 2.1. For graphs, locality is a stronger condition than the sparsity. Sparsity of a graph only requires that $|\mathsf{E}| = \mathcal{O}(\mathsf{b})$, or the average degree $= \mathcal{O}(1)$. But locality requires more than that. The slow growth of the neighborhood volume in localized graph results in a large average path length (1.5.2). This excludes some sparse graphs, including the small-world network [116], where short-cuts are allowed. For graphs with short-cuts, the neighborhood volume typically grows exponentially with the radius.

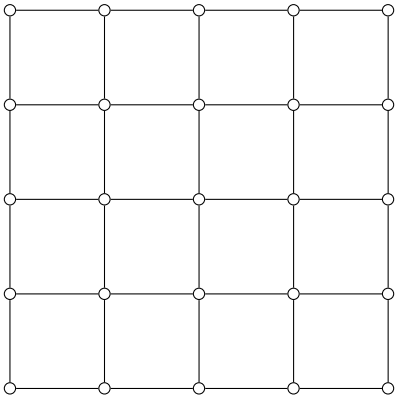

Figure 2.2: Two-dimensional lattice model

2.1.3 Markov property

We proceed to define the MRF on a localized graph G by the Markov property.

Consider an undirected graph $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ with b vertices. Attach each vertex $i \in [\mathsf{b}]$ with a measurable space $(\mathcal{X}_i, \mathcal{F}_i)$. Denote their product space as $(\mathcal{X}, \mathcal{F}) = (\bigotimes_{i\in[\mathsf{b}]} \mathcal{X}_i, \bigotimes_{i\in[\mathsf{b}]} \mathcal{F}_i)$.

Definition 2.2. A MRF on G is a collection of random variables $X = (X_i)_{i\in[\mathsf{b}]}$ in the space $(\mathcal{X}, \mathcal{F})$ such that

$$\forall i \in [\mathsf{b}], \quad X_i \perp\!\!\!\perp X_{[\mathsf{b}]\setminus\mathcal{N}_i} \mid X_{\mathcal{N}_i\setminus\{i\}}, \tag{2.1.4}$$

where $\mathcal{N}_i$ is the neighborhood of i (2.1.1).

(2.1.4) states that the random variable X_i is conditionally independent of all other random variables given its immediate neighbors. If (2.1.4) holds, we call G a *dependency graph* of the MRF X. Note by definition, we do not require the dependency graph to be minimal, i.e. it might include redundant edges.

In the following, we focus on Euclidean spaces, i.e. $(\mathcal{X}_i, \mathcal{F}_i) = (\mathbb{R}^{d_i}, \mathcal{B}(\mathbb{R}^{d_i}))$, where $\mathcal{B}(\mathbb{R}^{d_i})$ is the usual Borel σ-algebra on $\mathbb{R}^{d_i}$. Denote

$$x = (x_1, \ldots, x_{\mathsf{b}}) \in \mathbb{R}^d, \quad \text{where } x_i \in \mathbb{R}^{d_i}, \quad d = \sum_{i=1}^{\mathsf{b}} d_i. \tag{2.1.5}$$

Denote $\pi = \mathsf{Law}(X)$ as the probability distribution of the MRF X on $\mathbb{R}^d$. Unless mentioned otherwise, we always assume that π is absolutely continuous with respect to the Lebesgue measure λ on $\mathbb{R}^d$. Without abuse of notations, still denote its

density as $\pi(x)$, i.e.

$$\pi(x) = \frac{\mathrm{d}\pi}{\mathrm{d}\lambda}(x) = \frac{1}{Z}\exp\left(-U(x)\right), \tag{2.1.6}$$

where $U(x)$ is the *potential function*, and $Z = \int_{\mathbb{R}^d} \exp\left(-U(x)\right) \mathrm{d}x$ is the normalizing constant, or the *partition function*. Distribution of the form (2.1.6) is usually called *Gibbs distribution*.

Notice the conditional independence (2.1.4) can be written as

$$\pi(x_i, x_{[\mathsf{b}]\setminus\mathcal{N}_i} | x_{\mathcal{N}_i\setminus\{i\}}) = \pi(x_i \mid x_{\mathcal{N}_i\setminus\{i\}}) \cdot \pi(x_{[\mathsf{b}]\setminus\mathcal{N}_i} \mid x_{\mathcal{N}_i\setminus\{i\}}). \tag{2.1.7}$$

Example 2.2 (1D Ginzburg-Landau)**.** The Ginzburg-Landau (GL) model is a widely used model in statistical physics [72]. A discrete 1D GL model describes a chain of real-valued spins $\{x_j\}_{j=1}^n$, where each $x_j \in \mathbb{R}$ interacts locally with its neighbors. Its Gibbs distribution is given by

$$\pi(x) = \frac{1}{Z}\exp\left(\sum_{j=1}^{n} V(x_j) + \sum_{j=1}^{n-1} W(x_j, x_{j+1})\right),$$

where $V(x) = \frac{\lambda}{4}(x^2 - m^2)^2$ is the double-well potential and $W(x, y) = \frac{\beta}{2}(x - y)^2$ is the nearest-neighbor interaction. Fix $x_{j\pm1}$, we can factorize the distribution as

$$\begin{aligned}\pi(x_j, x_{[n]\setminus\mathcal{N}_j} \mid x_{j-1}, x_{j+1}) \propto{}& \exp\left(V(x_j) + W(x_{j-1}, x_j) + W(x_j, x_{j+1})\right) \\ &\cdot \exp\left(\sum_{i:i\neq j} V(x_i) + \sum_{i\notin\{j,j-1\}} W(x_i, x_{i+1})\right),\end{aligned}$$

from which we can directly verify the conditional independence.

2.1.4 Equivalent characterizations

Besides the conditional independence (2.1.4), the Markov property in MRF can be characterized in several equivalent ways.

Theorem 2.1. *Let* G *be an undirected graph. Suppose a probability measure* π *has nonnegative density* $\pi(x) \in C^2(\mathbb{R}^d)$*. The following statements are equivalent:*

(1) $X \sim \pi$ *is a MRF on* G*, i.e. (2.1.4) holds.*

(2) $\forall i, j \in [\mathsf{b}],\ i \not\sim j \Rightarrow \nabla^2_{ij} \log \pi(x) = 0.$

(3) $\log \pi(x)$ *admits a* ***clique factorization****, i.e.* $\exists \{u_C\}_{C\in\mathcal{C}}$ *s.t.*

$$-\log \pi(x) = \sum_{C\in\mathcal{C}} u_C(x_C), \tag{2.1.8}$$

where $\mathcal{C}$ *is a collection of cliques in* G*.*

Before stating the proof, we make some remarks on the above theorem. The second condition requires the Hessian of $\log \pi(x)$ to vanish in blocks corresponding to non-adjacent vertices, which is known to be equivalent to their conditional independence (see Lemma 2 in [103]). The sparse Hessian condition is also the key motivation for the localization method in sampling, as it introduces great sparse dependencies in the *score function* $s(x) := \nabla \log \pi(x)$. More discussions on its implication and applications will be given in Chapter 4.

The equivalence between the Markov property and the existence of a clique factorization is the renowned *Hammersley-Clifford theorem* [24]. Note the clique factorization is not unique.

Proof of Theorem 2.1. (1) $\Rightarrow$ (2). The conditional independence implies

$$\log \pi(x_i, x_{[\mathsf{b}]\setminus\mathcal{N}_i} | x_{\mathcal{N}_i\setminus\{i\}}) = \log \pi(x_i \mid x_{\mathcal{N}_i\setminus\{i\}}) + \log \pi(x_{[\mathsf{b}]\setminus\mathcal{N}_i} \mid x_{\mathcal{N}_i\setminus\{i\}}).$$

So that

$$\log \pi(x) = \log \pi(x_{\mathcal{N}_i}) + \log \pi(x_{[\mathsf{b}]\setminus\{i\}}) - \log \pi(x_{\mathcal{N}_i\setminus\{i\}}).$$

For any $j \not\sim i$, we have

$$\nabla^2_{ij} \log \pi(x) = \nabla_i \nabla_j \log \pi(x_{\mathcal{N}_i}) + \nabla_j \nabla_i \log \pi(x_{[\mathsf{b}]\setminus\{i\}}) - \nabla_i \nabla_j \log \pi(x_{\mathcal{N}_i\setminus\{i\}}) = 0.$$

(2) $\Rightarrow$ (3). We prove by induction on the number of vertices b, and note that $\log \pi(x)$ can be replaced by arbitrary function. The case $\mathsf{b} = 1$ is trivial. Assume it holds for $\mathsf{b} - 1$. For b vertices, the result is trivial if G is a complete graph. Otherwise, there exists a vertex k s.t. $|\mathcal{N}_k| < \mathsf{b}$. By (2),

$$\nabla^2_{k,[\mathsf{b}]\setminus\mathcal{N}_k} \log \pi(x) = 0 \;\Rightarrow\; \nabla_k \log \pi(x) = f(x_{\mathcal{N}_k}).$$

If $\pi \in C^3$, the function $f(x_{\mathcal{N}_k}) = \nabla_k \log \pi(x)$ also satisfies condition (2):

$$\forall i, j \in \mathcal{N}_k,\ i \not\sim j \;\Rightarrow\; \nabla^2_{ij} f(x_{\mathcal{N}_k}) = \nabla_k(\nabla^2_{ij} \log \pi(x)) = 0.$$

Since $|\mathcal{N}_k| < \mathsf{b}$, by the induction hypothesis, there exists a clique factorization

$$f(x_{\mathcal{N}_k}) = \sum_{C \in \mathcal{C}_k} u_C(x_C),$$

where $\mathcal{C}_k = \{C \cap \mathcal{N}_k : C \in \mathcal{C}\}$ for some clique collection $\mathcal{C}$. Note

$$\mathcal{C}_k^+ := \{C \cup \{k\} : C \in \mathcal{C}_k\}$$

is also a clique collection in G (since $j \sim k$ for all $j \in \mathcal{N}_k$). So that

$$\nabla_k \log \pi(x) = \sum_{C \in \mathcal{C}_k} u_C(x_C) \;\Rightarrow\; \log \pi(x) = \sum_{C^+ \in \mathcal{C}_k^+} u_{C^+}^+(x_{C^+}) + g(x_{-k}). \tag{2.1.9}$$

Here $u_{C^+}^+$ is any antiderivative of u_C w.r.t. x_k. Now $\forall i, j \in [\mathsf{b}], i \notin \mathcal{N}_j$,

$$0 = \nabla_{ij}^2 \log \pi(x) = \sum_{C^+ \in \mathcal{C}_k^+} \nabla_{ij}^2 u_{C^+}^+(x_{C^+}) + \nabla_{ij}^2 g(x_{-k}) = \nabla_{ij}^2 g(x_{-k}).$$

So that g also satisfies condition (2). By the induction hypothesis, g also admits a clique factorization. So that (2.1.9) provides a clique factorization of $\log \pi(x)$. When $\pi \notin C^3$, one can replace $\nabla_k \log \pi(x)$ by finite difference $\delta_{h_k} \log \pi(x)$, or use smooth mollifier; and the result still holds. This completes the induction.

(3) $\Rightarrow$ (1). For any i, note (2.1.8) implies

$$-\log \pi(x) = \sum_{C \in \mathcal{C}} u_C(x_C) = \sum_{C \in \mathcal{C}, i \in C} u_C(x_C) + \sum_{C \in \mathcal{C}, i \notin C} u_C(x_C).$$

The first term is only a function of $x_{\mathcal{N}_i}$, since $i \in C \in \mathcal{C} \Rightarrow C \subseteq \mathcal{N}_i$. One can write

$$\log \pi(x) = \log f(x_{\mathcal{N}_i}) + \log g(x_{[\mathsf{b}] \setminus \{i\}}).$$

$$\Rightarrow\; \pi(x_i | x_{\mathcal{N}_i \setminus \{i\}}) \propto f(x_{\mathcal{N}_i}), \quad \pi(x_{[\mathsf{b}] \setminus \{i\}} | x_{\mathcal{N}_i \setminus \{i\}}) \propto g(x_{[\mathsf{b}] \setminus \{i\}}).$$

So that (2.1.7) holds. This completes the proof. □

2.2 Localized distribution

In this section, we first define the localized distribution and study its important properties, among which two important properties, i.e. *dimension-independent*

marginal approximation and *exponential correlation decay*, will be studied in detail in the following chapter. Then we introduce its relaxations, the approximate locality and δ-locality.

Definition 2.3. *Localized distribution* is a MRF on a localized graph G.

Besides the Markov property in MRF, the quantitative locality structure in localized graph provides additional information in the localized distribution. For instance, the slow growth of the neighborhood volume (2.1.3) implies that the interaction between random variables at large distance must propogate through a long path. This is the key intuition for the dimension-independent marginal approximation and the exponential correlation decay.

2.2.1 Important properties

Reconstruction from marginals

The Markov property of localized distribution implies that the Hessian of its log density is very sparse. Intuitively, this implies estimation of its density is much easier than the general case. One way to characterize it is that the localized distribution can be reconstructed from its low-dimensional marginals. Note in general, one cannot uniquely reconstruct a distribution from its marginals, see [113].

Theorem 2.2. *A localized distribution π can be reconstructed from its neighborhood marginals $\{\pi_{\mathcal{N}_i}\}_{i\in[\mathsf{b}]}$.*

The above theorem only uses the Markov property, so that it holds for any MRF. But for localized distribution, the neighborhood marginals are guaranteed to be low-dimensional. It suggests that learning a localized distribution is essentially a low-dimensional problem. We will discuss this in detail in Chapter 4.

Proof of Theorem 2.2. By the Markov property, $\pi(x_i|x_{-i}) = \pi(x_i|x_{\mathcal{N}_i\setminus\{i\}})$. Here we denote $x_{-i} := x_{[\mathsf{b}]\setminus\{i\}}$. So that the conditionals $\{\pi(x_i|x_{-i})\}_{i\in[\mathsf{b}]}$ can be obtained from marginals $\{\pi_{\mathcal{N}_i}\}_{i\in[\mathsf{b}]}$. The above theorem then directly follows from Lemma 2.1. □

Lemma 2.1. *Let $u : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function. Then up to a constant, $u \mapsto \{\nabla_i u\}_{i\in[\mathsf{b}]}$ is one-to-one. As a corollary, $\pi \mapsto \{\pi(x_i|x_{-i})\}_{i\in[\mathsf{b}]}$ is one-to-one.*

Proof. We prove by induction on b. When $\mathsf{b} = 1$, $u(x) = \int_{[0,x]} \nabla_1 u(y)\mathrm{d}y + \text{const}$. For $\mathsf{b} \geq 2$, Let v be any fixed antiderivative of $\nabla_1 u$ w.r.t. x_1, i.e. $\nabla_1 v = \nabla_1 u$, then

$$v(x) - u(x) = u_1(x_2, x_3, \ldots, x_d).$$

u_1 is a function in $\mathbb{R}^{d-d_1}$, so that by induction hypothesis, it is uniquely determined (up to a constant) by

$$\{\nabla_j u_1 = \nabla_j v - \nabla_j u\}_{j=2,\ldots,d}.$$

Since v is fixed, $\{\nabla_i u\}_{i\in[\mathsf{b}]}$ uniquely determines u_1, and thus u (up to a constant). This completes the induction.

For the corollary, suppose first π has C^1 density $\pi(x)$. Consider the function $\log \pi(x)$. Notice

$$\nabla_i \log \pi(x) = \nabla_i \log \pi(x_i|x_{-i}).$$

So that up to a constant $\{\log \pi(x_i|x_{-i})\}_{i\in[\mathsf{b}]}$ uniquely determines $\log \pi(x)$. But the constant can be fixed by the normalization condition

$$\int \pi(x)\mathrm{d}x = 1.$$

When $\pi \notin C^1$, one can replace $\nabla_i \log \pi(x)$ by finite difference $\delta_{h_i} \log \pi(x)$, or use smooth mollifier. This completes the proof. □

Dimension-independent marginal approximation

Theorem 2.2 indicates that to estimate a localized distribution, it suffices to estimate its low-dimensional marginals. Due to the local dependencies, the error of approximating low-dimensional marginals is usually dimensional independent. It is natural to consider the marginal version of existing distribution inequalities, such as the renown Otto-Villani inequality [86]

$$\mathsf{W}_2(\mu, \nu) \leq C_\mu \sqrt{\mathsf{I}(\mu\|\nu)}, \tag{2.2.1}$$

where W_2 is the 2-Wasserstein distance, and $\mathsf{I}(\mu\|\nu) = \mathbb{E}_\mu \left\|\nabla \log \frac{\mu}{\nu}\right\|^2$ denotes the Fisher information. Our target is to establish a marginal version of the above inequality, which should be dimension-independent.

We will discuss such marginal inequalities in detail in Section 3.1. The following theorem is from [33], where a marginal Otto-Villani inequality is established using

the δ-locality condition introduced in Section 3.1.3.

Theorem 2.3. *Consider two distributions $\pi, \pi' \in \mathcal{P}_1(\mathbb{R}^d)$. Assume π' is δ-localized (see Definition 3.1), then the marginal W_1 distance of π, π' satisfies*

$$\max_{i\in[\mathsf{b}]} \mathsf{W}_1(\pi_i, \pi_i') \le \delta \cdot \max_{j\in[\mathsf{b}]} \|\nabla_j \log \pi' - \nabla_j \log \pi\|_{L^1(\pi)}\,, \tag{2.2.2}$$

where π_i and π_i' denote the marginals of π and π' on x_i respectively.

The above inequality provides a dimension independent uniform control of the marginal errors in terms of the difference in the score's individual components. This is not achieved if one simply uses the joint distribution error bounds, since those are usually dimension dependent as in (2.2.1). In high-dimensional problems, dimension-dependent bounds are usually meaningless for marginal error control.

We would like to comment on the importance of the marginal inequalities. In high-dimensional problems, usually not all the components are of interest, and one usually only needs the statistics of a few components [44, 64, 111, 46]. For instance, in image deblurring problems [46], the uniform marginal bound ensures that the error is evenly distributed across the image, rather than concentrating on certain pixels and creating unwanted artifacts in the image.

Exponential correlation decay

One of the most important properties of localized distribution is the *exponential correlation decay*, which states that the correlation between two random variables decays exponentially with their graph distance. Detailed discussions on the exponential correlation decay will be given in Section 3.2. Here we state a key result (Theorem 3.5).

Theorem 2.4. *Suppose π has dependency graph G and is log-concave and smooth, i.e. $\exists 0 < m \le M < \infty$ s.t. $mI \preceq -\nabla^2 \log \pi(x) \preceq MI$. 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*

$$|\mathsf{Cov}_{x\sim\pi}(f(x_i), g(x_j))| \le \frac{1}{m}\left(1 - \frac{m}{M}\right)^{\mathsf{d}_{\mathsf{G}}(i,j)} |f|_{\mathrm{Lip}}\, |g|_{\mathrm{Lip}}\,. \tag{2.2.3}$$

We comment the above theorem *does not* assumes locality of the graph G. The exponential correlation decay is merely a consequence of the Markov property.

2.2.2 Approximate locality

In many applications, the locality structure is not exact, but only approximately holds. It is therefore important to consider relaxations of the locality condition.

The first natural relaxation is based on the sparse Hessian condition (see (2) in Theorem 2.1). Motivated by the ubiquitous exponential decay phenomenon, we introduce in [56] the following *approximate locality* condition.

Definition 2.4. A distribution π is called approximately localized w.r.t. G, if there exist dimensional independent constants $c_\pi, C_\pi > 0$ such that

$$\left\|\nabla^2_{ij} \log \pi\right\|_\infty \leq C_\pi \exp\left(-c_\pi \mathsf{d}_\mathsf{G}(i,j)\right). \tag{2.2.4}$$

Here $\|\cdot\|_\infty$ denotes the L^∞-norm.

Here, by dimensional independence we mean c_π, C_π do not scale with the problem dimension for a certain class of target distributions.

Note the score function of approximate localized distributions can be efficiently approximated by a low-dimensional function

$$s_j(x) = \nabla_j \log \pi(x) \approx \widehat{s}_j(x_{\mathcal{N}_j^r}).$$

Here r is the *localization radius* and $\mathcal{N}_j^r$ is the r-neighborhood (2.1.2). Note by (2.2.4), the approximation error decays exponentially with the radius r, while the dimension of $\widehat{s}_j$ only grows polynomially with r if G is localized.

Another relaxation is the δ-*locality* condition [33], which is inspired by the Stein's method. It provides a *quantitative* characterization of the locality structure. It is a more general condition, but is difficult to directly verify in practice. We will discuss it in detail in Section 3.1.3.

Finally, we mention a possible relaxation of the localized distribution, which is based on the exponential correlation decay (see Theorem 3.5). That is, we require that for any i, j and 1-Lipschitz functions $f : \mathbb{R}^{d_i} \to \mathbb{R}$ and $g : \mathbb{R}^{d_j} \to \mathbb{R}$, it holds

$$\left|\mathsf{Cov}_{x\sim\pi}\left(f(x_i), g(x_j)\right)\right| \leq C_\pi \exp\left(-c_\pi \mathsf{d}_\mathsf{G}(i,j)\right).$$

Here C_π, c_π are dimensional independent constants. Such relaxation captures an important feature of locality structure, is easier to verify in practice, and also applies to a wider range of distributions, including empirical distributions.

Chapter 3

Marginal Stein's Method

In this chapter, we introduce a novel *marginal Stein's method* that relates the locality structure to quantitative properties of high dimensional distributions. This method originates from the classical Stein's method [104], a powerful tool for quantifying distances between probability distributions. It considers a *Stein equation* associated with a test function, and provides a way to bound the test error via the solution of the equation. To apply Stein's method to derive bounds on marginal distributions, we introduce the *marginal Stein equation*, where the test function only depends on certain marginal variables. By a careful gradient estimate of the marginal Stein equation, we derive a marginal transport inequality in Section 3.1 that provides dimension independent bounds on the marginal distance. Some generalizations of this marginal inequality are also discussed. This method can go beyond bounds on marginal distributions, and it can be used to derive bounds on certain integrals against localized distributions. In Section 3.2, we establish the exponential decay of correlation between different components of localized distributions using the marginal Stein's method. Section 3.3 introduces the key technical analysis in Marginal Stein's method, i.e. the gradient estimate of the marginal Stein equation, which crucially quantifies the locality structure. Section 3.4 interprets the marginal Stein's method from a Langevin semigroup perspective, which presents its own theoretical interest.

3.1 Marginal transport inequality

3.1.1 Stein's method

Stein's method is a useful approach for quantifying distances between probability distributions. First developed in [104] for Gaussian approximation, it has been extended to various distributions, including Poisson [17], binomial [105], diffusion process [4], and high dimensional settings [96, 23]. We refer to Stein's monograph [105] for a comprehensive review. Here we focus on Stein's method for general continuous distributions.

Consider two continuous distributions $\pi, \pi' \in \mathcal{P}_1(\mathbb{R}^d)$. Depending on the choice of the probability distances $\mathsf{d}(\pi, \pi')$, we take test functions ϕ from certain function class $\mathcal{F}$. For instance,

- $\mathcal{F} = \{f : |f|_{\mathrm{Lip}} \le 1\}$, then $\mathsf{d}(\pi, \pi')$ is the 1-Wasserstein distance $\mathsf{W}_1(\pi, \pi')$.
- $\mathcal{F} = \{f : \|f\|_\infty \le \frac{1}{2}\}$, then $\mathsf{d}(\pi, \pi')$ is the TV distance $\mathsf{TV}(\pi, \pi')$.
- $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \le 1\}$, where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), then $\mathsf{d}(\pi, \pi')$ is the maximum mean discrepancy $\mathsf{MMD}(\pi, \pi')$.

Fix a test function $\phi \in \mathcal{F}$, consider the *Stein equation*

$$\mathcal{L}_\pi u_\phi := \Delta u_\phi + \nabla \log \pi \cdot \nabla u_\phi = \phi - \mathbb{E}_\pi[\phi]. \tag{3.1.1}$$

Here $\mathcal{L}_\pi$ is called the *Stein operator*, which in this case is exactly the generator of the Langevin dynamics [83, 3] associated with π. Suppose that for certain π and function class $\mathcal{F}$, one can derive the gradient estimate

$$\sup_{\phi \in \mathcal{F}} \|\nabla u_\phi\|_\infty \le C_\pi.$$

Here $\|\cdot\|_\infty$ denotes the L^∞ norm. Then it holds that

$$\begin{aligned}
\sup_{\phi \in \mathcal{F}} [\mathbb{E}_{\pi'}[\phi] - \mathbb{E}_\pi[\phi]] &= \sup_{\phi \in \mathcal{F}} \mathbb{E}_{\pi'}[\phi - \mathbb{E}_\pi[\phi]] = \sup_{\phi \in \mathcal{F}} \mathbb{E}_{\pi'}[\mathcal{L}_\pi u_\phi] \\
&= \sup_{\phi \in \mathcal{F}} \mathbb{E}_{\pi'}[(\nabla \log \pi - \nabla \log \pi') \cdot \nabla u_\phi] \\
&\le \sup_{\phi \in \mathcal{F}} \|\nabla \log \pi - \nabla \log \pi'\|_{L^1(\pi')} \|\nabla u_\phi\|_\infty \\
&\le C_\pi \|\nabla \log \pi - \nabla \log \pi'\|_{L^1(\pi')}.
\end{aligned}$$

Here the second line follows from the integration by parts (or *Stein's lemma* [4])

$$\begin{aligned}\mathbb{E}_{\pi'}[\mathcal{L}_\pi u_\phi] &= \int (\Delta u_\phi(x) + \nabla \log \pi(x) \cdot \nabla u_\phi(x))\, \pi'(x) \mathrm{d}x \\ &= \int (-\nabla u_\phi(x) \cdot \nabla \pi'(x) + \nabla \log \pi(x) \cdot \nabla u_\phi(x)\ \pi'(x))\, \mathrm{d}x \qquad (3.1.2) \\ &= \int (\nabla \log \pi(x) - \nabla \log \pi'(x)) \cdot \nabla u_\phi(x)\ \pi'(x) \mathrm{d}x.\end{aligned}$$

We can see from above that the key step is to derive the gradient estimate of Stein equation for specific distribution π and function class $\mathcal{F}$. The result then follows from the standard argument of Stein's method.

3.1.2 Marginal Stein equation

To derive bounds on the marginal distance, consider test function ϕ that only depends on the i-th component x_i, i.e.

$$\phi(x) = \phi_i(x_i), \quad \phi_i \in \mathcal{F}_i.$$

Notice the obvious relation

$$\mathbb{E}_{\pi_i'}[\phi_i] - \mathbb{E}_{\pi_i}[\phi_i] = \mathbb{E}_{\pi'}[\phi] - \mathbb{E}_{\pi}[\phi].$$

The right hand side can be controlled using the Stein's method. Then it suffices to derive the gradient estimate of the *marginal Stein equation*

$$\mathcal{L}_\pi u_\phi(x) := \Delta u_\phi(x) + \nabla \log \pi(x) \cdot \nabla u_\phi(x) = \phi_i(x_i) - \mathbb{E}_\pi[\phi_i(x_i)]. \qquad (3.1.3)$$

Note the left hand side is generally a function of x, but the right hand side is only a function of x_i. Well-posedness of (3.1.3) requires certain conditions. In this thesis, we focus the following scenario:

- π satisfies the *Poincaré inequality* [3], i.e. there exists a constant $\mathsf{C}_{\mathsf{PI}} > 0$ s.t.

$$\forall u \in H^1(\pi), \quad \mathsf{Cov}_\pi(u) \le \mathsf{C}_{\mathsf{PI}} \mathbb{E}_\pi[\|\nabla u\|^2]. \qquad (3.1.4)$$

- ϕ_i is a Lipschitz function, i.e., $|\phi_i|_{\mathrm{Lip}} < \infty$.

Under these conditions, we can show by Lax-Milgram theorem that the marginal Stein equation (3.1.3) has a unique solution u_ϕ in the space

$$H_0^1(\pi) := \{u \in H^1(\pi) : \mathbb{E}_\pi[u] = 0\}. \tag{3.1.5}$$

Under this setting, we can derive the W_1 bounds on the marginal distance. Extension to other distances is left for future work.

3.1.3 δ-localized distributions

Deriving the gradient estimate of the marginal Stein equation (3.1.3) is non-trivial. For generality, we propose in [33] to directly use the gradient estimate condition to identify a class of distributions that satisfy the marginal transport inequality.

Definition 3.1. A distribution $\pi \in \mathcal{P}_1(\mathbb{R}^d)$ is called **δ-localized** for some constant $\delta > 0$, *independent of* d, if for any $i \in [\mathsf{b}]$ and 1-Lipschitz function $\phi_i : \mathbb{R}^{d_i} \to \mathbb{R}$, the solution $u(x)$ to the marginal Stein equation

$$\Delta u(x) + \nabla \log \pi(x) \cdot \nabla u(x) = \phi_i(x_i) - \mathbb{E}_\pi[\phi_i(x_i)],$$

satisfies the gradient estimate

$$\|\nabla u\|_{\infty,1} := \sum_{j=1}^{\mathsf{b}} \|\nabla_j u\|_{L^\infty} \le \delta. \tag{3.1.6}$$

We will show in Theorem 3.3 that the δ-locality condition directly implies the marginal transport inequality. From the proof of Theorem 3.3, we can see that

$\nabla_j u(x)$ *quantifies* how modifications of $\nabla_j \log \pi(x)$ affect the marginal π_i.

We keep the technical definition here for the sake of generality. In the following, we consider two classes of distributions, i.e. the localized distributions and the distributions with certain diagonal dominance condition, and claim that they are δ-localized with δ independent of d.

Localized distributions are δ-localized

Theorem 3.1. *Let* G *be a* (s, ν)*-local graph. Suppose* $\pi \in \mathcal{P}(\mathbb{R}^d)$ *is localized on* G*, and satisfies for some* $0 < m \le M < \infty$*,*

$$\forall x \in \mathbb{R}^d, \quad mI \preceq -\nabla^2 \log \pi(x) \preceq MI.$$

Then π is δ-localized with $\delta = \frac{\mathsf{s}\nu!\kappa^\nu}{m}$ where $\kappa = \frac{M}{m}$.

Proofs are delayed to Section 3.3.3.

Remark 3.1. The condition number κ plays an important role in localization:

- κ is known to be crucial to preserve the band structure in matrix inversion [38, 9]. In probability language, consider a Gaussian distribution $\mathsf{N}(0, C)$, then a moderate κ ensures the equivalence of local correlation (C is nearly banded) and conditional dependencies (C^{-1} is nearly banded).

- We comment that the condition number κ of typical localized distributions is independent of dimension d. This is in contrast to the distributions for fixed-domain models with finer resolution. The key difference is different types of high-dimensionality. An illustrative example is the 1d lattice model:

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

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

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

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

   The condition number is thus

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

   Examples include using a finer discretization of a PDE problem.

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

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

Examples include spatial extension of a physical system.

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.

Diagonal dominant distributions are δ-localized

Another condition that implies δ-locality is a diagonal dominance condition studied in [46], which is motivated by an image deblurring problem. In this case, the dependency graph of the distribution is not necessarily local. The locality is guaranteed by the diagonal dominance, which can be interpreted as that any block x_i is mostly correlated with itself rather than with other blocks x_j.

Theorem 3.2. *Consider $\pi \in \mathcal{P}_1(\mathbb{R}^d)$. Suppose $H(x) := -\nabla^2 \log \pi(x)$ is c-uniformly diagonal block dominant, i.e. there exists a matrix $M \in \mathbb{R}^{\mathsf{b}\times\mathsf{b}}_{\geq 0}$ s.t. $\forall i, j \in [\mathsf{b}], i \neq j$,*

$$H_{ii}(x) \succeq M_{ii} I_{d_i}, \quad \|H_{ij}(x)\| \leq M_{ij}; \quad \sum_{j:j\neq i} M_{ij} + c \leq M_{ii},$$

where $H_{ij}(x)$ denotes the (i,j)-th subblock of $H(x)$. Then π is c^{-1}-localized.

Proofs are delayed to Section 3.3.4.

Remark 3.2. (1) The diagonal dominance condition takes a similar form as the Dobrushin condition [39], where it assumes that the sum of the influence coefficients

$$\rho_{ij} := \sup_{x_{-i}, y_{-i}: x_{[\mathsf{b}]\setminus\{i,j\}} = y_{[\mathsf{b}]\setminus\{i,j\}}} \mathsf{TV}\left(\pi_{i|-i}(\cdot|x_{-i}), \pi_{i|-i}(\cdot|y_{-i})\right).$$

is bounded by a constant $c < 1$, i.e. $\max_i \sum_{j:j\neq i} \rho_{ij} \leq c$. Although M_{ij} and ρ_{ij} have similar interpretations, i.e. they measure the correlation between x_i and x_j, it is hard to find direct connection between them. The motivation for the two conditions is different: the Dobrushin condition is to ensure the uniqueness of Gibbs measure, while the diagonal dominance condition is to ensure the distribution is effectively localized. We also point out that the Dobrushin condition in general is hard to verify, while the diagonal dominance condition is easier to check in practice.

(2) The conditions above imply that π is log-concave. Denote $M' \in \mathbb{R}^{\mathsf{b}\times\mathsf{b}}$ s.t. $M'_{ij} := 2M_{ii}\mathbf{1}_{i=j} - M_{ij}$, then M' is c-diagonal dominant. Gerŝgorin discs

theorem [62] implies that the smallest eigenvalue of M' is lower bounded by c, and thus

$$\begin{aligned} u^{\mathrm{T}} H(x) u &= \sum_{i,j} u_i^{\mathrm{T}} H_{ij}(x) u_j \geq \sum_i M_{ii} \|u_i\|^2 - \sum_{i \neq j} M_{ij} \|u_i\| \|u_j\| \\ &= \sum_{i,j} M'_{ij} \|u_i\| \|u_j\| \geq c \sum_i \|u_i\|^2 = c \|u\|^2 . \end{aligned}$$

3.1.4 Marginal transport inequality

Theorem 3.3. *Consider two distributions $\pi, \pi' \in \mathcal{P}_1(\mathbb{R}^d)$. Assume π' is δ-localized, then the marginal W_1 distance of π, π' satisfies*

$$\max_{i \in [\mathsf{b}]} \mathsf{W}_1(\pi_i, \pi'_i) \leq \delta \cdot \max_{j \in [\mathsf{b}]} \|\nabla_j \log \pi' - \nabla_j \log \pi\|_{L^1(\pi)} , \tag{3.1.7}$$

where π_i and π'_i denote the marginals of π and π' on x_i respectively.

Proof. By Kantorovich duality [113],

$$\mathsf{W}_1(\pi_i, \pi'_i) = \sup_{\phi_i \in \mathrm{Lip}_1} \left[\mathbb{E}_{\pi_i}[\phi_i] - \mathbb{E}_{\pi'_i}[\phi_i] \right] .$$

Let $u(x)$ solve Stein equation

$$\mathcal{L}_{\pi'} u(x) = \phi_i(x_i) - \mathbb{E}_{\pi'}[\phi_i(x_i)].$$

Since π' is δ-localized, we have

$$\|\nabla u\|_{\infty,1} \leq \delta.$$

Denote $\phi(x) = \phi_i(x_i)$, then by Stein's Lemma (3.1.2), we have

$$\begin{aligned} \mathbb{E}_{\pi_i}[\phi_i] - \mathbb{E}_{\pi'_i}[\phi_i] &= \mathbb{E}_{\pi}[\phi] - \mathbb{E}_{\pi'}[\phi] \\ &= \mathbb{E}_{\pi}[\phi - \mathbb{E}_{\pi'}[\phi]] = \mathbb{E}_{\pi}[\mathcal{L}_{\pi'} u] \\ &= \mathbb{E}_{\pi}[(\nabla \log \pi' - \nabla \log \pi) \cdot \nabla u] \\ &= \sum_{j=1}^{\mathsf{b}} \mathbb{E}_{\pi}[(\nabla_j \log \pi' - \nabla_j \log \pi) \cdot \nabla_j u]. \end{aligned} \tag{3.1.8}$$

Therefore, we obtain

$$\begin{aligned}\max_i \mathsf{W}_1(\pi_i, \pi_i') &\le \max_i \max_{\phi_i \in \mathrm{Lip}_1} \left[\mathbb{E}_{\pi_i}[\phi_i] - \mathbb{E}_{\pi_i'}[\phi_i]\right]. \\ &\le \max_i \max_{\phi_i \in \mathrm{Lip}_1} \sum_{j=1}^{\mathsf{b}} \|\nabla_j u\|_\infty \cdot \max_j \|\nabla_j \log \pi' - \nabla_j \log \pi\|_{L^1(\pi)} \\ &= \delta \cdot \max_j \|\nabla_j \log \pi' - \nabla_j \log \pi\|_{L^1(\pi)}.\end{aligned}$$

This completes the proof. □

Remark 3.3. To control the marginal error, one can directly apply Otto-Villani inequality [86] on the marginals and obtain

$$\mathsf{W}_2(\pi_i, \pi_i') \le C_{\pi_i} \sqrt{\mathbb{E}_{\pi_i}\left[\|\nabla \log \pi_i - \nabla \log \pi_i'\|^2\right]}.$$

The main issue of this approach is that we often only have access to π and π' but not to their marginals. Evaluating the marginals is often computationally very challenging as it involves integrating out the other components. This makes the inequality less useful in practice.

3.1.5 Generalizations of marginal transport inequality

Various generalizations of the marginal transport inequality are possible for applications in different scenarios. They can be similarly derived using the marginal Stein's method. We introduce two examples here.

Marginal error of a specific block

The marginal transport inequality only provides a ℓ_∞-bound over the marginal blocks, which is due to that we use a δ-locality condition that mixes all blocks. For localized distributions, the proof of Theorem 3.1 already reveals the exponential decay of $\|\nabla_j u\|$ in terms of $\mathsf{d}_\mathsf{G}(i,j)$ (cf. (3.3.10)). A direct consequence is the following marginal transport inequality for a specific block.

Proposition 3.1. *Under the conditions in Theorem 3.1, it holds that*

$$\mathsf{W}_1(\pi_i, \pi_i') \le \frac{1}{m} \sum_{j \in [\mathsf{b}]} \left(1 - \frac{m}{M}\right)^{\mathsf{d}_\mathsf{G}(i,j)} \|\nabla_j \log \pi' - \nabla_j \log \pi\|_{L^1(\pi)}.$$

This is a special case of Proposition 3.2. The above proposition says that the approximation error of a certain marginal of localized distributions mainly depends on the error of its neighboring components of the score function. This provides a refined control compared to Theorem 3.3.

Marginal error of multiple blocks

The marginal transport inequality only considers the marginal distance of π and π' on one block x_i. It is natural to extend this result to the case of multiple blocks.

Theorem 3.4. *Consider two distributions $\pi, \pi' \in \mathcal{P}_1(\mathbb{R}^d)$. Assume π' satisfies any one of the conditions in Theorem 3.1 or Theorem 3.2, then for any index set $I \subseteq [\mathsf{b}]$, the W_1 distance of π, π' on the marginal x_I satisfies*

$$\mathsf{W}_1(\pi_I, \pi'_I) \leq \delta |I| \cdot \max_{j \in [\mathsf{b}]} \|\nabla_j \log \pi' - \nabla_j \log \pi\|_{L^1(\pi)} . \tag{3.1.9}$$

Here δ can be taken as the same as in Theorem 3.1 or Theorem 3.2.

Proofs are delayed to Section 3.5.1. Theorem 3.4 provides further control on the correlation between different blocks in π', which cannot be directly derived from Theorem 3.3. When π and π' are both Gaussians, Theorem 3.3 only guarantees that the diagonal blocks of the covariance matrix of π' are close to those of π, while Theorem 3.4 further guarantees that the off-diagonal blocks are also close.

3.2 Exponential correlation decay

In localized systems, interactions between different sites are short ranged, and any influence from a distant site must propagate through the intermediate space. This implies that perturbations at one site affect distant sites only weakly, hence the correlation decays. In this section, we will quantify such decay and show that it is exponential in the distance in the low temperature or weakly coupled regime. Such exponential decay is a ubiquitous phenomenon reported in probability [71, 15], statistical physics [14, 41] and quantum mechanics [7, 53]. We will use the marginal Stein's method to derive the correlation exponential decay, where the gradient of the solution to the marginal Stein equation precisely encodes the correlation structure. We will also discuss a generalized version of the exponential decay result.

3.2.1 Exponential correlation decay

Theorem 3.5. *Suppose π has dependency graph G and is log-concave and smooth, i.e. $\exists 0 < m \leq M < \infty$ s.t. $mI \preceq -\nabla^2 \log \pi(x) \preceq MI$. 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*

$$|\mathsf{Cov}_\pi (f(x_i), g(x_j))| \leq \frac{1}{m} \left(1 - \frac{m}{M}\right)^{\mathsf{d}_\mathsf{G}(i,j)} |f|_{\mathrm{Lip}} |g|_{\mathrm{Lip}} . \tag{3.2.1}$$

Note the above theorem does not assume the sparsity or locality of the graph. It is merely a consequence of the Markov property. The proof is based on the marginal Stein's method, and a key result (Theorem 3.6) on the gradient estimate of the marginal Stein equation, which quantifies the exponential decay of the correlation in localized systems.

Proof of Theorem 3.5. By subtracting the mean, we assume without loss of generality that $\mathbb{E}_\pi[f(x_i)] = \mathbb{E}_\pi[g(x_j)] = 0$. Then

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

Consider the marginal Stein equation

$$\mathcal{L}_\pi u_f(x) = f(x_i).$$

By Theorem 3.6, the following gradient estimate of u_f holds:

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

By integration by parts, it holds that

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

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

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

This completes the proof. □

Theorem 3.6. *Suppose π has dependency graph G and is log-concave and smooth, i.e. $\exists 0 < m \leq M < \infty$ s.t. $mI \preceq -\nabla^2 \log \pi(x) \preceq MI$. For any i and Lipschitz function $f : \mathbb{R}^{d_i} \to \mathbb{R}$, let $u(x)$ solve the marginal Stein equation*

$$\mathcal{L}_\pi u(x) = f(x_i) - \mathbb{E}_\pi[f(x_i)].$$

The following gradient estimate holds:

$$\|\nabla_j u\|_\infty \leq \frac{1}{m} \left(1 - \frac{m}{M}\right)^{\mathsf{d}_\mathsf{G}(i,j)} |f|_{\mathrm{Lip}} \, . \tag{3.2.2}$$

Proof. The proof is based on a refined analysis of that of Theorem 3.1. By Lemma 3.1, the gradient of the solution to the marginal Stein equation is given by

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

where X_t^x is the path solution of the Langevin dynamics

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

Since f is Lipschitz, we obtain

$$\|\nabla_j u(x)\| \leq |f|_{\mathrm{Lip}} \int_0^\infty \mathbb{E}\|\nabla_{x_j} X_{t,i}^x\| \mathrm{d}t.$$

As in the proof of Theorem 3.1, (3.3.10) holds:

$$\|\nabla_{x_j} X_t^x\| \leq \mathrm{e}^{-Mt} \sum_{r=\mathsf{d}_\mathsf{G}(i,j)}^\infty \frac{t^r (M-m)^r}{r!}.$$

Therefore,

$$\begin{aligned}\|\nabla_j u(x)\| &\le |f|_{\mathrm{Lip}} \int_0^\infty \mathbb{E}\|\nabla_{x_j} X_{t,i}^x\| \mathrm{d}t \\ &\le |f|_{\mathrm{Lip}} \int_0^\infty \mathrm{e}^{-Mt} \sum_{r=\mathsf{d}_{\mathsf{G}}(i,j)}^\infty \frac{t^r (M-m)^r}{r!} \mathrm{d}t \\ &= |f|_{\mathrm{Lip}} \frac{1}{M} \sum_{r=\mathsf{d}_{\mathsf{G}}(i,j)}^\infty \Big(1-\frac{m}{M}\Big)^r = \frac{1}{m}\Big(1-\frac{m}{M}\Big)^{\mathsf{d}_{\mathsf{G}}(i,j)} |f|_{\mathrm{Lip}}\,.\end{aligned}$$

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

Recall in the proof of Theorem 3.3, $\|\nabla_j u\|$ controls how much the marginal distribution π_j changes when we perturb the score $\nabla_i \log \pi$. The above theorem directly implies a refined version of Theorem 3.3:

Proposition 3.2. *Consider two distributions $\pi, \pi' \in \mathcal{P}_1(\mathbb{R}^d)$. Suppose π has dependency graph G and is log-concave and smooth, i.e. $\exists 0 < m \le M < \infty$ s.t. $mI \preceq -\nabla^2 \log \pi(x) \preceq MI$. Then it holds that*

$$\mathsf{W}_1(\pi_i, \pi_i') \le \frac{1}{m} \sum_{j \in [\mathsf{b}]} \Big(1-\frac{m}{M}\Big)^{\mathsf{d}_{\mathsf{G}}(i,j)} \|\nabla_j \log \pi' - \nabla_j \log \pi\|_{L^1(\pi)}\,. \tag{3.2.3}$$

Proof. The result directly follows from (3.1.8) and Theorem 3.6. □

3.2.2 Generalization

In Theorem 3.5, the functions f and g are assumed to only depend on x_i and x_j. A direct generalization is possible by allowing f and g to depend on all variables, but with 'concentration' on x_i and x_j. These observables arise in practical problems such as spatial statistics, where they are often expressed as local functionals of the entire field. For these observables, we prove

Theorem 3.7. *Suppose π is localized on a (s, ν)-local graph G, and is log-concave and smooth, i.e. $\exists 0 < m \le M < \infty$ s.t. $mI \preceq -\nabla^2 \log \pi(x) \preceq MI$. Let $i, j \in [\mathsf{b}]$, and suppose $f, g : \mathbb{R}^d \to \mathbb{R}$ satisfy xz*

$$\|\nabla_k f\|_\infty \le L_f \exp(-c_f \mathsf{d}_{\mathsf{G}}(i,k)), \quad \|\nabla_k g\|_\infty \le L_g \exp(-c_g \mathsf{d}_{\mathsf{G}}(j,k)).$$

Then it holds that

$$|\mathsf{Cov}_\pi\left(f(x), g(x)\right)| \leq L_f L_g \psi_{3(\nu+1)}(\mathsf{d}_\mathsf{G}(i,j)) \exp(-c\mathsf{d}_\mathsf{G}(i,j)). \tag{3.2.4}$$

Here $\psi_{3(\nu+1)} \in \mathbb{P}_{3(\nu+1)}$ *is a polynomial depending on* s, ν *and* c*, and* c *is defined as*

$$c = \min\left\{c_f, c_g, -\log\left(1 - \frac{m}{M}\right)\right\},$$

Proof. By subtracting the mean, we assume without loss of generality that $\mathbb{E}_\pi[f] = \mathbb{E}_\pi[g] = 0$. Then

$$\mathsf{Cov}_\pi\left(f(x), g(x)\right) = \int f(x)g(x)\pi(x)\mathrm{d}x.$$

Let u solve Stein equation $\mathcal{L}_\pi u = f$. By Lemma 3.1, the solution is given by

$$u(x) = -\int_0^\infty \mathbb{E}[f(X_t^x)]\mathrm{d}t,$$

By (3.3.10) and the assumption on f, we have

$$\begin{aligned}
\|\nabla_l u(x)\| &\leq \sum_{k\in[\mathsf{b}]} \int_0^\infty \mathbb{E}\left[\left\|\nabla_{x_l} X_{t,k}^x\right\| \|\nabla_k f(X_t^x)\|\right] \mathrm{d}t \\
&\leq \sum_{k\in[\mathsf{b}]} \int_0^\infty \mathrm{e}^{-Mt} \sum_{r=\mathsf{d}_\mathsf{G}(k,l)}^\infty \frac{t^r(M-m)^r}{r!} \cdot L_f \exp(-c_f \mathsf{d}_\mathsf{G}(i,k))\mathrm{d}t \\
&= \frac{L_f}{M} \sum_{k\in[\mathsf{b}]} \sum_{r=\mathsf{d}_\mathsf{G}(k,l)}^\infty \left(1 - \frac{m}{M}\right)^r \cdot \exp(-c_f \mathsf{d}_\mathsf{G}(i,k)) \\
&= \frac{L_f}{m} \sum_{k\in[\mathsf{b}]} \left(1 - \frac{m}{M}\right)^{\mathsf{d}_\mathsf{G}(k,l)} \cdot \exp(-c_f \mathsf{d}_\mathsf{G}(i,k)).
\end{aligned}$$

Therefore, as in the proof of Theorem 3.5, we have

$$
\begin{aligned}
&\Big|\int f(x)g(x)\pi(x)\mathrm{d}x\Big| \\
=&\left|\sum_{l\in[\mathsf{b}]}\int \nabla_l u(x)\cdot\nabla_l g(x)\pi(x)\mathrm{d}x\right| \\
\le&\frac{L_f L_g}{m}\sum_{k,l\in[\mathsf{b}]}\Big(1-\frac{m}{M}\Big)^{\mathsf{d}_{\mathsf{G}}(k,l)}\cdot\exp(-c_f\mathsf{d}_{\mathsf{G}}(i,k))\exp(-c_g\mathsf{d}_{\mathsf{G}}(j,l)) \\
\le&\frac{L_f L_g}{m}\exp(-c\mathsf{d}_{\mathsf{G}}(i,j))\sum_{r=0}^{\infty}\exp(-cr) \\
&\qquad\cdot\#\{(k,l)\in[\mathsf{b}]^2\mid \mathsf{d}_{\mathsf{G}}(i,k)+\mathsf{d}_{\mathsf{G}}(k,l)+\mathsf{d}_{\mathsf{G}}(j,l)=\mathsf{d}_{\mathsf{G}}(i,j)+r\}.
\end{aligned}
$$

Here we denote

$$
c=\min\left\{c_f,c_g,-\log\Big(1-\frac{m}{M}\Big)\right\}.
$$

Since G is (s,ν)-localized, we have

$$
\begin{aligned}
&\#\{(k,l)\in[\mathsf{b}]^2:\mathsf{d}_{\mathsf{G}}(i,k)+\mathsf{d}_{\mathsf{G}}(k,l)+\mathsf{d}_{\mathsf{G}}(j,l)=\mathsf{d}_{\mathsf{G}}(i,j)+r\} \\
\le&\sum_{d_1+d_2+d_2=\mathsf{d}_{\mathsf{G}}(i,j)+r}(1+\mathsf{s}d_1^{\nu})\,(1+\mathsf{s}d_2^{\nu})\,(1+\mathsf{s}d_3^{\nu}) \\
\le&\binom{\mathsf{d}_{\mathsf{G}}(i,j)+r+2}{2}\left(1+\mathsf{s}\left(\mathsf{d}_{\mathsf{G}}(i,j)+r\right)^{\nu}\right)^3 \\
\le&C\mathsf{s}^3\left(\mathsf{d}_{\mathsf{G}}(i,j)+r\right)^{3\nu+2},
\end{aligned}
$$

for some universal constant $C>0$. Note

$$
\begin{aligned}
&\sum_{r=0}^{\infty}\exp(-cr)\left(\mathsf{d}_{\mathsf{G}}(i,j)+r\right)^{3\nu+2} \\
\le&\sum_{r=0}^{\infty}\exp(-cr)2^{3\nu+2}\left(\left(\mathsf{d}_{\mathsf{G}}(i,j)\right)^{3\nu+2}+r^{3\nu+2}\right) \\
\le&2^{3\nu+2}\left(c^{-1}(\mathsf{d}_{\mathsf{G}}(i,j))^{3\nu+2}+\mathrm{e}^c c^{-3(\nu+1)}\Gamma(3(\nu+1))\right).
\end{aligned}
$$

Here we use

$$\begin{aligned}\sum_{r=0}^{\infty}\exp(-cr)r^{3\nu+2} &\le \int_0^{\infty}\exp(-cx)(x+1)^{3\nu+2}\mathrm{d}x\\ &= \mathrm{e}^c\int_1^{\infty}\exp(-cx)x^{3\nu+2}\mathrm{d}x\\ &\le \mathrm{e}^c c^{-3(\nu+1)}\Gamma(3(\nu+1)).\end{aligned}$$

Therefore, we obtain

$$\left|\int f(x)g(x)\pi(x)\mathrm{d}x\right| \le L_f L_g \psi_{3(\nu+1)}(\mathsf{d}_{\mathsf{G}}(i,j))\exp(-c\mathsf{d}_{\mathsf{G}}(i,j)).$$

Here we denote

$$\psi_{3(\nu+1)}(x) = C\frac{2^{3\nu+2}}{m}\mathsf{s}^3\left(c^{-1}x^{3\nu+2} + \mathrm{e}^c c^{-3(\nu+1)}\Gamma(3(\nu+1))\right).$$

This completes the proof. □

3.3 Gradient estimate of marginal Stein equation

We now prove the gradient estimate of the marginal Stein equation (3.1.3). The idea is to use the explicit solution (3.3.1) by Dynkin's formula, and represent its gradient as an expectation of the derivate of the path solution of the Langevin dynamics (3.3.2). The main technical part is to control the diffusion speed of the Langevin dynamics using the Dyson series [43]. The idea originates from the polynomial approximation of the inverse of a banded matrix in [38, 7, 6], and we generalize it to the case of time-dependent banded matrix. We mention that [52] derive a decay result for this case (termed *time-ordered exponential*), but their bound depends on the total dimension, which arises from a combinatorial term in the path-integral formula. Our result avoids the dimension dependence by using Dyson series, see [33] for more discussions. We also document our earlier proof using PDE analysis approach in [46] for the diagonal dominant case.

3.3.1 Explicit solution of Stein equation

Lemma 3.1. *Suppose $\pi \in \mathcal{P}(\mathbb{R}^d)$ is strongly log-concave. For any $i \in [\mathsf{b}]$ and 1-Lipschitz function $\phi_i : \mathbb{R}^{d_i} \to \mathbb{R}$, the solution of the marginal Stein equation*

(3.1.3) *is given by (up to a constant)*

$$u(x) = -\int_0^\infty \mathbb{E}\left[\phi_i(X_{t,i}^x) - \mathbb{E}_{x'\sim\pi}[\phi_i(x_i')]\right] \mathrm{d}t. \tag{3.3.1}$$

where X_t^x *is the path solution of the overdamped Langevin dynamics*

$$\mathrm{d}X_t^x = \nabla \log \pi(X_t^x)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \quad X_0^x = x. \tag{3.3.2}$$

As a corollary, the gradient estimate holds

$$\|\nabla_j u(x)\| \le \int_0^\infty \mathbb{E}\|\nabla_{x_j} X_{t,i}^x\| \mathrm{d}t. \tag{3.3.3}$$

Remark 3.4. When the Stein operator $\mathcal{L}_\pi$ is a generator of a process, it is known that the according Stein equation admits explicit solutions (3.3.1) (see [4]).

Proof. Let X_t^x solves the Langevin dynamics (3.3.2). By Dynkin's formula [83],

$$\begin{aligned} \mathbb{E}[u(X_T^x)] - u(x) &= \mathbb{E}\int_0^T (\nabla \log \pi(X_t^x)\cdot \nabla u(X_t^x) + \Delta u(X_t^x))\,\mathrm{d}t \\ &= \int_0^T \mathbb{E}\left[\phi_i(X_{t,i}^x) - \mathbb{E}_{x_i'\sim\pi}[\phi_i(x_i')]\right] \mathrm{d}t. \end{aligned} \tag{3.3.4}$$

Since π is strongly log-concave, it is well-known that $\mathsf{Law}(X_t^x)$ converges to π exponentially [3]. It implies that the limit $T\to\infty$ exists for both sides in (3.3.4), and the limit is

$$\int u(x)\pi(x)\mathrm{d}x - u(x) = \int_0^\infty \mathbb{E}\left[\phi_i(X_{t,i}^x) - \mathbb{E}_{x_i'\sim\pi}[\phi_i(x_i')]\right] \mathrm{d}t.$$

This gives (3.3.1) up to a constant. Taking derivative w.r.t x_j gives

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

Note it is valid due to the exponential decay of $\nabla_{x_j} X_{t,i}^x$. Since ϕ_i is 1-Lipschitz,

$$\left\|\nabla_{x_j} u(x)\right\| \le \int_0^\infty \mathbb{E}\left[\|\nabla_{x_j} X_{t,i}^x\| \left\|\nabla\phi_i(X_{t,i}^x)\right\|\right] \mathrm{d}t \le \int_0^\infty \mathbb{E}\|\nabla_{x_j} X_{t,i}^x\| \mathrm{d}t.$$

This completes the proof. □

3.3.2 A key lemma

We now prove the following key technical lemma, which essentially controls the diffusion speed in graphs. The proof is based on the polynomial approximation trick in [38] and Dyson series [43].

Lemma 3.2. *Let $H_t \in \mathbb{R}^{d\times d}$ be a time-dependent positive definite matrix satisfying:*

1. *H_t has dependency graph G, i.e. $H_t(i,j) = 0$ if $\mathsf{d}_{\mathsf{G}}(i,j) > 1$.*
2. *$\exists M > 0$ s.t. $\forall t \geq 0,\ 0 \preceq H_t \preceq MI$.*

Here $I \in \mathbb{R}^{d\times d}$ denotes the identity matrix. Consider the matrix ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}G_t = -H_t G_t, \quad G_0 = I. \tag{3.3.5}$$

Then for any $t \geq 0$, it holds that

$$\|G_t(i,j)\| \leq \exp(-tM) \sum_{k=\mathsf{d}_{\mathsf{G}}(i,j)}^{\infty} \frac{t^k M^k}{k!}. \tag{3.3.6}$$

Proof. **I. Scaling.** First note by a scaling argument, it suffices to consider $t = 1$. Consider $t = t_0$. Let G_t solves (3.3.5), then $G_{t_0 s}$ solves

$$\frac{\mathrm{d}}{\mathrm{d}s}G_{t_0 s} = -t_0 H_{t_0 s} G_{t_0 s}, \quad G_0 = I.$$

If the theorem holds for $t = 1$, then we obtain that at $s = 1$ (note $\|t_0 H_{t_0 s}\| \leq M t_0$),

$$\|G_{t_0}(i,j)\| \leq \exp(-Mt_0) \sum_{k=\mathsf{d}_{\mathsf{G}}(i,j)}^{\infty} \frac{M^k t_0^k}{k!}.$$

II. Dyson series solution. By variation of constants formula, we have

$$G_t = I - \int_0^t H_s G_s \mathrm{d}s.$$

Applying this identity recursively, we obtain

$$\begin{aligned} G_t &= I - \int_0^t H_s \left(I - \int_0^s H_u G_u \mathrm{d}u \right) \mathrm{d}s \\ &= I - \int_0^t H_s \mathrm{d}s + \int_0^t \int_0^s H_s H_u G_u \mathrm{d}u \mathrm{d}s = \cdots \\ &= I + \sum_{n=1}^{N-1} (-1)^n \int_{[0,t]^n} H_{t_n} \cdots H_{t_1} \mathbf{1}_{t_1 \le t_2 \le \cdots \le t_n} \mathrm{d}t_1 \cdots \mathrm{d}t_n \\ &\quad + (-1)^N \int_{[0,t]^N} H_{t_N} \cdots H_{t_1} G_{t_1} \mathbf{1}_{t_1 \le t_2 \le \cdots \le t_N} \mathrm{d}t_1 \cdots \mathrm{d}t_N. \end{aligned}$$

For simplicity, denote

$$X_0(t) = I, \quad X_n(t) := \frac{n!}{t^n} \int_{[0,t]^n} H_{t_n} \cdots H_{t_1} \mathbf{1}_{t_1 \le t_2 \le \cdots \le t_n} \mathrm{d}t_1 \cdots \mathrm{d}t_n, \ n \ge 1. \quad (3.3.7)$$

$$R_N(t) = \int_{[0,t]^N} H_{t_N} \cdots H_{t_1} G_{t_1} \mathbf{1}_{t_1 \le t_2 \le \cdots \le t_N} \mathrm{d}t_1 \cdots \mathrm{d}t_N.$$

Notice $X_n(t)$ is the average of 'H^n' in $[0,t]$. Then

$$G_t = \sum_{n=0}^{N-1} (-1)^n \frac{t^n}{n!} X_n(t) + R_N(t).$$

Now we prove that $\lim_{N\to\infty} R_N(t) = 0$. First notice $\|G_t\|_2 \le 1$, since

$$\frac{\mathrm{d}}{\mathrm{d}t} G_t^{\mathrm{T}} G_t = -2 G_t^{\mathrm{T}} H_t G_t \ \Rightarrow \ G_t^{\mathrm{T}} G_t = I - 2 \int_0^t G_s^{\mathrm{T}} H_s G_s \mathrm{d}s \preceq I.$$

So that as $N \to \infty$,

$$\begin{aligned} \|R_N(t)\|_2 \le & \int_{[0,t]^N} \|H_{t_N}\|_2 \cdots \|H_{t_1}\|_2 \, \|G_{t_1}\|_2 \, \mathbf{1}_{t_1 \le t_2 \le \cdots \le t_N} \mathrm{d}t_1 \cdots \mathrm{d}t_n \\ &\le M^N \int_{[0,t]^N} \mathbf{1}_{t_1 \le t_2 \le \cdots \le t_N} \mathrm{d}t_1 \cdots \mathrm{d}t_N = \frac{M^N t^N}{N!} \to 0. \end{aligned}$$

This proves that the Dyson series converges, and

$$G_t = \sum_{n=0}^{\infty} (-1)^n \frac{t^n}{n!} X_n(t).$$

III. Representation of polynomials. Denote the matrix process space

$$\mathscr{X} = \mathrm{span}\{X_n(t), n \geq 0\} = \overline{\Big\{ \sum_{k=0}^{n} a_k X_n(t) : a_k \in \mathbb{C} \Big\}}.$$

We define the representation of any polynomial P in $\mathscr{X}$ as

$$P[X](t) = \sum_{k=0}^{n} a_k X_k(t), \quad \text{if } P(x) = \sum_{k=0}^{n} a_k x^k.$$

Note that it can be extended from polynomials to any analytic functions. In Lemma 3.3, we show that the representation has an equivalent definition: if $P(x) = a_n \prod_{k=1}^{n}(x - x_k)$, then

$$\begin{aligned} P[X](t) = \frac{a_n}{t^n} \sum_{\sigma \in S_n} \int_0^t (H_{t_1} - x_{\sigma_1} I) \int_0^{t_1} (H_{t_2} - x_{\sigma_2} I) \cdots \\ \cdots \int_0^{t_{n-1}} (H_{t_n} - x_{\sigma_n} I) \mathrm{d}t_n \cdots \mathrm{d}t_2 \mathrm{d}t_1. \end{aligned} \tag{3.3.8}$$

Here S_n is the permutation group of degree n.

IV. Banded matrix approximation. By Taylor expansion,

$$\exp(-x) = \exp(-M) \sum_{n=0}^{\infty} \frac{\hat{X}_n(x)}{n!}, \quad \hat{X}_n(x) = (M - x)^n.$$

Represent the series in $\mathscr{X}$, we obtain

$$G_1 = \exp(-M) \sum_{n=0}^{\infty} \frac{1}{n!} \hat{X}_n[X](1),$$

$$\hat{X}_n[X](1) = \sum_{\sigma \in S_n} \int_{[0,1]^n} (MI - H_{t_1}) \cdots (MI - H_{t_n}) \mathbf{1}_{t_1 \leq t_2 \leq \cdots \leq t_n} \mathrm{d}t_1 \cdots \mathrm{d}t_n.$$

Here we use the alternative representation (3.3.8). We can truncate the Dyson series of G_1 as

$$G_1 = \exp(-M) \sum_{r=0}^{n} \frac{1}{r!} \hat{X}_r[X](1) + \exp(-M) \sum_{r=n+1}^{\infty} \frac{1}{r!} \hat{X}_r[X](1).$$

Consider the off-diagonal entry $G_1(i,j)$. Take $n = \mathsf{d}_{\mathsf{G}}(i,j) - 1$, then since all the path in G connecting i and j has length no less than $\mathsf{d}_{\mathsf{G}}(i,j) > n$, it must hold that

$$\forall 1 \le r \le n, \quad [(MI - H_{t_1}) \cdots (MI - H_{t_r})]\,(i,j) = 0 \;\Rightarrow\; \hat{X}_r[X](1)(i,j) = 0,$$

Therefore,

$$G_1(i,j) = \exp(-M) \sum_{r=n+1}^{\infty} \frac{1}{r!} \hat{X}_r[X](1)(i,j).$$

Since $0 \preceq H_{t_k} \preceq MI$, it holds that $\|MI - H_{t_k}\|_{\mathrm{op}} \le M$, and thus

$$\|\hat{X}_n[X](1)\|_{\mathrm{op}} \le M^n n! \int_{[0,1]^n} \mathbf{1}_{t_1 \le t_2 \le \cdots \le t_n} \mathrm{d}t_1 \cdots \mathrm{d}t_n = M^n.$$

$$\Rightarrow\; \|G_1(i,j)\| \le \exp(-M) \sum_{r=n+1}^{\infty} \frac{M^r}{r!} = \exp(-M) \sum_{r=\mathsf{d}_{\mathsf{G}}(i,j)}^{\infty} \frac{M^r}{r!}.$$

This verifies the case when $i \neq j$. For $i = j$, the result directly follows from $\|G_1(i,i)\| \le \|G_1\| \le 1$. □

Lemma 3.3. *The two representations of polynomials in $\mathscr{X}$ are equivalent, i.e.*

$$P[X](t) = P\{X\}(t), \quad \forall P \in \mathbb{P}.$$

where we denote $P\{X\}(t)$ for polynomial $P = a_n \prod_{k=1}^{n}(x - x_k)$ as

$$P\{X\}(t) = \frac{a_n}{t^n} \sum_{\sigma \in S_n} \int_0^t (H_{t_1} - x_{\sigma_1} I) \int_0^{t_1} (H_{t_2} - x_{\sigma_2} I) \cdots\cdots \int_0^{t_{n-1}} (H_{t_n} - x_{\sigma_n} I) \mathrm{d}t_n \cdots \mathrm{d}t_2 \mathrm{d}t_1.$$

Proof. We prove by induction. First note $n = 1$ is obvious,

$$(a_1(x - x_1))\{X\}(t) = \frac{a_1}{t} \int_0^t (H_{t_1} - x_1 I) \mathrm{d}t_1 = a_1(X_1(t) - x_1 I) = (a_1(x - x_1))[X](t).$$

Now consider $n \ge 2$. Notice $P[X](t), P\{X\}(t)$ can both be viewed as mulitlinear

maps on $x_1, \ldots, x_n$. Take the partial derivative w.r.t. x_n,

$$
\begin{aligned}
&\nabla_{x_n}\left(P\{X\}(t)\right) \\
=\ & \frac{a_n}{t^n}\sum_{k=1}^{n}\sum_{\sigma\in S_n,\sigma_k=n}\int_0^t (H_{t_1}-x_{\sigma_1}I)\cdots\int_0^{t_{k-1}}\nabla_{x_n}(H_{t_k}-x_nI)\cdots\int_0^{t_{n-1}}(H_{t_n}-x_{\sigma_n}I)\mathrm{d}t_n\cdots\mathrm{d}t_1 \\
=\ & -\frac{a_n}{t^n}\sum_{k=1}^{n}\sum_{\sigma\in S_n,\sigma_k=n}\int_0^t (H_{t_1}-x_{\sigma_1}I)\cdots\int_0^{t_{k-1}} I\cdots\int_0^{t_{n-1}}(H_{t_n}-x_{\sigma_n}I)\mathrm{d}t_n\cdots\mathrm{d}t_1 \\
=\ & -\frac{a_n}{t^n}\sum_{k=1}^{n}\sum_{\sigma\in S_n,\sigma_k=n}\int_0^t (H_{t_1}-x_{\sigma_1}I)\cdots(t_{k-1}-t_{k+1})\cdots\int_0^{t_{n-1}}(H_{t_n}-x_{\sigma_n}I)\mathrm{d}t_n\cdots\widehat{\mathrm{d}t}_k\cdots\mathrm{d}t_1 \\
=\ & -\frac{a_n}{t^n}\sum_{k=1}^{n}\sum_{\sigma\in S_{n-1}}\int_0^t (H_{t_1}-x_{\sigma_1}I)\cdots(t_{k-1}-t_k)\cdots\int_0^{t_{n-2}}(H_{t_{n-1}}-x_{\sigma_{n-1}}I)\mathrm{d}t_{n-1}\cdots\mathrm{d}t_1 \\
=\ & -\frac{a_n}{t^n}\cdot t\sum_{\sigma\in S_{n-1}}\int_0^t (H_{t_1}-x_{\sigma_1}I)\cdots\int_0^{t_{n-2}}(H_{t_{n-1}}-x_{\sigma_{n-1}}I)\mathrm{d}t_{n-1}\cdots\mathrm{d}t_1 \\
=\ & -a_n\left(\prod_{k=1}^{n-1}(x-x_k)\right)\{X\}(t) = -a_n\left(\prod_{k=1}^{n-1}(x-x_k)\right)[X](t).
\end{aligned}
$$

Here the first equality follows from discussion on the position of x_n. The third equality follows from integrating the variable t_k and notice the constraint $t_{k+1} \leq t_k \leq t_{k-1}$. The forth equality follows from relabeling the index and taking $t_n = 0$. The last equality follows from the induction hypothesis. By symmetry, the relation holds for any x_i:

$$
\nabla_{x_i}\left(P\{X\}(t)\right) = -\Big(a_n\prod_{j:j\neq i}(x-x_j)\Big)[X](t) = (\nabla_{x_i}P)[X](t),
$$

since $\nabla_{x_i}P(x) = -a_n\prod_{j:j\neq i}(x-x_j)$. Now notice the representation $P[X]$ is linear in the coefficients of P, it is direct to verify

$$
\nabla_{x_i}(P[X](t)) = (\nabla_{x_i}P)[X](t) = \nabla_{x_i}\left(P\{X\}(t)\right).
$$

Finally notice when $x_1 = \cdots = x_n = 0$,

$$
x^n\{X\}(t) = \frac{n!}{t^n}\int_0^t H_{t_1}\int_0^{t_1}H_{t_2}\cdots\int_0^{t_{n-1}}H_{t_n}\mathrm{d}t_1\cdots\mathrm{d}t_n = X_n(t) = x^n[X](t).
$$

Now the two multi-linear maps agree on one point, and also on all the partial derivatives. So that they must be identical. This shows the equivalence holds for n, and by induction, it holds for all n. □

3.3.3 Proof of Theorem 3.1

Proof of Theorem 3.1. Under the conditions in Theorem 3.1, Lemma 3.1 holds. By (3.3.3), it remains to control $\nabla_{x_j} X_{t,i}^x$. Taking derivative w.r.t. x in (3.3.2), we get

$$\mathrm{d}\nabla X_t^x = -H_t \cdot \nabla X_t^x \mathrm{d}t, \quad H_t := -\nabla^2 \log \pi(X_t^x). \tag{3.3.9}$$

Note here $\nabla X_t^x = \nabla_x X_t^x \in \mathbb{R}^{d\times d}$. Denote $G_t = \mathrm{e}^{mt}\nabla X_t^x$ and $\tilde{H}_t = H_t - mI$, then it holds that

$$\frac{\mathrm{d}}{\mathrm{d}t} G_t = \mathrm{e}^{mt}\left(m\nabla X_t^x - H_t \nabla X_t^x\right) = -\tilde{H}_t G_t, \quad G_0 = \nabla X_0^x = I.$$

By assumption, $0 \preceq \tilde{H}_t \preceq (M-m)I$, and $\tilde{H}_t(i,j) = 0$ if $\mathsf{d}_{\mathsf{G}}(i,j) > 1$. By Lemma 3.2,

$$\mathrm{e}^{mt}\|\nabla_{x_j} X_{t,k}^x\| = \|G_t(k,j)\| \le \mathrm{e}^{-(M-m)t} \sum_{r=\mathsf{d}_{\mathsf{G}}(j,k)}^{\infty} \frac{t^r (M-m)^r}{r!}. \tag{3.3.10}$$

The estimate holds for any initial condition x. For different vertices $j \in [\mathsf{b}]$, consider different initial conditions $x^{(j)}$, and take summation over $j \in [\mathsf{b}]$, we obtain

$$\begin{aligned}
&\sum_{j\in[\mathsf{b}]} \|\nabla_{x_j^{(j)}} X_{t,k}^{x^{(j)}}\| \\
&\le \mathrm{e}^{-mt}\mathrm{e}^{-(M-m)t} \sum_{j\in[\mathsf{b}]} \sum_{r=\mathsf{d}_{\mathsf{G}}(j,k)}^{\infty} \frac{t^r (M-m)^r}{r!} \\
&= \mathrm{e}^{-mt}\mathrm{e}^{-(M-m)t} \left(\sum_{r=0}^{\infty} \frac{t^r (M-m)^r}{r!} + \sum_{k=1}^{\infty} \sum_{j:0<\mathsf{d}_{\mathsf{G}}(j,k)\le r} \frac{t^r (M-m)^r}{r!} \right) \\
&\le \mathrm{e}^{-mt} + \mathrm{e}^{-Mt} \sum_{r=1}^{\infty} \mathsf{s} r^{\nu} \frac{t^r (M-m)^r}{r!}.
\end{aligned}$$

Here we use the sparsity condition $|\mathcal{N}_j^r \backslash \{j\}| \le \mathsf{s} r^\nu$ (2.1.3). From (3.3.3), we obtain

$$\begin{aligned}\sum_{j\in[\mathsf{b}]} \|\nabla_j u(x^{(j)})\| &\le \sum_{j\in[\mathsf{b}]} \int_0^\infty \mathbb{E}\|\nabla_{x_j^{(j)}} X_{t,k}^{x^{(j)}}\| \mathrm{d}t \\ &= \mathbb{E}\int_0^\infty \sum_{j\in[\mathsf{b}]} \|\nabla_{x_j^{(j)}} X_{t,k}^{x^{(j)}}\| \mathrm{d}t \\ &\le \int_0^\infty \left(\mathrm{e}^{-mt} + \mathsf{s}\mathrm{e}^{-Mt} \sum_{r=1}^\infty r^\nu \frac{t^r (M-m)^r}{r!} \right) \mathrm{d}t \\ &= \frac{1}{m} + \frac{\mathsf{s}}{M} \sum_{r=1}^\infty r^\nu \left(1 - \frac{m}{M}\right)^r.\end{aligned}$$

By Lemma 3.6, it holds that (denote $\kappa = \frac{M}{m}$)

$$\begin{aligned}\sum_{j\in[\mathsf{b}]} \|\nabla_j u(x^{(j)})\| &\le \frac{1}{m} + \frac{\mathsf{s}}{M} \nu! \left(\frac{m}{M}\right)^{-\nu-1} \left(1 - \frac{m}{M}\right) \\ &= \frac{1}{m}\left(1 + \mathsf{s}\nu!\kappa^\nu (1 - \kappa^{-1})\right) \le \frac{\mathsf{s}\nu!\kappa^\nu}{m}.\end{aligned}$$

Taking supremum over $x^{(j)}$, we obtain the gradient estimate with $\delta = \frac{\mathsf{s}\nu!\kappa^\nu}{m}$. □

3.3.4 Proof of Theorem 3.2

We provide two proofs of Theorem 3.2. The first uses similar arguments as in Theorem 3.1. The the second proof (from [46]) is based on the maximum principle of the elliptic PDE.

Stochastic analysis approach

Proof I of Theorem 3.2. Similar as the proof of Theorem 3.1, it remains to control ∇X_t^x in (3.3.9). To control the 2-norm of $\nabla_{x_j} X_{t,k}^x \in \mathbb{R}^{d_k \times d_j}$, consider fixing a test vector $v_j \in \mathbb{R}^{d_j}$ s.t. $\|v_j\| = 1$. For $i, j \in [\mathsf{b}]$, denote $g_t^x(k,j) = \nabla_{x_j} X_{t,k}^x \cdot v_j \in \mathbb{R}^{d_k}$. Then

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} g_t^x(k,j) &= \frac{\mathrm{d}}{\mathrm{d}t} \nabla_{x_j} X_{t,k}^x v_j = -H_t \cdot \nabla_{x_j} X_t^x \cdot v_j \\ &= -\sum_{l=1}^{\mathsf{b}} H_t(k,l) \nabla_{x_j} X_{t,l}^x \cdot v_j = -\sum_{l=1}^{\mathsf{b}} H_t(k,l) g_t^x(l,j).\end{aligned}$$

Here we denote $H_t(k,l) \in \mathbb{R}^{d_k \times d_l}$ as the (k,l)-th subblock of H_t. Then by the diagonal dominance assumption,

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|g_t^x(k,j)\right\|^2 = & -\sum_{l=1}^{\mathsf{b}}(g_t^x(k,j))^{\mathrm{T}} H_t(k,l) g_t^x(l,j) \\ \leq & -M_{kk}\left\|g_t^x(k,j)\right\|^2 + \sum_{l:l\neq k} M_{kl}\left\|g_t^x(k,j)\right\|\left\|g_t^x(l,j)\right\|.\end{aligned}$$

Notice $\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left\|g_t^x(k,j)\right\|^2 = \left\|g_t^x(k,j)\right\|\frac{\mathrm{d}}{\mathrm{d}t}\left\|g_t^x(k,j)\right\|$, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\|g_t^x(k,j)\right\| \leq -M_{kk}\left\|g_t^x(k,j)\right\| + \sum_{l:l\neq k} M_{kl}\left\|g_t^x(l,j)\right\|.$$

Note this inequality holds for any index $j,k \in [\mathsf{b}]$, test vector v_j and initial condition x. For different indices $j \in [\mathsf{b}]$, consider different initial conditions $x^{(j)}$, and denote the matrix $G_t \in \mathbb{R}^{\mathsf{b}\times\mathsf{b}}$ where

$$G_t(k,j) = \|g_t^{x^{(j)}}(k,j)\|.$$

Then the above inequality can be written compactly in a matrix form

$$\frac{\mathrm{d}}{\mathrm{d}t}G_t \leq -\tilde{M}G_t, \quad \tilde{M}_{ij} := 2M_{ii}\mathbf{1}_{i=j} - M_{ij}.$$

Here $\leq$ is in the entrywise sense. Note the initial condition is $G_0 = I$, since

$$G_0(k,j) = \|g_0^{x^{(j)}}(k,j)\| = \|\nabla_{x_j^{(j)}} X_{0,k}^{x^{(j)}} v_j\| = \|\nabla_{x_j^{(j)}} x_k^{(j)} v_j\| = \delta_{jk}\|v_j\| = \delta_{jk}.$$

By assumption, $\forall i,j \in [\mathsf{b}], i \neq j$, $\tilde{M}_{ij} = -M_{ij} \leq 0$, and

$$\sum_{j:j\neq i}|\tilde{M}_{ij}| + c = \sum_{j:j\neq i} M_{ij} + c \leq M_{ii} = \tilde{M}_{ii}.$$

Thus we can apply Lemma 3.7 and obtain

$$\|G_t\|_\infty \leq \mathrm{e}^{-ct}\|G_0\|_\infty = \mathrm{e}^{-ct} \;\Rightarrow\; \max_k \sum_j \|\nabla_{x_j^{(j)}} X_{t,k}^{x^{(j)}} v_j\| \leq \mathrm{e}^{-ct}.$$

Since v_j is arbitrary, we obtain that

$$\max_k \sum_j \|\nabla_{x_j^{(j)}} X_{t,k}^{x^{(j)}}\| \le \mathrm{e}^{-ct}.$$

Recall (3.3.3), this implies

$$\begin{aligned}\sum_j \|\nabla_j u(x^{(j)})\| &\le \sum_j \int_0^\infty \mathbb{E}\|\nabla_{x_j^{(j)}} X_{t,i}^{x^{(j)}}\| \mathrm{d}t \\ &= \mathbb{E}\int_0^\infty \sum_j \|\nabla_{x_j^{(j)}} X_{t,i}^{x^{(j)}}\| \mathrm{d}t \\ &\le \mathbb{E}\int_0^\infty \mathrm{e}^{-ct}\mathrm{d}t = c^{-1}.\end{aligned}$$

Now as $x^{(j)}$ is arbitrary, we obtain the gradient estimate with $\delta = c^{-1}$. □

PDE analysis approach

Proof II of Theorem 3.2. Let $v : \mathbb{R}^d \to \mathbb{R}^{d_j}$ be a vector-valued function. Denote $\mathcal{L}_\pi v$ as the entrywise application of the operator $\mathcal{L}_\pi$ (cf. (3.1.3)) on v_i, i.e.

$$\mathcal{L}_\pi v := (\mathcal{L}_\pi v_1, \dots \mathcal{L}_\pi v_{d_j})^{\mathrm{T}}.$$

Note it suffices to prove (3.1.6) for $\phi_i \in C^1 \cap \mathrm{Lip}_1$. Since this space is dense in Lip_1, so for general $\phi_i \in \mathrm{Lip}_1$, we can take a sequence of $\phi_i^{(k)} \in C^1 \cap \mathrm{Lip}_1$ that converges to ϕ_i. If (3.1.6) holds uniformly for $\phi_i^{(k)}$, then passing to the limit shows that it holds for any $\phi_i \in \mathrm{Lip}_1$.

Now fix any $\phi_i \in C^1 \cap \mathrm{Lip}_1$. It is straightforward to verify by standard elliptic theory that the solution exists (up to a constant) and $u \in C^3$. Taking the gradient w.r.t. x_j in (3.1.3), we obtain

$$\mathcal{L}_\pi(\nabla_j u)(x) + \sum_k \nabla_{jk}^2 \log \pi(x) \nabla_k u(x) = \delta_{ij} \nabla_j \phi_i(x_i).$$

Recall $H(x) = -\nabla^2 \log \pi(x)$. Multiplying the above equality from left by $(\nabla_j u(x))^{\mathrm{T}}$,

and by $\phi_i \in \mathrm{Lip}_1$ and the diagonal dominance assumption, we have

$$\begin{aligned}
&(\nabla_j u(x))^{\mathrm{T}} \mathcal{L}_\pi(\nabla_j u)(x) \\
=& \sum_k (\nabla_j u(x))^{\mathrm{T}} H_{jk}(x) \nabla_k u(x) + \delta_{ij} (\nabla_j u(x))^{\mathrm{T}} \nabla_j \phi_i(x_i) \\
\geq& M_{jj} \|\nabla_j u(x)\|^2 - \sum_{k:k\neq j} M_{ij} \|\nabla_j u(x)\| \|\nabla_k u(x)\| - \delta_{ij} \|\nabla_j u(x)\|.
\end{aligned} \tag{3.3.11}$$

The key is to show the maximum principle still holds for the operator $\mathcal{L}_\pi$ when acting on a vector-valued function. Consider x where $\|\nabla_j u(x)\|_2$ reaches its maximum, i.e. $\|\nabla_j u(x)\|_2 = \|\nabla_j u\|_{L^\infty}$. The first order optimality condition reads

$$0 = \nabla \left(\|\nabla_j u(x)\|_2^2 \right) = 2 \nabla \nabla_j u(x) \cdot \nabla_j u(x),$$

and the second order optimality condition reads

$$0 \geq \Delta \left(\|\nabla_j u(x)\|_2^2 \right) = 2 \|\nabla \nabla_j u(x)\|_{\mathrm{F}}^2 + 2 (\nabla_j u(x))^{\mathrm{T}} \Delta \nabla_j u(x).$$

Thus, $(\nabla_j u(x))^{\mathrm{T}} \Delta \nabla_j u(x) \leq 0$. Under these conditions,

$$(\nabla_j u(x))^{\mathrm{T}} \mathcal{L}_\pi (\nabla_j u(x)) = (\nabla_j u(x))^{\mathrm{T}} \Delta \nabla_j u(x) + \nabla \log \pi(x) \cdot \nabla \nabla_j u(x) \cdot \nabla_j u(x) \leq 0.$$

Hence, at the maximum point (3.3.11) reads

$$0 \geq M_{jj} \|\nabla_j u(x)\|^2 - \sum_{k:k\neq j} M_{jk} \|\nabla_j u(x)\| \|\nabla_k u(x)\| - \delta_{ij} \|\nabla_j u(x)\|.$$

If $\|\nabla_j u\|_{L^\infty} > 0$, it holds that

$$\delta_{ij} \geq M_{jj} \|\nabla_j u(x)\| - \sum_{k:k\neq j} M_{jk} \|\nabla_k u(x)\|.$$

Taking summation over $j \in \mathcal{I}_+ := \{j \in [\mathsf{b}] : \|\nabla_j u\|_{L^\infty} > 0\}$ gives

$$\begin{aligned}
1 \geq& \sum_{j\in\mathcal{I}_+} \delta_{ij} \geq \sum_{j\in\mathcal{I}_+} \left[M_{jj} \|\nabla_j u(x)\| - \sum_{k:k\neq j} M_{jk} \|\nabla_k u(x)\| \right] \\
\geq& \sum_{j\in[\mathsf{b}]} M_{jj} \|\nabla_j u(x)\| - \sum_{k\neq j} M_{jk} \|\nabla_k u(x)\| \\
=& \sum_{j\in[\mathsf{b}]} \left(M_{jj} - \sum_{k:k\neq j} M_{kj} \right) \|\nabla_j u(x)\| \geq c \sum_{j\in[\mathsf{b}]} \|\nabla_j u(x)\|.
\end{aligned}$$

Here we use the diagonal dominance assumption. The conclusion follows. □

3.4 Locality in Langevin semigroup

In the previous sections, we developed the marginal Stein's method, which is used to prove marginal transport inequalities and the exponential decay of correlations for localized distributions. In this section, we show that this method can be interpreted as a quantification of the locality in the Langevin semigroup. It is well known that when the target distribution is strongly log-concave, the associated Langevin semigroup is exponential contractive in the H^1-norm. To study the locality property, we adjust the norm to locality-aware variants. We show that the Langevin semigroup is *eventually exponentially contractive* under these new norms. We will discuss two applications of this eventual exponential contraction, the δ-locality and the convergence of Langevin dynamics under the $\mathsf{W}_{1,\infty}$-distance.

3.4.1 Langevin semigroup

Consider the (overdamped) Langevin dynamics

$$\mathrm{d}X_t^x = \nabla \log \pi(X_t^x)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \quad X_0^x = x, \tag{3.4.1}$$

where π is a target distribution and W_t is a standard Brownian motion. The Langevin semigroup $\{P_t\}_{t\geq 0}$ is defined as the transition semigroup of the Langevin dynamics, i.e., for $t \geq 0$,

$$P_t u(x) = \mathbb{E}[u(X_t^x)]. \tag{3.4.2}$$

It is straightforward to verify that $\{P_t\}_{t\geq 0}$ is a Markov semigroup, i.e.

- (Semigroup) $P_0 = \mathrm{id}$, $P_{s+t} = P_s \circ P_t$ for all $s, t \geq 0$.
- (Markovian) $P_t\mathbf{1} = \mathbf{1}$, and $P_t f \geq 0$ if $f \geq 0$.

Note its infinitesimal generator is $\mathcal{L}_\pi$, i.e.,

$$\mathcal{L}_\pi u(x) = \lim_{t\to 0} \frac{P_t u(x) - u(x)}{t} = \nabla \log \pi(x) \cdot \nabla u(x) + \Delta u(x). \tag{3.4.3}$$

Exponential contraction

Suppose π is strongly log-concave, i.e.

$$-\nabla^2 \log \pi(x) \succeq mI.$$

It is known that [3] that the Langevin semigroup is exponentially contractive. Here we use the gradient bound (see Section 3.3 in [3]) to illustrate it.

For any $u \in H_0^1(\pi) = \{u \in H^1(\pi) : \mathbb{E}_\pi[u] = 0\}$ (3.1.5), notice

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} \|\nabla P_t u\|_\pi^2 &= 2 \langle \nabla P_t u, \nabla \mathcal{L}_\pi P_t u \rangle_\pi \\ &= 2 \langle \nabla P_t u, \mathcal{L}_\pi (\nabla P_t u) \rangle_\pi + 2 \left\langle \nabla P_t u, \nabla^2 \log \pi(x) \cdot \nabla P_t u \right\rangle_\pi \\ &\leq -2 \left\| \nabla^2 P_t u \right\|_\pi^2 - 2m \|\nabla P_t u\|_\pi^2 \leq -2m \|\nabla P_t u\|_\pi^2 .\end{aligned}$$

Here we denote $\mathcal{L}v = (\mathcal{L}v_1, \dots \mathcal{L}v_d)$ if $v = (v_1, \dots, v_d)$ is a vector-valued function, and use the fact that $\langle f, \mathcal{L}_\pi g \rangle_\pi = -\langle \nabla f, \nabla g \rangle_\pi$. Then we obtain

$$\|\nabla P_t u\|_\pi^2 \leq \mathrm{e}^{-2mt} \|\nabla u\|_\pi^2 .$$

This implies that the P_t is exponentially contractive in the H^1-norm. In the operator form, we have

$$\|P_t\|_{H_0^1(\pi) \to H_0^1(\pi)} := \sup_{0 \neq u \in H_0^1(\pi)} \frac{\|P_t u\|_{H_0^1(\pi)}}{\|u\|_{H_0^1(\pi)}} \leq \mathrm{e}^{-mt}. \tag{3.4.4}$$

Gradient estimate of Stein equation

The solution to Stein equation can be formally written as

$$u = \mathcal{L}_\pi^{-1} (\phi - \mathbb{E}_\pi[\phi]) . \tag{3.4.5}$$

Under the settings in this thesis (see Section 3.1.2), $\mathcal{L}_\pi^{-1}$ can be regarded as an operator in $H_0^1(\pi)$. The Poincaré inequality implies

$$\|\mathcal{L}_\pi^{-1}\|_{H_0^1(\pi) \to H_0^1(\pi)} \leq \mathsf{C}_{\mathsf{PI}}.$$

This can be verified directly: for $u, \phi \in H_0^1(\pi)$,

$$\begin{aligned}\|u\|_{H_0^1(\pi)}^2 :=\ & \|\nabla u\|_{L^2(\pi)}^2 = -\langle \mathcal{L}_\pi u, u\rangle_\pi = -\langle \phi - \mathbb{E}_\pi[\phi], u\rangle_\pi \\ \leq\ & \|\phi - \mathbb{E}_\pi[\phi]\|_{L^2(\pi)} \|u\|_{L^2(\pi)} \leq \mathsf{C}_{\mathsf{PI}} \|\nabla \phi\|_{L^2(\pi)} \|\nabla u\|_{L^2(\pi)}.\end{aligned}$$

Under stronger assumption, i.e. the log-concavity condition, this bound can be directly obtained from the exponential contraction. Note formally,

$$\mathcal{L}_\pi^{-1} = -\int_0^\infty \mathrm{e}^{t\mathcal{L}_\pi} \mathrm{d}t = -\int_0^\infty P_t \mathrm{d}t. \tag{3.4.6}$$

Due to the exponential contraction, the above equation holds as a strict identity in the H^1-norm, i.e. for $\phi \in H_0^1(\pi)$,

$$u(x) = \mathcal{L}_\pi^{-1}\phi = -\int_0^\infty P_t\phi(x) \mathrm{d}t = -\int_0^\infty \mathbb{E}_\pi[\phi(X_t^x)] \mathrm{d}t.$$

This is precisely Lemma 3.1. As a result,

$$\left\|\mathcal{L}_\pi^{-1}\right\|_{H_0^1(\pi)\to H_0^1(\pi)} \leq \int_0^\infty \|P_t\|_{H_0^1(\pi)\to H_0^1(\pi)} \,\mathrm{d}t \leq \int_0^\infty \mathrm{e}^{-mt} \mathrm{d}t = \frac{1}{m}. \tag{3.4.7}$$

3.4.2 Eventual exponential contraction

To study the locality structure, we introduce two different metrics as variants of the H^1-norm used in the previous section. The first is the ℓ^1(Lip)-(semi)norm used in Definition 3.1:

$$|u|_{\ell^1(\mathrm{Lip})} := \sum_{j\in[\mathsf{b}]} \|\nabla_j u\|_{L^\infty}. \tag{3.4.8}$$

The other is a weaker version, the $|\cdot|_{\mathrm{Lip},\infty}$-(semi)norm:

$$|u|_{\mathrm{Lip},\infty} = \operatorname*{ess\,sup}_x \sum_{j\in[\mathsf{b}]} \|\nabla_j u(x)\|. \tag{3.4.9}$$

The reason for the name $|\cdot|_{\mathrm{Lip},\infty}$ is given in Lemma 3.5. Under the new norm, we show that the Langevin semigroup is *eventually* exponentially contractive when the target distribution is strongly log-concave and localized. This is in contrast to the exponential contraction in the H^1-norm, which is gauranteed by a positive spectral gap. While under the new norm, the generator no longer exhibits a spectral gap, and one can at most establish eventual exponential contraction. This behavior arises from the interplay between two opposing effects: the exponential decay due

to the strong log-concavity and the slower diffusion due to the locality of the graph.

We prove the following theorem.

Theorem 3.8. *Let* G *be a* (s, ν)*-local graph. Suppose* $\pi \in \mathcal{P}(\mathbb{R}^d)$ *is localized on* G*, and satisfies for some* $0 < m \le M < \infty$*,*

$$\forall x \in \mathbb{R}^d, \quad mI \preceq -\nabla^2 \log \pi(x) \preceq MI.$$

Then it holds that for all $t \ge 0$*,*

$$\max\left\{ \|P_t\|_{|\cdot|_{\ell^1(\mathrm{Lip})} \to |\cdot|_{\ell^1(\mathrm{Lip})}}, \|P_t\|_{|\cdot|_{\mathrm{Lip},\infty} \to |\cdot|_{\mathrm{Lip},\infty}} \right\} \le \mathsf{M}_t, \tag{3.4.10}$$

where

$$\mathsf{M}_t := \mathrm{e}^{-mt}\left(1 + \mathsf{s} p_\nu(t(M-m))\right), \quad p_k(x) := \mathrm{e}^{-x} \sum_{r=0}^{\infty} r^k \frac{x^r}{r!}. \tag{3.4.11}$$

As a corollary, denote $\kappa = \frac{M}{m}$*, and we have*

$$\max\left\{ \left\|\mathcal{L}_\pi^{-1}\right\|_{|\cdot|_{\ell^1(\mathrm{Lip})} \to |\cdot|_{\ell^1(\mathrm{Lip})}}, \left\|\mathcal{L}_\pi^{-1}\right\|_{|\cdot|_{\mathrm{Lip},\infty} \to |\cdot|_{\mathrm{Lip},\infty}} \right\} \le \frac{\mathsf{s}\nu!\kappa^\nu}{m}. \tag{3.4.12}$$

Remark 3.5. The bound M_t is not necessarily monotone in t. As M_t is a product of a polynomial and an exponential function in t, it typically first increases and then decreases. This explains the term eventual exponential contraction.

We also comment that this is not an artifact of the proof. The exponential term comes from the strong log-concavity, while the polynomial term describes the diffusion in the graph, which is due to the polynomial growth of the neighborhood volume in the graph.

Proof. Let $u \in H_0^1(\pi)$. By definition, we have

$$P_t u(x) = \mathbb{E}[u(X_t^x)],$$

where X_t^x is the path solution of the Langevin dynamics with initial state x. Note

$$\nabla_j P_t u(x) = \sum_{k \in [\mathsf{b}]} \mathbb{E}[\nabla_{x_j} X_{t,k}^x \cdot \nabla_k u(X_t^x)].$$

By the same argument as in the proof of Theorem 3.1, we have (cf. (3.3.10))

$$\|\nabla_{x_j} X^x_{t,k}\| \le \mathrm{e}^{-Mt} \sum_{r=\mathsf{d}_{\mathsf{G}}(j,k)}^{\infty} \frac{t^r (M-m)^r}{r!}.$$

Therefore, using the (s,ν)-locality, we have

$$\begin{aligned}
|P_t u|_{\ell^1(\mathrm{Lip})} &:= \sum_{j\in[\mathsf{b}]} \|\nabla_j P_t u\|_{L^\infty} \\
&\le \sum_{j,k\in[\mathsf{b}]} \mathbb{E}\left[\left\|\nabla_{x_j} X^x_{t,k}\right\| \|\nabla_k u(X^x_t)\|\right] \\
&\le \sum_{k\in[\mathsf{b}]} \|\nabla_k u\|_{L^\infty} \left[\sum_{j\in[\mathsf{b}]} \mathrm{e}^{-Mt} \sum_{r=\mathsf{d}_{\mathsf{G}}(j,k)}^{\infty} \frac{t^r (M-m)^r}{r!}\right] \\
&= \mathrm{e}^{-Mt} \sum_{k\in[\mathsf{b}]} \|\nabla_k u\|_{L^\infty} \left[\sum_{r=0}^{\infty} \frac{t^r (M-m)^r}{r!} + \sum_{r=1}^{\infty} \sum_{j:1\le \mathsf{d}_{\mathsf{G}}(j,k)\le r} \frac{t^r (M-m)^r}{r!}\right] \\
&\le \mathrm{e}^{-Mt} \sum_{k\in[\mathsf{b}]} \|\nabla_k u\|_{L^\infty} \left[\mathrm{e}^{(M-m)t} + \sum_{r=1}^{\infty} \mathsf{s} r^{\nu} \frac{t^r (M-m)^r}{r!}\right] \\
&\le |u|_{\ell^1(\mathrm{Lip})} \mathrm{e}^{-mt} \left(1 + \mathsf{s} p_\nu(t(M-m))\right).
\end{aligned}$$

where we denote

$$p_k(x) := \mathrm{e}^{-x} \sum_{r=0}^{\infty} r^k \frac{x^r}{r!}. \tag{3.4.13}$$

It is verified in Lemma 3.4 that p_k is a monic polynomial of degree k. We obtain

$$\|P_t\|_{|\cdot|_{\ell^1(\mathrm{Lip})} \to |\cdot|_{\ell^1(\mathrm{Lip})}} \le \mathrm{e}^{-mt} \left(1 + \mathsf{s} p_\nu(t(M-m))\right) =: \mathsf{M}_t.$$

For the $|\cdot|_{\mathrm{Lip},\infty}$-norm, the proof is similar:

$$\begin{aligned}
|P_t u|_{\mathrm{Lip},\infty} :=&\ \operatorname*{ess\,sup}_x \sum_{j\in[\mathsf{b}]} \|\nabla_j P_t u(x)\| \\
\le&\ \operatorname*{ess\,sup}_x \sum_{j,k\in[\mathsf{b}]} \mathbb{E}\left[\left\|\nabla_{x_j} X^x_{t,k}\right\| \|\nabla_k u(X^x_t)\|\right] \\
\le&\ \operatorname*{ess\,sup}_x \mathbb{E} \sum_{k\in[\mathsf{b}]} \|\nabla_k u(X^x_t)\| \left[\sum_{j\in[\mathsf{b}]} \mathrm{e}^{-Mt} \sum_{r=\mathsf{d}_\mathsf{G}(j,k)}^{\infty} \frac{t^r (M-m)^r}{r!}\right] \\
\le&\ \operatorname*{ess\,sup}_x \mathbb{E} \sum_{k\in[\mathsf{b}]} \|\nabla_k u(X^x_t)\| \cdot \mathsf{M}_t \\
\le&\ \mathsf{M}_t \mathbb{E}\left[\operatorname*{ess\,sup}_x \sum_{k\in[\mathsf{b}]} \|\nabla_k u(X^x_t)\|\right] \le \mathsf{M}_t \,|u|_{\mathrm{Lip},\infty}\,.
\end{aligned}$$

Thus we proved that $\|P_t\|_{|\cdot|_{\mathrm{Lip},\infty}\to|\cdot|_{\mathrm{Lip},\infty}} \le \mathsf{M}_t$.

For the corollary, we note that the operator $\mathcal{L}_\pi^{-1}$ can be expressed as

$$\mathcal{L}_\pi^{-1} = -\int_0^\infty P_t \mathrm{d}t.$$

Therefore, by the definition of M_t and Lemma 3.6, we have

$$\begin{aligned}
\left\|\mathcal{L}_\pi^{-1}\right\|_{|\cdot|_{\ell^1(\mathrm{Lip})}\to|\cdot|_{\ell^1(\mathrm{Lip})}} \le&\ \int_0^\infty \|P_t\|_{|\cdot|_{\ell^1(\mathrm{Lip})}\to|\cdot|_{\ell^1(\mathrm{Lip})}} \mathrm{d}t \le \int_0^\infty \mathsf{M}_t \mathrm{d}t \\
=&\ \int_0^\infty \left(\mathrm{e}^{-mt} + \mathsf{s}\mathrm{e}^{-Mt} \sum_{r=1}^{\infty} r^\nu \frac{t^r (M-m)^r}{r!}\right) \mathrm{d}t \\
=&\ \frac{1}{m} + \mathsf{s} \sum_{r=1}^{\infty} \frac{r^\nu (M-m)^r}{r!} \int_0^\infty \mathrm{e}^{-Mt} t^r \mathrm{d}t \\
=&\ \frac{1}{m} + \frac{\mathsf{s}}{M} \sum_{r=1}^{\infty} r^\nu \left(1-\frac{m}{M}\right)^r \\
\le&\ \frac{1}{m} + \frac{\mathsf{s}}{M} \nu! \left(\frac{m}{M}\right)^{-\nu-1} \left(1-\frac{m}{M}\right) \qquad \text{(Lemma 3.6)} \\
=&\ \frac{1}{m}\left(1 + \mathsf{s}\nu!\kappa^\nu(1-\kappa^{-1})\right) \le \frac{\mathsf{s}\nu!\kappa^\nu}{m}.
\end{aligned}$$

Here we denote $\kappa = \frac{M}{m}$. The proof is similar for the $|\cdot|_{\mathrm{Lip},\infty}$-norm. □

Lemma 3.4. *$p_k(x) := \mathrm{e}^{-x} \sum_{r=0}^\infty r^k \frac{x^r}{r!}$ is a monic polynomial of degree k.*

Proof. It can be directly verified that

$$\sum_{r=0}^{\infty} r^k \frac{x^r}{r!} = \left(x\frac{\mathrm{d}}{\mathrm{d}x}\right)^k \left(\sum_{r=0}^{\infty} \frac{x^r}{r!}\right) = \left(x\frac{\mathrm{d}}{\mathrm{d}x}\right)^k \mathrm{e}^x.$$

By definition, $\sum_{r=0}^{\infty} r^k \frac{x^r}{r!} = \mathrm{e}^x p_k(x)$, so that

$$\mathrm{e}^x p_{k+1}(x) = \left(x\frac{\mathrm{d}}{\mathrm{d}x}\right)(\mathrm{e}^x p_k(x)) = x\mathrm{e}^x \left(p_k(x) + p_k'(x)\right).$$

Therefore,

$$p_{k+1}(x) = x\left(p_k(x) + p_k'(x)\right), \quad p_0(x) = 1.$$

The conclusion follows by standard induction. □

3.4.3 Applications

δ-locality

A direct result of the above theorem is the δ-locality of Markov random field on local graph (Theorem 3.1). Note the test function $\phi(x) = \phi_i(x_i), |\phi_i|_{\mathrm{Lip}} \le 1$ satisfies

$$|\phi|_{\ell^1(\mathrm{Lip})} = \sum_{j\in[\mathsf{b}]} \|\nabla_j \phi\|_{L^\infty} = |\phi_i|_{\mathrm{Lip}} \le 1.$$

Therefore, the solution u to Stein equation $\mathcal{L}_\pi u = \phi - \mathbb{E}_\pi[\phi]$ satisfies

$$\sum_{j\in[\mathsf{b}]} \|\nabla_j u\|_{L^\infty} = |u|_{\ell^1(\mathrm{Lip})} \le \left\|\mathcal{L}_\pi^{-1}\right\|_{\ell^1(\mathrm{Lip})\to\ell^1(\mathrm{Lip})} |\phi|_{\ell^1(\mathrm{Lip})} \le \frac{\mathsf{s}\nu!\kappa^\nu}{m}.$$

Convergence of Langevin dynamics under $\mathsf{W}_{1,\infty}$

In this section, we establish the convergence of the Langevin dynamics under the $\mathsf{W}_{1,\infty}$-distance. $\mathsf{W}_{p,\infty}$-distance, introduced in [20], is the p-Wasserstein distance with ℓ_∞-norm as the underlying metric. Specifically, we define

$$\mathsf{W}_{p,\infty}(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \left(\int \|x-y\|_{2,\infty}^p \,\mathrm{d}\gamma(x,y)\right)^{1/p}. \tag{3.4.14}$$

Here $\Pi(\mu,\nu)$ denotes the set of couplings between μ and ν. Note we use the norm

$$\|x-y\|_{2,\infty} := \max_{j\in[\mathsf{b}]} \|x_j - y_j\|, \tag{3.4.15}$$

as the base metric instead of $\|\cdot\|_\infty$ in the original definition in [20]. This is for consistency with the block decomposition $x = (x_1, \ldots, x_{\mathsf{b}})$. One can simply take $d_j \equiv 1$ to go back to the original definition.

Note $\mathsf{W}_{p,\infty}$ is stronger than the marginal W_p distance, i.e.

$$\max_{j\in[\mathsf{b}]} \mathsf{W}_p(\mu_j, \nu_j) \le \mathsf{W}_{p,\infty}(\mu, \nu).$$

Simply note that

$$\forall j \in [\mathsf{b}], \quad \mathsf{W}_{p,\infty}(\mu,\nu) \ge \inf_{\gamma\in\Pi(\mu,\nu)} \left(\int \|x_j - y_j\|^p \,\mathrm{d}\gamma(x,y) \right)^{1/p} = \mathsf{W}_p(\mu_j, \nu_j).$$

We focus on the case $p = 1$, which admits a duality representation.

Lemma 3.5. *For $\mu, \nu \in \mathcal{P}_1(\mathbb{R}^d)$, it holds that*

$$\mathsf{W}_{1,\infty}(\mu,\nu) = \sup_{|\phi|_{\mathrm{Lip},\infty}\le 1} \left[\mathbb{E}_\mu[\phi] - \mathbb{E}_\nu[\phi]\right], \tag{3.4.16}$$

where we denote the $|\cdot|_{\mathrm{Lip},\infty}$-seminorm

$$|\phi|_{\mathrm{Lip},\infty} := \sup_{x\neq y} \frac{|\phi(x) - \phi(y)|}{\|x - y\|_{2,\infty}} = \operatorname*{ess\,sup}_x \sum_{j\in[\mathsf{b}]} \|\nabla_j \phi(x)\|. \tag{3.4.17}$$

Proof. (3.4.16) directly follows the Kantorovich-Rubinstein duality [113]. For the second equality, first note $\forall x, y \in \mathbb{R}^d$,

$$\begin{aligned}
\phi(x) - \phi(y) =& \int_0^1 \sum_{j\in[\mathsf{b}]} \nabla_j \phi((1-t)x + ty) \cdot (x_j - y_j) \mathrm{d}t \\
\le& \|x - y\|_{2,\infty} \int_0^1 \sum_{j\in[\mathsf{b}]} \|\nabla_j \phi((1-t)x + ty)\| \,\mathrm{d}t \\
\le& \|x - y\|_{2,\infty} \operatorname*{ess\,sup}_x \sum_{j\in[\mathsf{b}]} \|\nabla_j \phi(x)\|.
\end{aligned}$$

This implies that $|\phi|_{\mathrm{Lip},\infty} \le \operatorname{ess\,sup}_x \sum_{j\in[\mathsf{b}]} \|\nabla_j \phi(x)\|$. Conversely, for any $x \in \mathbb{R}^d$, denote $t \in \mathbb{R}^d$ s.t. $t_j = \nabla_j \phi(x) / \|\nabla_j \phi(x)\|$. Note $\|t\|_{2,\infty} = 1$, and we have for sufficiently small $h > 0$,

$$\phi(x + ht) - \phi(x) = \sum_{j=1}^d h t_j^{\mathrm{T}} \nabla_j \phi(x) + o(h) = h \sum_{j=1}^d \|\nabla_j \phi(x)\| + o(h).$$

$$\Rightarrow \ |\phi|_{\mathrm{Lip},\infty} \geq \frac{\phi(x+ht)-\phi(x)}{h} = \sum_{j=1}^{d} \|\nabla_j \phi(x)\| + o(1).$$

Since x and h are arbitrary, we have $|\phi|_{\mathrm{Lip},\infty} \geq \operatorname{ess\,sup}_x \sum_{j\in[\mathsf{b}]} \|\nabla_j \phi(x)\|$. □

Next we state the convergence theorem.

Theorem 3.9. *Let* G *be a* (s,ν)*-local graph. Suppose* $\pi \in \mathcal{P}(\mathbb{R}^d)$ *is localized on* G*, and satisfies for some* $0 < m \leq M < \infty$*,*

$$\forall x \in \mathbb{R}^d, \quad mI \preceq -\nabla^2 \log \pi(x) \preceq MI.$$

Consider the Langevin dynamics for π *with initial distribution* μ_0*, and let* μ_t *be the distribution of* X_t^x*. Then for all* $t \geq 0$*,*

$$\mathsf{W}_{1,\infty}(\mu_t,\pi) \leq \mathsf{M}_t \mathsf{W}_{1,\infty}(\mu_0,\pi), \tag{3.4.18}$$

where M_t *is defined in* (3.4.11)*.*

Proof. Let $\phi \in H_0^1(\pi)$ be any test function s.t. $|\phi|_{\mathrm{Lip},\infty} \leq 1$. By Theorem 3.8,

$$|P_t\phi|_{\mathrm{Lip},\infty} \leq \mathsf{M}_t \, |\phi|_{\mathrm{Lip},\infty} \leq \mathsf{M}_t.$$

Therefore, by Lemma 3.5, we have

$$\begin{aligned}
\mathsf{W}_{1,\infty}(\mu_t,\pi) &= \sup_{|\phi|_{\mathrm{Lip},\infty}\leq 1} [\mathbb{E}_{\mu_t}[\phi] - \mathbb{E}_\pi[\phi]] \\
&= \sup_{|\phi|_{\mathrm{Lip},\infty}\leq 1} [\mathbb{E}_{\mu_0}[P_t\phi] - \mathbb{E}_\pi[P_t\phi]] \\
&\leq \mathsf{M}_t \sup_{|\psi|_{\mathrm{Lip},\infty}\leq 1} [\mathbb{E}_{\mu_0}[\psi] - \mathbb{E}_\pi[\psi]] = \mathsf{M}_t \mathsf{W}_{1,\infty}(\mu_0,\pi).
\end{aligned}$$

This completes the proof. □

Remark 3.6. Since under $\mathsf{W}_{1,\infty}$-norm, the Langevin dynamics is not contractive in the usual sense, one cannot expect one-step coupling to work without further conditions. This is the main reason for the multistep coupling used in [20]. In fact, [20] essentially uses a discrete version of Theorem 3.9 by taking t larger so that $\mathsf{M}_t < 1$.

3.5 Proofs

3.5.1 Proof of Theorem 3.4

Proof of Theorem 3.4. The proof is based on that of Theorem 3.1 and Theorem 3.3. We only present the different parts for the multiple block case.

For any index set $I \subseteq [\mathsf{b}]$, denote $d_I = \sum_{i \in I} d_i$. Then by definition,

$$\mathsf{W}_1(\pi_I, \pi_I') = \sup_{\phi_I \in \mathrm{Lip}_1(\mathbb{R}^{d_I})} \int \phi_I(x_I) \left(\pi_I(x_I) - \pi_I'(x_I)\right) \mathrm{d}x_I.$$

Given ϕ_I, let $u(x)$ solve the marginal Stein equation

$$\mathcal{L}_{\pi'} u(x) = \phi_I(x_I) - \mathbb{E}_{\pi'}[\phi_I(x_I)].$$

Similarly as in Lemma 3.1, it holds that

$$\|\nabla_j u(x)\| \le \int_0^\infty \mathbb{E}\left[\|\nabla_j X_{t,I}^x\| \left\|\nabla \phi_I(X_{t,I}^x)\right\|\right] \mathrm{d}t \le \int_0^\infty \mathbb{E}\|\nabla_j X_{t,I}^x\| \mathrm{d}t,$$

where X_t^x is the solution to the Langevin dynamics for π' with initial condition x. Notice $\|\nabla_j X_{t,I}^x\| \le \sum_{i \in I} \|\nabla_j X_{t,i}^x\|$, we obtain that

$$\|\nabla u\|_{\infty,1} = \sum_{j \in [\mathsf{b}]} \|\nabla_j u(x)\| \le \sum_{j \in [\mathsf{b}]} \mathbb{E} \int_0^\infty \sum_{i \in I} \|\nabla_j X_{t,i}^x\| \mathrm{d}t \le \sum_{i \in I} \delta = \delta |I|.$$

Using the same arguments in Theorem 3.3, we obtain

$$\begin{aligned} \mathsf{W}_1(\pi_I, \pi_I') &\le \|\nabla u\|_{\infty,1} \cdot \max_j \|\nabla_j \log \pi' - \nabla_j \log \pi\|_{L^1(\pi)} \\ &\le \delta |I| \cdot \max_j \|\nabla_j \log \pi' - \nabla_j \log \pi\|_{L^1(\pi)}. \end{aligned}$$

This completes the proof. □

3.5.2 Lemmas

Lemma 3.6. *For any $t \ge 0$ and $x \in (0,1)$, it holds that*

$$\sum_{k \ge 1} k^t (1-x)^k < 2\Gamma(t+1) x^{-t-1}(1-x).$$

When $t \in \mathbb{N}$, the factor 2 can be omitted, i.e. $\sum_{k \ge 1} k^t (1-x)^k \le t! x^{-t-1}(1-x)$.

Proof. We prove by induction. For $t = 0$, it holds that

$$\sum_{k\geq 1}(1-x)^k = \frac{1-x}{x}.$$

For $t \in (0,1)$, first notice by Abel transformation,

$$x\sum_{k\geq 1} k^t(1-x)^k = \sum_{k\geq 1} k^t\left[(1-x)^k - (1-x)^{k+1}\right] = \sum_{k\geq 1}\left(k^t - (k-1)^t\right)(1-x)^k.$$

Therefore, since $k^t - (k-1)^t \leq t(k-1)^{t-1}$ when $t \in (0,1)$ and $k \geq 2$,

$$\begin{aligned}
\sum_{k\geq 1} k^t(1-x)^k &\leq x^{-1}\left[(1-x) + \sum_{k\geq 2} t(k-1)^{t-1}(1-x)^k\right] \\
&\leq x^{-1}(1-x)\left[1 + t\sum_{k\geq 2}(k-1)^{t-1}\mathrm{e}^{-(k-1)x}\right] \\
&\leq x^{-1}(1-x)\left[1 + t\int_0^\infty y^{t-1}\mathrm{e}^{-yx}\mathrm{d}y\right] \\
&= x^{-1}(1-x)\left(1 + t\Gamma(t)x^{-t}\right) \\
&< 2\Gamma(t+1)x^{-t-1}(1-x).
\end{aligned}$$

Here we use $t\Gamma(t) = \Gamma(t+1)$ and $\Gamma(t+1)x^{-t} > 1$ in the last step. This verifies the case $t \in [0,1)$. Suppose the inequality holds for $t - 1 \geq 0$, then using the same methods,

$$\begin{aligned}
\sum_{k\geq 1} k^t(1-x)^k &= x^{-1}\sum_{k\geq 1}\left(k^t - (k-1)^t\right)(1-x)^k \\
&\leq x^{-1}\sum_{k\geq 1} tk^{t-1}(1-x)^k \\
&< x^{-1}t\cdot 2\Gamma(t)x^{-(t-1)-1}(1-x) \\
&= 2\Gamma(t+1)x^{-t-1}(1-x).
\end{aligned}$$

Here the first inequality follows from the elementary inequality $k^t - (k-1)^t \leq tk^{t-1}$ when $t \geq 1$, and the second inequality follows from induction hypothesis. The refined inequality for $t \in \mathbb{N}$ can be obtained similarly. This completes the proof. □

Lemma 3.7. *Suppose $G_t \in \mathbb{R}_{\geq 0}^{\mathsf{b}\times\mathsf{b}}$ is a time-dependent nonnegative matrix satisfying*

$$\frac{\mathrm{d}}{\mathrm{d}t}G_t \leq -MG_t,$$

where $\leq$ is in the entrywise sense, and $M \in \mathbb{R}^{\mathsf{b}\times\mathsf{b}}$ is c-diagonal dominant with non-positive off-diagonal entries, i.e. $\forall i, j \in [\mathsf{b}], i \neq j$,

$$\sum_{j:j\neq i} |M_{ij}| + c \leq M_{ii}; \quad M_{ij} \leq 0.$$

Then for any $t \geq 0$, it holds

$$\|G_t\|_\infty \leq \mathrm{e}^{-ct}\|G_0\|_\infty.$$

Proof. Denote $G_t^c := \mathrm{e}^{ct}G_t$, then

$$\frac{\mathrm{d}}{\mathrm{d}t}G_t^c = \mathrm{e}^{ct}\left(\frac{\mathrm{d}}{\mathrm{d}t}G_t + cG_t\right) \leq \mathrm{e}^{ct}\left(-MG_t + cG_t\right) = \left(-M + cI\right)G_t^c.$$

Multiple both sides by $\mathbf{1} = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^{\mathsf{b}}$ from right,

$$\frac{\mathrm{d}}{\mathrm{d}t}G_t^c\mathbf{1} \leq \left(-M + cI\right)G_t^c\mathbf{1}.$$

This operator preserves the inequality since it is equivalent to taking summation over row indices. We claim that

$$\forall t \geq 0, \quad \|G_t^c\mathbf{1}\|_\infty \leq \|G_0^c\mathbf{1}\|_\infty, \tag{3.5.1}$$

which is a reformulation of $\|G_t\|_\infty \leq \mathrm{e}^{-ct}\|G_0\|_\infty$, since $G_t^c \in \mathbb{R}^{\mathsf{b}}_{\geq 0}$ and

$$\mathrm{e}^{ct}\|G_t\|_\infty = \|G_t^c\|_\infty = \max_i \sum_j G_t^c(i,j) = \|G_t^c\mathbf{1}\|_\infty.$$

We prove (3.5.1) by contraction. Suppose (3.5.1) is false, then $\exists s \geq 0$ and i s.t.

$$\frac{\mathrm{d}}{\mathrm{d}t}(G_s^c\mathbf{1})_i > 0, \quad (G_s^c\mathbf{1})_i = \|G_s^c\mathbf{1}\|_\infty.$$

On the other hand, notice by assumption on M,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(G_s^c\mathbf{1})_i \leq\ & (-M_{ii}+c)\,(G_s^c\mathbf{1})_i + \sum_{j:j\neq i}(-M_{ij})(G_s^c\mathbf{1})_j \\ \leq\ & (-M_{ii}+c)\,(G_s^c\mathbf{1})_i + \sum_{j:j\neq i}(-M_{ij})(G_s^c\mathbf{1})_i \\ =\ & \left(-M_{ii}+c+\sum_{j:j\neq i}|M_{ij}|\right)(G_s^c\mathbf{1})_i \leq 0.\end{aligned}$$

Contradiction. This proves our result. □

Chapter 4

Localization Method in Sampling

The previous chapters discuss the theoretical properties of the locality structure. In essence, the locality structure induces a form of low-dimensionality, so that it is natural to study it algorithmically for high dimensional sampling problems. In this chapter, we will review some of the existing methods and discuss the general idea of localized sampling. We propose a framework to localize existing samplers, which turns a high-dimensional problem into many low-dimensional subproblems. We will then discuss its computational advantages, i.e. localized and parallelizable implementation; and its theoretical advantages, i.e. lower statistical complexity and controllable localization error. Specific examples of localized sampling will be discussed in the next two chapters.

4.1 Review on existing localized samplers

In this section, we review two existing methods for localized sampling, including the localized versions of SVGD [77, 115, 121] and Schrödinger bridge sampler [55, 57]. Detailed discussions on the localized MALA [95, 94, 81, 110] will be introduced in Chapter 5.

4.1.1 Message passing Stein variational gradient descent

Stein variational gradient descent (SVGD) [77] is a particle method to sample from a target distribution via a gradient flow [76, 22]. It uses an ensemble of particles $\{x^{(i)}\}_{i=1}^N$ to approximate the target distribution π, and evolves the particles

according to the following dynamics

$$\frac{\mathrm{d}}{\mathrm{d}t}x^{(l)} = v(x^{(l)}) := \frac{1}{N}\sum_{i=1}^{N}\left[\nabla \log \pi(x^{(i)})k(x^{(i)}, x^{(l)}) + \nabla_1 k(x^{(i)}, x^{(l)})\right],$$

where $k(x, y)$ is a positive definite kernel function. The mean field limit of the velocity field is (denote μ as the mean field measure of the particles)

$$v = \mathbb{E}_\mu\left[\nabla \log \pi(x)k(x, \cdot) + \nabla_1 k(x, \cdot)\right] = \mathbb{E}_\mu\left[k(x, \cdot)\nabla \log \frac{\pi(x)}{\mu(x)}\right].$$

Notice $-\nabla \log \frac{\pi}{\mu}$ is the first variation of the KL divergence, i.e.

$$\nabla \log \frac{\pi}{\mu} = -\frac{\delta}{\delta \mu}\mathsf{KL}\left(\mu \| \pi\right).$$

So that SVGD can be viewed as a gradient flow of $\mathsf{KL}\left(\cdot\|\pi\right)$ w.r.t. some kernelized metric (see [22]). Note also v admits a variational form [76]

$$v = \underset{\|\boldsymbol{\phi}\|_{\mathcal{H}}\leq 1}{\arg\max}\left[-\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathsf{KL}\left((\mathrm{id} + \varepsilon\boldsymbol{\phi})_{\#}\mu \| \pi\right)\right].$$

Here $\|\boldsymbol{\phi}\|_{\mathcal{H}}$ denotes the RKHS norm of the vector field $\boldsymbol{\phi}$. Here $\mathcal{H} = \mathcal{H}_0^{\otimes d}$ is the corresponding $d \times 1$ vector-valued RKHS, and $\mathcal{H}_0$ is the RKHS of the kernel k.

The use of a kernel k makes SVGD difficult to scale to high dimensions. For instance, the widely used Gaussian kernel $k(x, y) = \mathsf{N}(x - y; 0, \sigma^2 I)$ decays exponentially in the distance $\|x - y\|$, which is of $\mathcal{O}(\sqrt{d})$ in high dimensions. This makes the computation very sensitive to the error and the hyperparameters. It is also reported in [121] that SVGD tends to underestimate the marginal variance when dimensions are high.

To address the dimension issue, [115, 121] concurrently proposed a localized version of SVGD, which uses localized kernels inspired by the locality structure. [121] propose to consider a coordinate-wise velocity field that solves

$$\begin{aligned} v_j &= \underset{\|\phi_j\|_{\mathcal{H}_j}\leq 1}{\arg\max}\left[-\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathsf{KL}\left((\mathrm{id} + \varepsilon\boldsymbol{\phi}_j)_{\#}\mu \| \pi\right)\right], \\ \text{where} \quad \boldsymbol{\phi}_j(x) &= (0, \cdots, 0, \phi_j(x_j \mid x_{-j}), 0, \cdots, 0). \end{aligned} \tag{4.1.1}$$

Here $\mathcal{H}_j$ is some localized RKHS to be determined later. Since $\boldsymbol{\phi}_j$ is non-zero only

at the j-th component, it holds that

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathsf{KL}\left((\mathrm{id}+\varepsilon\boldsymbol{\phi}_j)_{\#}\mu\|\pi\right)=\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\mathsf{KL}\left((\mathrm{id}+\varepsilon\phi_j(\cdot\mid x_{-j}))_{\#}\mu(x_j\mid x_{-j})\|\pi(x_j\mid x_{-j})\right).$$

If π is localized, $\pi(x_j\mid x_{-j})=\pi(x_j\mid x_{\mathcal{N}_j})$; and as an approximation of π, μ should approximately satisfy the same property. So that the quantities to be optimized in (4.1.1) are approximately independent of $x_{-\mathcal{N}_j}$. Therefore, one can enforce ϕ_j to be independent of $x_{-\mathcal{N}_j}$. Take the *local kernel*

$$k_j=k_j(x_{\mathcal{N}_j},y_{\mathcal{N}_j}).$$

Consider the vector-valued RKHS $\mathcal{H}_j=\mathcal{H}_{j,0}^{\otimes d_j}$, where $\mathcal{H}_{j,0}$ is the RKHS of the kernel k_j. With this choice, the optimal v_j^* of (4.1.1) is

$$\begin{aligned}v_j^*(x_{\mathcal{N}_j})&=\mathbb{E}_{y\sim\mu}\left[k_j(y_{\mathcal{N}_j},x_{\mathcal{N}_j})\nabla_{y_j}\log\frac{\pi(y_j|y_{\mathcal{N}_j^-})}{\mu(y_j|y_{\mathcal{N}_j^-})}\right]\\&=\mathbb{E}_{y\sim\mu}\left[\nabla_{y_j}\log\pi(y_j|y_{\mathcal{N}_j^-})k_j(y_{\mathcal{N}_j},x_{\mathcal{N}_j})+\nabla_1 k_j(y_{\mathcal{N}_j},x_{\mathcal{N}_j})\right].\end{aligned}$$

The resulted localized SVGD is: $\forall j\in[\mathsf{b}]$ and $l\in[N]$,

$$\frac{\mathrm{d}}{\mathrm{d}t}x_j^{(l)}=\frac{1}{N}\sum_{i=1}^{N}\left[\nabla_{x_j^{(i)}}\log\pi(x_j^{(i)}|x_{\mathcal{N}_j^-}^{(i)})k_j(x_{\mathcal{N}_j}^{(i)},x_{\mathcal{N}_j}^{(l)})+\nabla_1 k_j(x_{\mathcal{N}_j}^{(i)},x_{\mathcal{N}_j}^{(l)})\right].\tag{4.1.2}$$

It is also called *graphical SVGD* in [115]. Note that the localized SVGD (4.1.2) can be regarded as a collection of b SVGD velocity fields (with only j-th component used) for the marginal distributions $\pi(x_{\mathcal{N}_j})$. This is a typical way of localizing samplers, that is, taking the transition kernel for j-th component as the j-th component of the original transition kernel for the marginal distribution $\pi(x_{\mathcal{N}_j})$.

It is almost obvious that the localized SVGD (4.1.2) no longer suffers from the dimension issues, since the localized flow is low dimensional for each j. All the computations are also localized and thus parallelizable. Numerical experiments in [115, 121] validate the effectiveness of the localized SVGD compared to the vanilla SVGD in high dimensions.

4.1.2 Localized Schrödinger bridge sampler

Schrödinger bridge (SB) sampler [55] aims to learn a transition kernel $P_\epsilon(x, y)$ that is invariant under the target distribution π, given data $\{X^{(i)}\}_{i=1}^N$ sampled from π. The learned transition kernel $P_\epsilon(x, y)$ is then used to sample from π.

[55] proposes to construct P_ϵ by approximating the SB solution Q_ϵ w.r.t. π and a reference transition kernel T_ϵ. Specifically, [55] takes T_ϵ as the Gaussian kernel $T_\epsilon(x, y) = \mathsf{N}(y; x, 2\epsilon I)$, and the according SB problem [88, 21] is to find a transition kernel $Q_\epsilon(x, y)$ s.t.

$$Q_\epsilon = \underset{Q:\pi Q=\pi}{\arg\min}\, \mathsf{KL}\left(\pi(x)Q(x, y)\|\pi(x)T_\epsilon(x, y)\right).$$

The solution is of the form

$$Q_\epsilon(x, y) = \varphi_\epsilon(x)T_\epsilon(x, y)\psi_\epsilon(y)\pi(y),$$

and $\varphi_\epsilon, \psi_\epsilon$ can be solved by the Sinkhorn algorithm [36, 55]. Note $\varphi_\epsilon = \psi_\epsilon$ since T_ϵ is symmetric. Finally, P_ϵ can be taken as the Gaussian approximation of Q_ϵ, i.e.

$$P_\epsilon^{\mathsf{SB}}(x, y) = T_\epsilon(m_\epsilon(x), y), \quad m_\epsilon(x) = \mathbb{E}_{\pi\otimes Q_\epsilon}\left[y|x\right].$$

Once ψ_ϵ is obtained, m_ϵ can be explicitly computed by

$$m_\epsilon(x) = \int yQ_\epsilon(x, y)\mathrm{d}y = \mathbb{E}_{y\sim\pi}[w_\epsilon(x, y)y],$$

$$w_\epsilon(x, y) = \frac{\mathrm{d}Q_\epsilon(x, \cdot)}{\mathrm{d}\pi}(y) = \varphi_\epsilon(x)T_\epsilon(x, y)\psi_\epsilon(y) = \frac{T_\epsilon(x, y)\psi_\epsilon(y)}{\int T_\epsilon(x, y)\psi_\epsilon(y)\pi(y)\mathrm{d}y}.$$

The discretized version can be easily obtained for numerical implementation. Note when ϵ is small, P_ϵ^{SB} will be a good approximation of Q_ϵ, and thus can be approximately used as a MCMC kernel to sample π. Also note $\{\psi_\epsilon(X^{(i)})\}_{i=1}^N$ can be computed offline using the Sinkhorn algorithm, and in sampling, only $\omega_\epsilon(x, X^{(i)})$ needs to be computed, which is relatively cheaper. We mention that P_ϵ^{SB} can be regarded as an approximation of $\exp(\epsilon\mathcal{L}_\pi)$ (see (3.1.1)), which is more stable compared to the Euler-Maruyama discretization of the Langevin dynamics [55].

However, learning a generic high dimensional transition kernel P_ϵ faces the curse of dimensionality. To reduce the sample complexity, [56] proposes the localized SB

(LSB) sampler by using the localized transition kernel:

$$P_\epsilon^{\mathsf{LSB}}(x, y) = \prod_{s\in[d]} \mathsf{N}(y_s; m_{\epsilon,s}(x_{\mathcal{N}_s}), 2\epsilon I), \tag{4.1.3}$$

where $\mathcal{N}_s$ denotes the neighbors of s, and $m_{\epsilon,s}$ is obtained by

$$m_{\epsilon,s}(x_{\mathcal{N}_s}) = \int y_s Q_{\epsilon,s}(x_{\mathcal{N}_s}, y_{\mathcal{N}_s}) \mathrm{d}y_{\mathcal{N}_s}.$$

Here $Q_{\epsilon,s}$ solves the SB problem for the marginal distribution $\pi(x_{\mathcal{N}_s})$ with

$$T_{\epsilon,s}(x_{\mathcal{N}_s}, y_{\mathcal{N}_s}) = \mathsf{N}(y_{\mathcal{N}_s}; x_{\mathcal{N}_s}, 2\epsilon I).$$

Detailed derivation can be found in [56].

As in the localized SVGD, the LSB learns a collection of d SB samplers for the marginal distributions $\pi(x_{\mathcal{N}_s})$, and use the learned transition kernels to sample in a Gibbs way. Since the LSB turns a high-dimensional kernel learning problem into d low-dimensional ones, the sample complexity is significantly reduced.

4.2 Framework for designing localized samplers

We summarize the localized sampling methods as a general framework. Formally, consider a classical sampler in the form of a transition kernel

$$x \sim \mathsf{P}(x_0, x), \quad x_0 \sim \pi_0.$$

Here π_0 is the initial distribution or the prior distribution, which is easier to be sampled from, and $\mathsf{P}(x_0, x)$ is the sampling kernel that transforms π_0 to the target distribution π. The kernel P is usually a progressive Markov kernel, i.e.

$$\mathsf{P} = \mathsf{P}_1 \circ \mathsf{P}_2 \circ \cdots \circ \mathsf{P}_T.$$

For instance, for unadjusted Langevin algorithm (ULA), the transition kernel is

$$\mathsf{P}_t(x, y) = \mathsf{N}(y; x + \tau\nabla \log \pi(x), 2\tau I), \quad \forall 1 \le t \le T.$$

Due to the locality structure, the transition kernel of many samplers can be written or approximated as

$$\mathsf{P}_t^{\mathrm{loc}}(x, y) = \prod_{i \in [\mathsf{b}]} \mathsf{P}_{t,i}(x_i, y_i \mid x_{\mathcal{N}_i^r}). \tag{4.2.1}$$

Here we use the block decomposition form, $r \in \mathbb{Z}_+$ is the localization radius, and $\mathcal{N}_i^r$ is the r-neighborhood (2.1.2). The simplest way to construct (4.2.1) is to take $\mathsf{P}_{t,i}(x_i, y_i \mid x_{\mathcal{N}_i^r})$ as the conditional transition kernel for the i-th component of the full marginal transition kernel $\mathsf{P}_{t,i}(x_{\mathcal{N}_i^r}, y_{\mathcal{N}_i^r})$ obtained by the original sampler on the marginal distribution $\pi(x_{\mathcal{N}_i^r})$. The localized SVGD (with time discretization) and the localized SB sampler can both be fitted into this framework. It can be observed directly from (4.1.2) and (4.1.3). And we will see in the following two chapters that the framework also applies to other localized samplers.

The advantanges of the localized sampler (4.2.1) include:

- **Localized and parallelizable**: to compute $\mathsf{P}_{t,i}$ in the localized sampler, only local information of x is required (i.e. $x_{\mathcal{N}_i^r}$). This reduces the computational complexity of the sampler, and also allows parallel implementation. Examples of local and parallel implementation can be found in Section 5.3.4.

- **Lower statistical complexity**: the transition kernel is usually learned from the data, and since the localized sampler is a collection of b low-dimensional samplers, the sample complexity is significantly reduced compared to the original sampler. Examples of the statistical analysis of the localized sampler can be found in Section 6.2.3.

However, the localized sampler will introduce localization error in the sampled distribution. The localization error can be controlled using the marginal Stein's method. Due to the exponential correlation decay in the locality structure, the localization error typically decays exponentially in the localization radius r. An example of the localization error analysis can be found in Section 6.2.2. The exponential decay can also be numerically observed and validated.

We also comment that in practice, we can tune the localization radius r in the localized sampler to balance the localization error and the statistical error, similar to the bias-variance trade-off. We will show both theoretically and numerically that an appropriate localization radius r can indeed reduce the overall error in the sampled distribution, see Section 6.2.3 and Section 6.3.

Chapter 5

Localized Metropolis-adjusted Langevin Algorithm

In this chapter, we discuss the MALA-within-Gibbs (MLwG) sampler, which is the localized Metropolis-adjusted Langevin algorithm (MALA). The vanilla MALA, as a typical Markov chain Monte Carlo (MCMC) method, is slow in high dimensional problems; the step length of MALA should be $\tau = \mathcal{O}(d^{-1/3})$ for d dimensional problems to obtain a non-degenerate acceptance rate. MLwG aims to mitigate the dimension problem by exploiting the locality structure. By using MALA step within the Gibbs sampler, the step length of MLwG can be chosen independently of the dimension d. The acceptance rate and the convergence rate are also guaranteed to be dimension independent. Through an image deblurring problem, we show that MLwG can be implemented in a local and parallel way. A dimension-free approximation result is also discussed.

5.1 MALA-within-Gibbs

5.1.1 Metropolis-adjusted Langevin algorithm

Metropolis-adjusted Langevin algorithm (MALA) [95, 94] samples a target distribution $\pi \in \mathcal{P}(\mathbb{R}^d)$ by using Langevin dynamics corrected by a Metropolis-Hastings step. It consists of two steps:

- Langevin step: draw a proposal from the Langevin dynamics, i.e.,

$$z = x + \tau \nabla \log \pi(x) + \sqrt{2\tau}\xi, \quad \xi \sim \mathcal{N}(0, I),$$

- Metropolis step: accept the proposal z with probability

$$\alpha(x,z) = \min\left\{1, \frac{\pi(z)Q(z,x)}{\pi(x)Q(x,z)}\right\},$$

where we denote the proposal kernel as

$$Q(x,z) = \mathsf{N}(z; x + \tau\nabla \log \pi(x), 2\tau I). \tag{5.1.1}$$

The algorithm is summarized in Algorithm 1.

Algorithm 1 MALA sampler

Input: Initial state $x^0 \in \mathbb{R}^d$, step size $\tau > 0$, number of iterations T.

for $n = 0$ to $T-1$ **do**

Draw proposal

$$z^n = x^n + \tau\nabla \log \pi(x^n) + \sqrt{2\tau}\xi^n, \quad \xi^n \sim \mathcal{N}(0, I).$$

Compute acceptance probability (cf. (5.1.1))

$$\alpha(x^n, z^n) = \min\left\{1, \frac{\pi(z^n)Q(z^n,x^n)}{\pi(x^n)Q(x^n,z^n)}\right\}.$$

Draw a uniform random variable $\zeta^n \sim \mathcal{U}[0,1]$.

if $\zeta^n < \alpha(x^n, z^n)$ **then**

Accept the proposal: $x^{n+1} = z^n$.

else

Reject the proposal: $x^{n+1} = x^n$.

end if

end for

Output: Sampled chain $\{x^n\}_{n=0}^T$.

The transition kernel of MALA is given by

$$\mathsf{P}_{\mathrm{MALA}}(x,z) = \alpha(x,z)Q(x,z) + \delta_x(z)\int (1-\alpha(x,z))Q(x,z)\mathrm{d}z. \tag{5.1.2}$$

It can be directly verified that π is the stationary distribution of $\mathsf{P}_{\mathrm{MALA}}$. Under strong log-concavity assumption, it is established in [95] that MALA converges exponentially fast to the target distribution π.

To obtain a non-degenerate acceptance rate in high dimensional problems, it is known [94, 89] that the step length τ should scale as $\tau = \mathcal{O}(d^{-1/3})$. When the dimension d is large, the step length τ becomes small, resulting in a slow

convergence rate.

5.1.2 MALA-within-Gibbs

MALA-within-Gibbs (MLwG) is a localized version of MALA, which is a Gibbs sampler with blockwise MALA update. In Gibbs sampling, one draws samples in a blockwise manner from the conditional distributions of the target distribution. To be specific, one first determine the block index j in a deterministic or stochastic way, and then draw a sample x' s.t.

$$x'_{-j} = x_{-j}, \quad x'_j \sim P_j(x_j, x'_j \mid x_{-j}),$$

and $P_j(x_j, x'_j \mid x_{-j})$ is a transition kernel that is invariant under the conditional distribution $\pi_j(x_j \mid x_{-j})$. It is straightforward to verify that π is the stationary distribution of the Gibbs sampler [49].

In MLwG, P_j is specified as the (block) MALA transition kernel, i.e., one first draw a proposal z s.t.

$$z_{-j} = x_{-j}, \quad z_j = x_j + \tau \nabla_j \log \pi(x) + \sqrt{2\tau}\xi_j, \quad \xi_j \sim \mathcal{N}(0, I_{d_j}).$$

Then, one accepts the proposal with probability

$$\alpha_j(x, z) = \min\left\{1, \frac{\pi(z)Q_j(z_j, x_j \mid x_{-j})}{\pi(x)Q_j(x_j, z_j \mid x_{-j})}\right\},$$

where $Q_j(x_j, z_j \mid x_{-j})$ is the proposal kernel for the j-th block

$$Q_j(x_j, z_j \mid x_{-j}) = \mathsf{N}(z_j; x_j + \tau \nabla_j \log \pi(x_j, x_{-j}), 2\tau I_{d_j}). \tag{5.1.3}$$

We now establish some notation for MLwG. For simplicity, we consider the case where the blocks are updated sequentially. Other block updating rules like randomized sequences could also be employed, but are not discussed here. We call a complete iteration in which all blocks are updated a *cycle* and denote by $x^{n,j} \in \mathbb{R}^d$ the state *during* the n-th cycle *before* the update of the j-th block. To illustrate this notation, consider the following presentation of block updates:

$$x^0 = \underbrace{x^{1,1} \to x^{1,2} \to \cdots \to x^{1,\mathsf{b}+1}}_{\text{1st cycle}} = x^1 = \underbrace{x^{2,1} \to x^{2,2} \to \cdots \to x^{2,\mathsf{b}+1}}_{\text{2nd cycle}} = x^2 \to \cdots$$

Notice also that we introduce x^n to denote the state *after* the n-th cycle.

The algorithm is summarized in Algorithm 2.

Algorithm 2 MLwG sampler

Input: Initial state $x^0 \in \mathbb{R}^d$, step size $\tau > 0$, number of iterations N.

Set $x^{1,0} = x^0$.
for $n = 1$ to N **do**
  for $j = 1$ to b **do**
    Draw proposal $z^{n,j}$ s.t. $z^{n,j}_{-j} = x^{n,j}_{-j}$ and

$$z^{n,j}_j = x^{n,j}_j + \tau \nabla_j \log \pi(x^{n,j}) + \sqrt{2\tau}\xi^{n,j}_j, \quad \xi^{n,j}_j \sim \mathcal{N}(0, I_{d_j}).$$

    Compute acceptance probability (cf. (5.1.3))

$$\alpha(x^{n,j}, z^{n,j}) = \min\left\{1, \frac{\pi(z^{n,j})Q_j(z^{n,j}_j, x^{n,j}_j \mid x^{n,j}_{-j})}{\pi(x^{n,j})Q_j(x^{n,j}_j, z^{n,j}_j \mid x^{n,j}_{-j})}\right\}.$$

    Draw a uniform random variable $\zeta^{n,j} \sim \mathcal{U}[0,1]$.
    if $\zeta^{n,j} < \alpha(x^{n,j}, z^{n,j})$ **then**
      Accept the proposal: $x^{n,j+1} = z^{n,j}$.
    else
      Reject the proposal: $x^{n,j+1} = x^{n,j}$.
    end if
  end for
  Set $x^{n+1,0} = x^n = x^{n,\mathsf{b}+1}$.
end for

Output: Sampled chain $\{x^n\}_{n=0}^N$.

Discussions on the implementation strategies, e.g. the choice of step length τ, local and parallel computation in MLwG will be given in Section 5.3.

5.2 Dimensional-free properties

In this section, we show that if the target distribution is localized, the acceptance rate and the convergence rate of MLwG are independent of the total dimension d. This allows the step length τ to be chosen independently of d, which is a significant improvement over the vanilla MALA ($\tau = \mathcal{O}(d^{-1/3})$).

Before introducing the main results, we denote for brevity

$$v_j(x) := \nabla_j \log \pi(x). \tag{5.2.1}$$

5.2.1 Acceptance rate

Theorem 5.1. *Suppose $\pi \in \mathcal{P}(\mathbb{R}^d)$ is localized on a (s, ν)-local graph G. Assume*

$$\forall j \in [\mathsf{b}], \quad \|v_j\|_{L^\infty} \le \mathsf{M}, \quad \|\nabla v_j\|_{L^\infty} \le \mathsf{H}, \quad |\nabla v_j|_{\mathrm{Lip}} \le \mathsf{L}.$$

Then there exists some $M > 0$ depending only on $\mathsf{M}, \mathsf{H}, \mathsf{L}$ and s, s.t.

$$\mathbb{E}_{\xi_j^{n,j}} \left[\alpha(x^{n,j}, z^{n,j})\right] \ge 1 - M\tau^{3/2}. \tag{5.2.2}$$

Remark 5.1. (1) Due to the locality structure, $\mathsf{M}, \mathsf{H}, \mathsf{L}$ are typically dimensional independent. Since $v_j = \nabla_j \log \pi$ is only a function of $x_{\mathcal{N}_j}$, its derivates and itself are all low-dimensional functions.

(2) These boundedness assumptions are taken from [110]. It usually does not hold for unbounded support, but it is only introduced for simplicity of analysis and may not be required in practice. On the other hand, for unbounded support, one can consider the averaged acceptance rate where $x^{n,j}$ follows from some distribution that decays fast enough, and a certified lower bound is still obtainable.

(3) The bound in [110] is $1 - M\sqrt{\tau}$, and here we improve it to $1 - M\tau^{3/2}$. We note that the rate $3/2$ is optimal, which can be observed directly from the asymptotic expansion of the acceptance rate.

Proof of Theorem 5.1. The claim is an immediate result of Lemma 5.1. Consider a_j defined in (5.4.10). By definition, we have

$$\alpha(x^{n,j}, z^{n,j}) = \min\left\{1, \exp(a_j(x^{n,j}, \xi_j^{n,j}))\right\},$$

where $z_j^{n,j} = x_j^{n,j} + \tau v_j(x^{n,j}) + \sqrt{2\tau}\xi_j^{n,j}$. By Lemma 5.1,

$$\begin{aligned} 1 - \alpha(x^{n,j}, z^{n,j}) = &\ \exp\left(\min\{0, a_j(x^{n,j}, \xi_j^{n,j})\}\right) \\ &\le |a_j(x^{n,j}, \xi_j^{n,j})| \le \tau^{3/2}(\mathsf{M}_1 + \mathsf{M}_2\|\xi_j^{n,j}\|^3). \end{aligned}$$

Therefore, the result follows by taking expectation w.r.t. $\xi_j^{n,j}$. □

5.2.2 Convergence rate

MLwG also guarantees a dimension independent convergence rate for log-concave localized distributions. To adapt to the block structure, the following blockwise

log-concave condition is introduced (see Assumption 3.4 in [110]).

Definition 5.1. A distribution $\pi \in \mathcal{P}(\mathbb{R}^d)$ is called *blockwise $\lambda_{\boldsymbol{H}}$-log-concave* ($\lambda_{\boldsymbol{H}} > 0$) if there exists a symmetric matrix $\boldsymbol{H} \in \mathbb{R}^{\mathsf{b}\times\mathsf{b}}$ s.t. $\boldsymbol{H} \succeq \lambda_{\boldsymbol{H}} I_{\mathsf{b}}$ and $\forall i, j \in [\mathsf{b}], i \neq j$,

$$\forall x \in \mathbb{R}^d, \quad \nabla^2_{jj} \log \pi(x) \preceq -\boldsymbol{H}_{jj} I_{d_j}, \quad \left\|\nabla^2_{ij} \log \pi(x)\right\| \leq -\boldsymbol{H}_{ij}.$$

Note in the above defintiion, it is implicitly required that

$$\boldsymbol{H}_{jj} \geq 0; \quad \forall i \neq j, \ \boldsymbol{H}_{ij} \leq 0.$$

For more discussions on the blockwise log-concavity, we refer to [110]. In Section 5.3, we will verify this condition in an image deblurring problem with appropriate parameter choices.

Now we state the main theorem.

Theorem 5.2. *Suppose $\pi \in \mathcal{P}(\mathbb{R}^d)$ is blockwise $\lambda_{\boldsymbol{H}}$-log-concave, and is localized on a (s, ν)-local graph G. Assume*

$$\forall j \in [\mathsf{b}], \quad \|v_j\|_{L^\infty} \leq \mathsf{M}, \quad \|\nabla v_j\|_{L^\infty} \leq \mathsf{H}, \quad |\nabla v_j|_{\mathrm{Lip}} \leq \mathsf{L}.$$

Then there exist some $\rho, \tau_0 > 0$ that are independent of d s.t. for all $\tau \in (0, \tau_0]$, we can couple two MLwG chains $\{x^n\}_{n=0}^\infty$ and $\{y^n\}_{n=0}^\infty$ s.t.

$$\Big(\sum_{j\in[\mathsf{b}]} \left[\mathbb{E}\left\|x_j^n - y_j^n\right\|\right]^2\Big)^{1/2} \leq (1-\rho\tau)^n \Big(\sum_{j\in[\mathsf{b}]} \left[\mathbb{E}\left\|x_j^0 - y_j^0\right\|\right]^2\Big)^{1/2}. \tag{5.2.3}$$

As a corollary, take $y^n \sim \pi$, and we show that x^n converges to π exponentially fast with dimension independent rate.

Proofs are delayed to Section 5.4.1.

Remark 5.2. (1) The convergence result is based on the maximal coupling of two MLwG chains as in [110]. In brief, the two chains x^n, y^n are coupled to share in each step the same $\xi_j^{n,j}$ in the proposal and the random variable $\zeta^{n,j} \sim \mathcal{U}(0,1)$ used to determine the acceptance of proposals (See Algorithm 2).

(2) Under the locality assumption, the $\boldsymbol{H}$ matrix in the blockwise λ-log-concave condition can be made to satisfy $\forall i \not\sim j, \ \boldsymbol{H}_{ij} = 0$. We use this condition in the

proof for simplicity. One can also allow it to be nonzero, and the proof still holds with minor modifications.

5.3 Application in an image deblurring problem

In this section, we consider applying MLwG to an image deblurring problem. Due to the use of a total variation (TV) regularization [98, 46], the prior distribution π_0 is non-smooth, and thus one cannot directly use MALA proposal. To address this, we propose to use a local smoothing of the prior distribution, and then apply MLwG to the smoothed distribution.

The locality structure in the image deblurring problem guarantees that the approximation error can be uniformly bounded over the image, and is independent of the dimension. We will show how to implement an efficient local and parallel MLwG sampling algorithm by providing the local target densities and their gradients for the block updates. More details can be found in [46].

5.3.1 Problem setting

Consider the classic image deconvolution problem with TV regularization, e.g., [87], and assume that a blurred and noisy image $y \in \mathbb{R}^d$ is obtained by

$$y = Ax_{\text{true}} + \epsilon. \tag{5.3.1}$$

Here $x_{\text{true}} \in \mathbb{R}^d$ is the 'true' image, $\epsilon \sim \mathsf{N}(0, \lambda^{-1} I_d)$, and $A \in \mathbb{R}^{d\times d}$ is the convolution operator. For instance, one can construct A via the discrete point spread function (PSF). Assume that the discrete PSF has radius $r > 0$, then Ax convolves each pixel with the surrounding $(2r+1)^2$ pixels.

The inverse problem of (5.3.1) is to recover a solution that is close to x_{true} from the data y. Computing a solution is typically not straightforward due to the ill-posedness of the problem. For this reason, we employ the edge-preserving TV regularization introduced in [98], which is a commonly used regularization technique in image reconstruction. For discretized images, it reads

$$\|x\|_{\text{TV}} = \sum_{s=1}^{d} \sqrt{(D_s^{(v)}x)^2 + (D_s^{(h)}x)^2},$$

where $D^{(v)} \in \mathbb{R}^{d\times d}$ and $D^{(h)} \in \mathbb{R}^{d\times d}$ are finite difference matrices cooresponding to

the vertical and horizontal differences of the pixels resp. Specifically,

$$\begin{aligned} D^{(v)} &= I_n \otimes D_n, \\ D^{(h)} &= D_n \otimes I_n, \end{aligned} \quad \text{where} \quad D_n = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \\ & & & & -1 \end{bmatrix}_{[n\times n]}.$$

Here $n = \sqrt{d}$ and $\otimes$ denotes the Kronecker product. Note we use Dirichlet boundary conditions for the finite difference matrices. Other boundary conditions can be used by slight modification of the algorithm and the analysis. For simplicity, we only consider square images with uniform block decomposition, i.e.

- Image is of size $n \times n$ pixels, and thus $d = n^2$.
- Image is equally divided into $\mathsf{b} = b^2$ number of $m \times m$ blocks. Here $n = bm$.

We now formulate the Bayesian inverse problem. The prior distribution is given by the TV regularization, i.e. the TV prior

$$\pi_0(x) \propto \exp\left(-\mu \|x\|_{\mathrm{TV}}\right),$$

where $\mu > 0$ is some fixed parameter that controls the strength of the regularization. The likelihood function is determined by the model (5.3.1), i.e.

$$\pi(y|x) \propto \exp\left(-\frac{\lambda}{2}\|y - Ax\|_2^2\right).$$

The posterior distribution is then given by

$$\begin{aligned} \pi(x) := \pi(x|y) &\propto \exp\left(-l(x) - \varphi_0(x)\right), \\ l(x) = \frac{\lambda}{2}\|y - Ax\|_2^2, &\quad \varphi_0(x) = \mu\|x\|_{\mathrm{TV}}. \end{aligned} \tag{5.3.2}$$

Here we omit the dependence on y as it is fixed in the sampling task.

5.3.2 Posterior smoothing with dimension-free error

Since MLwG requires the gradient of the log density, we propose to approximate the non-smooth π in (5.3.2) by a smooth one π_ε. We show that the error between

π and π_ε is uniformly distributed among all the components, leading to a local dimension-independent error on the marginal distribution of any block x_i.

The non-smoothness of π originates from the potential φ_0 (cf. (5.3.2)). Hence, we replace φ_0 with a smoothed potential φ_ε for some small $\varepsilon > 0$, such that $\varphi_\varepsilon \to \varphi_0$ as $\varepsilon \to 0$. Various smoothing methods are possible, but it is crucial to ensure that the introduced error remains small. Here we consider the following approximation

$$\varphi_\varepsilon(x) := \mu \sum_{s=1}^{d} \sqrt{(D_s^{(v)}x)^2 + (D_s^{(h)}x)^2 + \varepsilon}. \tag{5.3.3}$$

Thus the smoothed posterior density becomes

$$\pi_\varepsilon(x) \propto \exp\left(-\frac{\lambda}{2}\|y - Ax\|_2^2 - \mu \sum_{s=1}^{d} \sqrt{(D_s^{(v)}x)^2 + (D_s^{(h)}x)^2 + \varepsilon}\right). \tag{5.3.4}$$

As we modify φ_0, which is a function in $\mathbb{R}^d$, the distance between π and π_ε in general depends on the dimension d. For instance, one can show $\mathsf{KL}(\pi\|\pi_\varepsilon) = \mathcal{O}(d\varepsilon)$. However, when examining the marginals of π and π_ε over small blocks x_j, we can show that the approximation error is *dimension-independent* with the help of the marginal transport inequality. We comment that such dimension independence is crucial for solving the image deblurring problem. It ensures that the smoothing error is evenly distributed across the image, rather than concentrating on certain pixels and creating unwanted artifacts in the image.

In light of Theorem 3.3, the key to is to verify that π_ε is δ-localized. We will mainly use Theorem 3.2, and we introduce some diagonal block dominance condition imposed on $C = A^{\mathrm{T}}A$.

Definition 5.2. A matrix $C \in \mathbb{R}^{d\times d}$ is called **c-diagonal block dominant** for some $c > 0$, if there exists a symmetric matrix $M \in \mathbb{R}^{\mathsf{b}\times\mathsf{b}}$ s.t. $\forall i, j \in [\mathsf{b}], i \neq j$,

$$C_{jj} \succeq M_{jj} I_{d_j}, \quad \|C_{ij}\|_2 \leq M_{ij}, \quad \sum_{k:k\neq j} M_{jk} + c \leq M_{jj}.$$

Remark 5.3. Definition 5.1 introduces a blockwise log-concavity condition similar to Definition 5.2. The c-diagonal block dominance here can be viewed as an ℓ_1 version of the blockwise log-concavity condition (which is an ℓ_2 condition).

Now we state the main theorem.

Theorem 5.3. *Consider the target distribution π (5.3.2) and its smooth approximation π_ε (5.3.4). Assume that $A^{\mathrm{T}}A$ is c-diagonal block dominant (Definition 5.2). Suppose $\frac{\lambda}{\mu} \geq \frac{36m}{c\sqrt{\varepsilon}}$. Then there exists a dimension-independent constant C s.t.*

$$\max_j \mathsf{W}_1(\pi_j, \pi_{\varepsilon,j}) \leq C\varepsilon. \tag{5.3.5}$$

Proofs are delayed to Section 5.4.2.

5.3.3 Dimension-free acceptance rate and convergence rate

In the following, we present two results, which show that the acceptance rate and convergence rate of MLwG for the smoothed distribution π_ε are independent of the dimension d. Since they are direct application of Theorem 5.1 and Theorem 5.2, we only state the results here.

Proposition 5.1. *Suppose A is bounded in the sense that $\exists C_A > 0$ s.t.*

$$\forall i, j \in [\mathsf{b}], \quad \|A_{ij}\| \leq C_A. \tag{5.3.6}$$

Then the acceptance rate of the MLwG proposal for π_ε is bounded below by

$$\mathbb{E}_{\xi_j^{n,j}}\left[\alpha(x^{n,j}, z^{n,j})\right] \geq 1 - M\tau^{3/2}.$$

Here M is a dimension-independent constant depending only on $m, C_A, \lambda, \mu, \varepsilon$, and $\max_j \|[Ax^{k,j} - y]_j\|$.

Remark 5.4. Here the lower bound depends on the state of x through the term $\max_j \|[Ax^{k,j} - y]_j\|$. However, since the pixels values are bounded in practice, this term is usually also bounded and dimension-independent.

Proposition 5.2. *Suppose A is bounded as in (5.3.6), $A^{\mathrm{T}}A$ is c-diagonally block dominant, and $\frac{\lambda}{\mu} \geq \frac{36m}{c\sqrt{\varepsilon}}$. Then there exist some $\rho, \tau_0 > 0$ that are independent of d s.t. for all $\tau \in (0, \tau_0]$, we can couple two MLwG chains $\{x^n\}_{n=0}^\infty$ and $\{y^n\}_{n=0}^\infty$ s.t.*

$$\Big(\sum_{j\in[\mathsf{b}]}\left[\mathbb{E}\left\|x_j^n - y_j^n\right\|\right]^2\Big)^{1/2} \leq (1-\rho\tau)^n\Big(\sum_{j\in[\mathsf{b}]}\left[\mathbb{E}\left\|x_j^0 - y_j^0\right\|\right]^2\Big)^{1/2}.$$

As a corollary, take $y^n \sim \pi$, and we show that x^n converges to π exponentially fast with dimension independent rate.

5.3.4 Local and parallel algorithm

Since the convolution operator A has a small radius r, and the difference operator $D^{(v)}, D^{(h)}$ are local, the posterior distribution π_ε is localized. We denote some notations to describe the locality structure.

- $\mathcal{I}_i \subseteq [d]$ denotes the index set of pixels in the i-th block.
- For $s, t \in [d]$, denote $s \sim t$ if $\partial^2_{st} \log \pi_\varepsilon(x) \equiv 0$.
- $\Theta_i = \{t : \exists s \in \mathcal{I}_i, t \sim s\}$ denote the index set of all the pixels that are neighbors of the i-th block.

In the (n, j)-th step of the MLwG algorithm (see Algorithm 2), we introduce the local negative log-density:

$$l^{n,j}_{\mathrm{loc}}(x_j | x^{n,j}_{-j}) := \frac{\lambda}{2} \sum_{s \in \Theta_j} \left| y_s - A_{sj} x_j - A_{s,-j} x^{n,j}_{-j} \right|^2 + \mu \sum_{s \in \Theta_j} \sqrt{(D^{(v)}_{s,j} x_j + D^{(v)}_{s,-j} x^{n,j}_{-j})^2 + (D^{(h)}_{s,j} x_j + D^{(h)}_{s,-j} x^{n,j}_{-j})^2 + \varepsilon}.$$

Note by definition,

$$\forall t \notin \Theta_j, \quad \partial_{x^{n,j}_t} l^{n,j}_{\mathrm{loc}}(x_j | x^{n,j}_{-j}) = 0.$$

Therefore, computing $l^{n,j}_{\mathrm{loc}}(x_j | x^{n,j}_{-j})$ does not require $x_{[d] \setminus \Theta_j}$. This allows for local and parallel implementation of MLwG. To be specific, note

$$\nabla_j \log \pi_\varepsilon(x^{n,j}) = -\nabla l^{n,j}_{\mathrm{loc}}(x^{n,j}_j | x^{n,j}_{-j}).$$

So that one can use the proposal

$$z^{n,j}_j = x^{n,j}_j - \tau \nabla l^{n,j}_{\mathrm{loc}}(x^{n,j}_j | x^{n,j}_{-j}) + \sqrt{2\tau} \xi^{n,j}_j, \quad \xi^{n,j}_j \sim \mathcal{N}(0, I_{d_j}),$$

which does not involve $x^{n,j}_{[d] \setminus \Theta_j}$. Similarly, to compute the acceptance rate, one uses

$$\alpha(x^{n,j}, z^{n,j}) = \min \left\{ 1, \frac{\exp(-l^{n,j}_{\mathrm{loc}}(z^{n,j}_j | x^{n,j}_{-j})) Q_j(z^{n,j}_j, x^{n,j}_j \mid x^{n,j}_{-j})}{\exp(-l^{n,j}_{\mathrm{loc}}(x^{n,j}_j | x^{n,j}_{-j})) Q_j(x^{n,j}_j, z^{n,j}_j \mid x^{n,j}_{-j})} \right\},$$

$$\text{where} \quad \log Q_j(x^{n,j}_j, z^{n,j}_j \mid x^{n,j}_{-j}) = -\frac{1}{4\tau} \left\| z^{n,j}_j - x^{n,j}_j + \tau \nabla l^{n,j}_{\mathrm{loc}}(x^{n,j}_j | x^{n,j}_{-j}) \right\|^2 + \text{const.}$$

And this can also be computed locally.

The local implementation facilitates updating several blocks in parallel during the loop over j in Algorithm 2. We define a parallel scheme via the index sets

$$\mathcal{U}_l \subseteq [\mathsf{b}], \quad l = 1, \ldots, L,$$

$$\text{s.t.} \quad \forall l,\ \forall i, j \in \mathcal{U}_l, \quad \Theta_i \cap \Theta_j = \varnothing.$$

So that blocks in $\mathcal{U}_l$ can be updated in parallel. The choices $\{\mathcal{U}_l\}_{l\in[L]}$ are not unique, and one can find parallel scheme with $L = 4$ for the imaging deblurring problem. An example of a parallel updating schedule is illustrated in Figure 5.1.

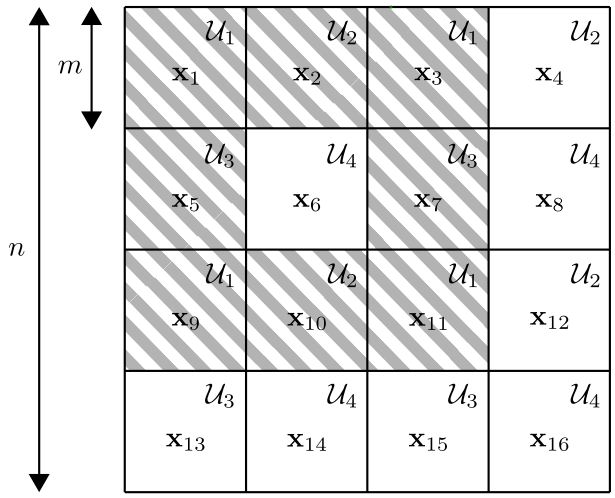

Figure 5.1: Block decomposition: example of a parallel scheme with 16 blocks.

For more details on the local and parallel algorithm, we refer to Section 5 in [46].

5.3.5 Numerical examples

We use the Cameraman image as the testing example. Figure 5.2 shows the true image and the blurred and noised image, which is obtained via the observation model (5.3.1). Here we take A corresponds to the discretization of a Gaussian blurring kernel with radius 8 and standard deviation 8. The noise is $\epsilon \sim \mathsf{N}(0, 10^{-4} \cdot I)$.

To solve the deblurring problem, we use MLwG to sample from the smoothed posterior distribution π_ε (5.3.4). We first check the effect of different choices of the smoothing parameter ε on the sampling performance. Second, we will verify numerically the dimension-independent acceptance rate and convergence rate of MLwG. Finally, we compare MLwG to MALA and show that the local and parallel

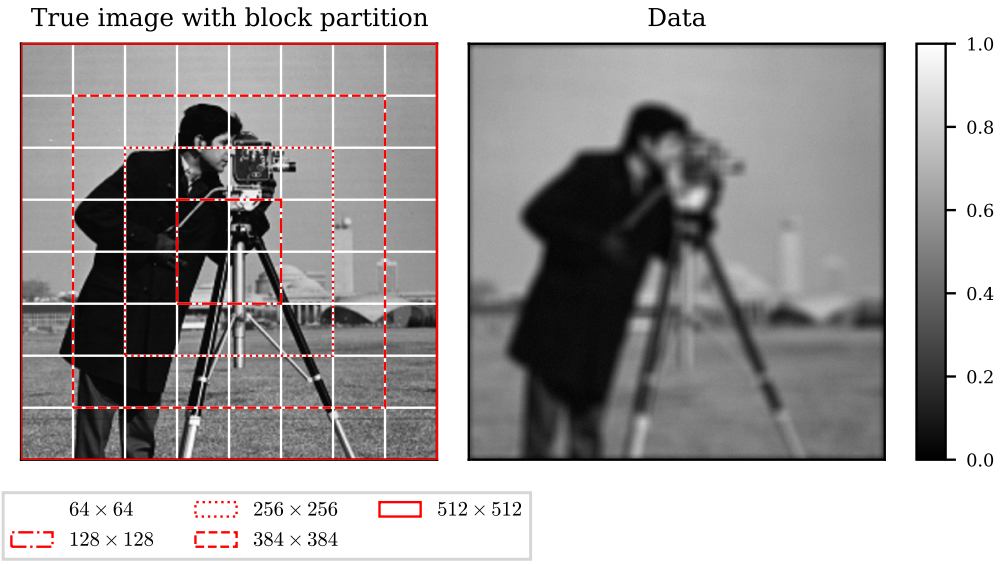

Figure 5.2: Image deblurring problem: Cameraman. Left: Cameraman image and partition into different sizes (red frames). All sections are again partitioned into blocks of equal size 64×64 (white frames). Right: Data obtained via Gaussian blur and additive Gaussian noise.

implementation outperforms MALA in increasing dimension in terms of both sample quality and wall-clock time.

In each experiment, we generate 5 independent chains of 2000 samples each, and apply thinning by recording every 200-th sample to reduce autocorrelation. We check our sample chains for convergence by means of the potential scale reduction factor (PSRF) [48], which compares the within-variance to the in-between variance of the chains. Empirically, one considers sample chains to be converged if PSRF< 1.1. We also compute the normalized effective sample size (nESS) and credible intervals (CIs), see for instance [82] for definitions.

To determine the hyperparameter μ in (5.3.2), we use the adaptive total variation approach in [85], and obtain $\mu = 35.80$ for the 512×512 image. We use this choice for all other problem sizes as well.

Influence of ε We compute MAP estimates for $\varepsilon \in \{10^{-3}, 10^{-5}, 10^{-7}\}$ with the majorization-minimization algorithm proposed in [108] and show the results in the left column of Figure 5.3. We can see that the restoration from $\varepsilon = 10^{-3}$ has smoother edges than the other two restorations which exhibit the typical cartoon-like structure of TV-regularized images. Further, the difference between the results from $\varepsilon = 10^{-5}$ and 10^{-7} is hardly visible.

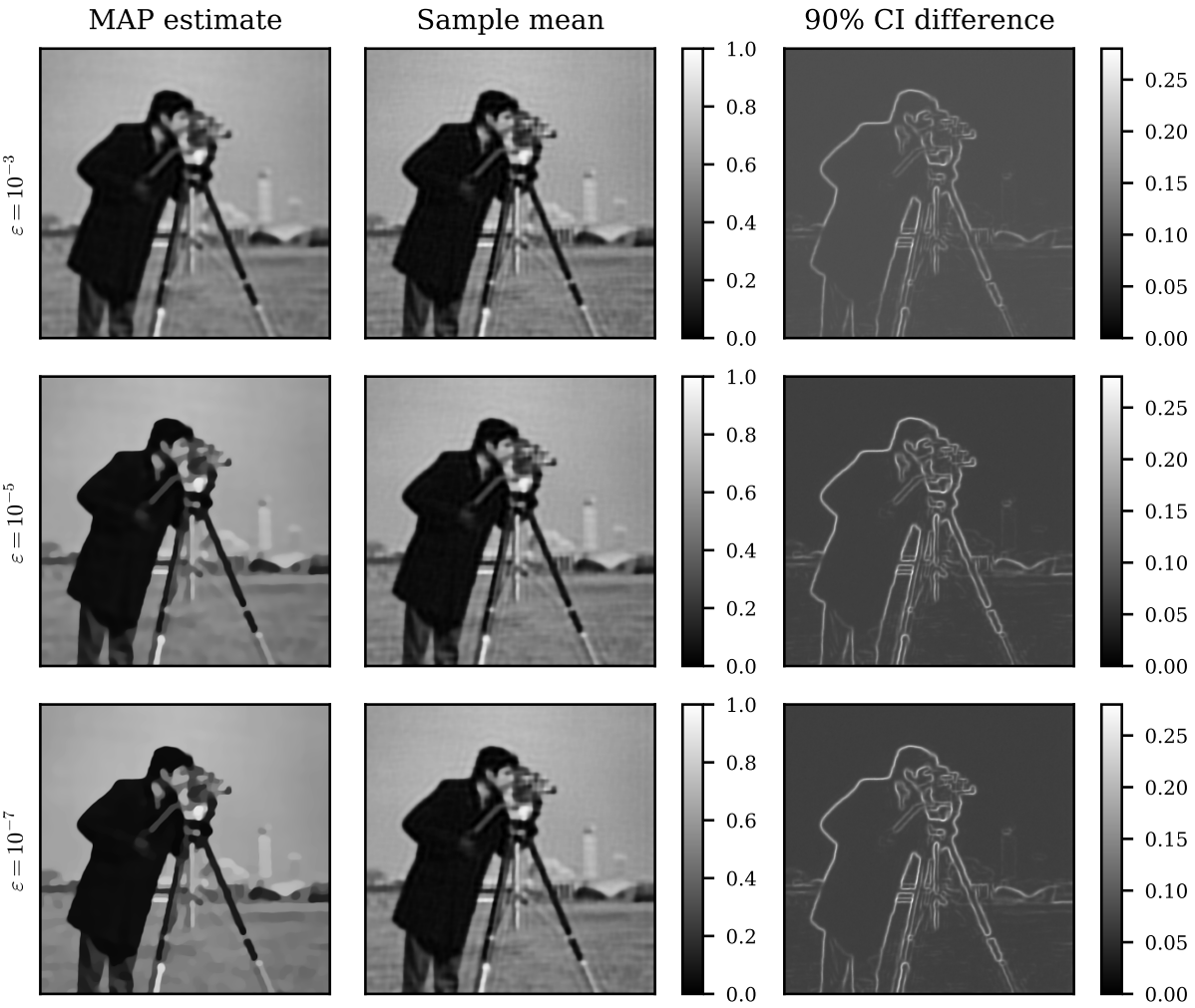

Figure 5.3: Image deblurring problem: influence of ε. MAP estimate, MLwG sample mean, and widths of the 90% sample CIs for $\varepsilon \in \{10^{-3}, 10^{-5}, 10^{-7}\}$.

Then we run MLwG for different ε with a diminishing step size adaptation during burn-in [78] targeting an acceptance rate of 0.547 in each block (see [94]). We show the sample means and widths of the 90% sample CIs in Figure 5.3. Here the sample means of $\varepsilon = 10^{-5}$ and 10^{-7} are visually more favorable than their corresponding MAP estimates. In contrast, the result of $\varepsilon = 10^{-3}$ contains visible artifacts. Moreover, the 90% sample CI difference is in general wider for $\varepsilon = 10^{-3}$ than for $\varepsilon = 10^{-5}$ and $\varepsilon = 10^{-7}$. However, the width of the 90% sample CIs are rather similar on the edges.

We show some quantitative results in Table 5.1. Here we note that $\varepsilon = 10^{-3}$ allows for a significantly larger mean step size in comparison to $\varepsilon = 10^{-5}$ or 10^{-7}. This results in less correlated samples, which is reflected in a larger nESS.

ε	min nESS [%]	τ [10^{-6}]	α [%]	max PSRF	med PSRF
10^{-3}	13.3	25.8	54.7	1.01	1.00
10^{-5}	3.2	7.5	54.4	1.03	1.00
10^{-7}	2.1	5.6	54.3	1.04	1.00

Table 5.1: Image deblurring problem: influence of ε. Sampling results of MLwG for $\varepsilon \in \{10^{-3}, 10^{-5}, 10^{-7}\}$. Here min nESS is the pixel-wise minimum of the mean nESS averaged over the 5 chains. The step size τ and acceptance rate α are averaged over all blocks and the 5 chains. The maximum and median PSRF are with respect to the pixels.

We conclude that relatively large values of ε make the posterior density smoother, allowing for larger step sizes and thus making the sampling more efficient in terms of nESS. However, at the same time, the results can be visually significantly different compared smaller ε that yields sharper edges. Based on these observations, and since the results for $\varepsilon = 10^{-5}$ and 10^{-7} are very close, we will fix $\varepsilon = 10^{-5}$ in the remaining experiments.

Dimension-independent acceptance rate To test the dimension-independent block acceptance rate, we partition the original 512×512 image into 4 sections of sizes 128×128, 256×256, 384×384 and 512×512. Furthermore, each section is partitioned into blocks of equal size 64×64. Thus the number of blocks in the sections of different sizes are 4, 16, 36, and 64 resp. The 4 deblurring problems are

shown on the left in Figure 5.2.

We run MLwG with a step size of $\tau = 7.44 \times 10^{-6}$ on the 4 deblurring problems with different sizes. The step size is taken from a pilot run on the 512×512 problem by targeting an acceptance rate of 0.547 in each block and then taking the average of all block step sizes. For all problem sizes, we use a burn-in period of $31,250$ samples. We plot the acceptance rate for each block in Figure 5.4 and see that the block acceptance rates are indeed dimension-independent.

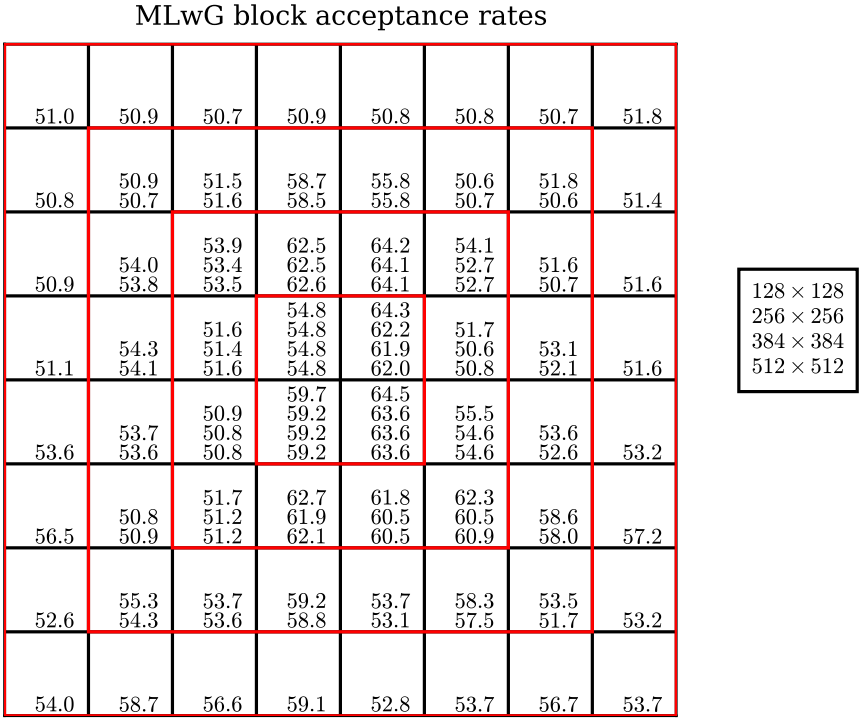

Figure 5.4: Image deblurring problem: acceptance rate. Block acceptance rates (%) of MLwG for different problem sizes, listed according to the problem sizes in the order shown on the right.

Comparison to MALA In this part, we compare the performance of MLwG with vanilla MALA. For MALA, we use again the diminishing step size adaptation from [78] during burn-in, where we target an acceptance rate of 0.547. The numbers of burn-in samples are listed in Table 5.2 and are chosen such that they increase linearly with the problem size. For MLwG, we use the same setting as in the previous tests.

We compare the sampling performance of MALA and MLwG in Table 5.2. In general, MLwG yields much larger nESS than MALA because it allows for a larger step size. Furthermore, the nESS of MLwG becomes even larger as the

problem size increases. We attribute this to the diminishing constraining effect of the boundary condition associated with the convolution operator on the inner blocks as the dimension increases. In addition, we note that for the given burn-in, MALA does not converge for the problem sizes 384×384 and 512×512, since the corresponding max PSRF> 1.1.

Problem size		**128×128**	**256×256**	**384×384**	**512×512**
min nESS [%]	MLwG	2.6	2.8	2.9	3.0
	MALA	1.4	0.8	0.6	0.4
$\boldsymbol{\tau}$ [$\mathbf{10^{-6}}$]	MLwG	7.4	7.4	7.4	7.4
	MALA	4.8	2.5	1.8	1.4
$\boldsymbol{\alpha}$ [%]	MLwG	60.8	57.7	55.5	54.3
	MALA	54.0	54.3	54.7	55.0
burn-in [$\mathbf{10^3}$]	MLwG	31.250	31.250	31.250	31.250
	MALA	125.000	500.000	1125.000	2000.000
max PSRF	MLwG	1.03	1.03	1.03	1.03
	MALA	1.06	1.08	1.20	1.20

Table 5.2: Comparison of MLwG and MALA for different problem dimensions. The min nESS is the pixel-wise minimum of the mean nESS taken over the 5 chains. For MLwG, the shown step size τ and acceptance rate α are the means taken over all blocks and the 5 chains. The maximum PSRF is with respect to the pixels.

Notice that the results from Table 5.2 also validate the dimension-independent convergence rate in Proposition 5.2 of MLwG. This is because MLwG produces roughly the same PSRF for all problem sizes with the same burn-in steps. In contrast, MALA requires significantly more burn-in steps with increasing dimension.

Finally, we compare the wall-clock and CPU time of the sample chains of MLwG and MALA. All chains are run on the same hardware, Intel® Xeon® E5-2650 v4 processors. Furthermore, we use the optimal number of cores for MLwG, such that all blocks with indices $i \in \mathcal{U}_l$ can be updated in parallel (see Section 5.3.4).

We show the computing times in seconds per 1000 samples in Figure 5.5 and observe that the wall-clock time of MLwG remains almost constant and does not increase with the problem dimension. This is because the main computational effort of updating the 64×64 blocks on each core remains constant and only more time is required for handling the increasing number of cores by the main process. For small problem sizes, the wall-clock time of MLwG is longer than that of MALA, because

of the overhead of the parallelized implementation and the additional convolutions of fixed pixels in the local block likelihoods. However, since several updates are run in parallel in MLwG, its wall-clock time is eventually shorter than that of MALA, see the time for problem size 512×512. Note that the total wall-clock time of MALA is actually significantly larger, since it requires much more burn-in. The benefits of MLwG obviously come at the cost of CPU time, which increases linearly with the number of cores.

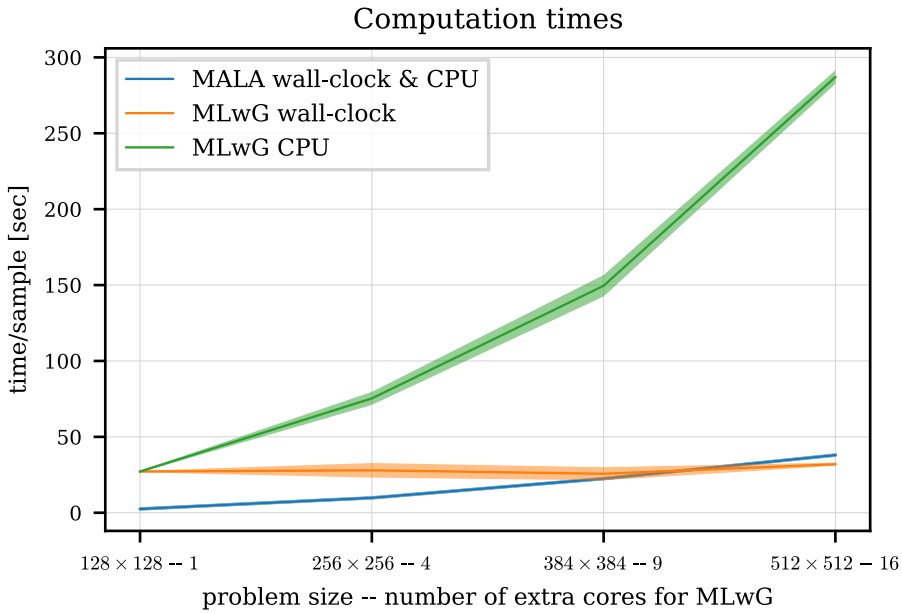

Figure 5.5: Image deblurring problem: wall-clock and CPU time of MLwG and MALA. For MLwG, we show the mean $\pm$ the standard deviation by means of the shaded area. The wall-clock and CPU time of MALA are approximately equal and are therefore not displayed separately. For MLwG, we used the number of cores indicated in the x-tick labels plus one additional core to handle the main process.

5.4 Proofs

5.4.1 Proof of Theorem 5.2

Proof of Theorem 5.2. Consider the maximal coupling of two MLwG chains $x^{n,j}$ and $y^{n,j}$, where the two chains share the same $\xi_j^{n,j}$ and $\zeta^{n,j}$ in Algorithm 2 for all n, j. The proof is divided into two parts: (i) derive the coupling inequality for one MALA step, and (ii) derive the contraction of one cycle.

Before the proof, we denote the filtration

$$\mathcal{F}^{n,j} = \text{span}\left\{x^0, y^0, \xi_i^{k,i}, \zeta^{k,i} \mid \forall k < n, \forall i \text{ or } k = n, \forall i < j\right\}.$$

Accordingly, denote $\mathbb{E}^{n,j}$ as the conditional expectation w.r.t. $\mathcal{F}^{n,j}$.

I. One MALA step analysis. Denote the proposals for the two chains as

$$z_{-j}^{n,j} = x_{-j}^{n,j}, \quad z_j^{n,j} = x_j^{n,j} + \tau v_j(x^{n,j}) + \sqrt{2\tau}\xi_j^{n,j},$$

$$w_{-j}^{n,j} = y_{-j}^{n,j}, \quad w_j^{n,j} = y_j^{n,j} + \tau v_j(y^{n,j}) + \sqrt{2\tau}\xi_j^{n,j}.$$

Depending on the acceptance of the proposals, we denote the events

- $\mathbf{1}_2$: both proposals are accepted.
- $\mathbf{1}_x$ or $\mathbf{1}_y$: only one proposal (x or y) is accepted.
- $\mathbf{1}_0$: both proposals are rejected.

By definition, we can decompose

$$\begin{aligned}
&\mathbb{E}^{n,j}\|x_j^{n,j+1} - y_j^{n,j+1}\| \\
&= \mathbb{E}_{\xi_j^{n,j},\zeta^{n,j}}\|x_j^{n,j+1} - y_j^{n,j+1}\| \\
&= \mathbb{E}_{\xi_j^{n,j}}\left[\mathbb{E}_{\zeta^{n,j}}[\mathbf{1}_2]\cdot\|z_j^{n,j} - w_j^{n,j}\|\right] + \mathbb{E}_{\xi_j^{n,j}}\left[\mathbb{E}_{\zeta^{n,j}}[\mathbf{1}_0]\cdot\|x_j^{n,j} - y_j^{n,j}\|\right] \\
&\quad + \mathbb{E}_{\xi_j^{n,j}}\left[\mathbb{E}_{\zeta^{n,j}}[\mathbf{1}_x]\cdot\|z_j^{n,j} - y_j^{n,j}\| + \mathbb{E}_{\zeta^{n,j}}[\mathbf{1}_y]\cdot\|x_j^{n,j} - w_j^{n,j}\|\right].
\end{aligned} \tag{5.4.1}$$

From now on, we omit for simplicity the superscript n, j in the notation. For the case where both proposals are accepted, denote $\eta_t = (1-t)x + ty,\ t \in [0,1]$, and we have

$$\begin{aligned}
z_j - w_j &= x_j - y_j + \tau(v_j(x) - v_j(y)) \\
&= x_j - y_j + \tau\sum_{i\in[\mathsf{b}]}\int_0^1 \nabla_i v_j(\eta_t)(x_i - y_i)\mathrm{d}t \\
&= \left(I + \tau\int_0^1 \nabla_j v_j(\eta_t)\mathrm{d}t\right)(x_j - y_j) + \tau\sum_{i:i\neq j}\int_0^1 \nabla_i v_j(\eta_t)(x_i - y_i)\mathrm{d}t.
\end{aligned}$$

Since π is blockwise λ-log-concave, by definition we have

$$\nabla_j v_j(\eta_t) \preceq -\boldsymbol{H}_{jj} I_{d_j}, \quad \|\nabla_i v_j(\eta_t)\| \leq -\boldsymbol{H}_{ij}\ (\forall i \neq j).$$

Therefore, if $\tau < \boldsymbol{H}_{jj}^{-1}$, we obtain that

$$\begin{aligned} \|z_j - w_j\| \leq\ & (1 - \tau \boldsymbol{H}_{jj}) \|x_j - y_j\| - \tau \sum_{i:i\neq j} \boldsymbol{H}_{ij} \|x_i - y_i\| \\ =\ & \|x_j - y_j\| - \tau \sum_{i\in[\mathsf{b}]} \boldsymbol{H}_{ij} \|x_i - y_i\| . \end{aligned} \tag{5.4.2}$$

For the case where only one proposal is accepted, we have

$$\|z_j - y_j\| \leq \|x_j - y_j\| + \tau \|v_j(x)\| + \sqrt{2\tau} \|\xi_j\| .$$

Similarly for $\|x_j - w_j\|$. So that we have

$$\max\{\|z_j - y_j\| , \|x_j - w_j\|\} \leq \|x_j - y_j\| + \tau \mathsf{M} + \sqrt{2\tau} \|\xi_j\| . \tag{5.4.3}$$

Therefore, combining (5.4.1) (5.4.2) and (5.4.3), we have

$$\begin{aligned} \mathsf{LHS} \leq\ & \mathbb{E}_{\xi_j}[\mathbb{E}_\zeta[\mathbf{1}_2]] \cdot \left[\|x_j - y_j\| - \tau \sum_{i\in[\mathsf{b}]} \boldsymbol{H}_{ij} \|x_i - y_i\| \right] \\ & + \mathbb{E}_{\xi_j}[\mathbb{E}_\zeta[\mathbf{1}_0] \cdot \|x_j - y_j\|] \\ & + \mathbb{E}_{\xi_j}\left[(\mathbb{E}_\zeta[\mathbf{1}_x] + \mathbb{E}_\zeta[\mathbf{1}_y]) \cdot \left[\|x_j - y_j\| + \tau \mathsf{M} + \sqrt{2\tau} \|\xi_j\| \right] \right] \\ =\ & \|x_j - y_j\| - \tau \mathbb{E}_{\xi_j}[\mathbb{E}_\zeta[\mathbf{1}_2]] \cdot \sum_{i\in[\mathsf{b}]} \boldsymbol{H}_{ij} \|x_i - y_i\| \\ & + \mathbb{E}_{\xi_j}\left[(\mathbb{E}_\zeta[\mathbf{1}_x] + \mathbb{E}_\zeta[\mathbf{1}_y]) \cdot \left[\tau \mathsf{M} + \sqrt{2\tau} \|\xi_j\| \right] \right] . \end{aligned} \tag{5.4.4}$$

Here we use the fact that $\mathbf{1}_2 + \mathbf{1}_x + \mathbf{1}_y + \mathbf{1}_0 \equiv 1$. By definition of the acceptance probability, we have

$$1 - \mathbb{E}_\zeta[\mathbf{1}_2] \leq 1 - \alpha(x, z) + 1 - \alpha(y, w).$$

$$\mathbb{E}_\zeta[\mathbf{1}_x] + \mathbb{E}_\zeta[\mathbf{1}_y] = |\alpha(x, z) - \alpha(y, w)| .$$

Recall the definition of a_j in (5.4.10). By Lemma 5.1, we have

$$\begin{aligned} & 1 - \alpha(x, z) + 1 - \alpha(y, w) \\ =\ & 1 - \exp(\min\{0, a_j(x, \xi)\}) + 1 - \exp(\min\{0, a_j(y, \xi)\}) \\ \leq\ & |a_j(x, \xi)| + |a_j(y, \xi)| \leq 2\tau^{3/2}(\mathsf{M}_1 + \mathsf{M}_2 \|\xi_j\|^3). \end{aligned}$$

$$
\begin{aligned}
|\alpha(x,z)-\alpha(y,w)| =&\ |\exp\left(\min\{0,a_j(x,\xi)\}\right)-\exp\left(\min\{0,a_j(y,\xi)\}\right)| \\
\le&\ |a_j(x,\xi)-a_j(y,\xi)| \\
\le&\ \tau(\mathsf{M}_3+\mathsf{M}_4\|\xi_j\|^2)\sum_{k\in\mathcal{N}_j}\|x_k-y_k\|.
\end{aligned}
$$

So that

$$
1\ge\mathbb{E}_{\xi_j}\left[\mathbb{E}_\zeta[\mathbf{1}_2]\right]\ge\mathbb{E}_{\xi_j}\left[1-\tau^{3/2}(\mathsf{M}_1+\mathsf{M}_2\|\xi_j\|^3)\right]=1-\mathsf{C}_1\tau^{3/2}.
$$

$$
\begin{aligned}
\mathbb{E}_{\xi_j}\left[(\mathbb{E}_\zeta[\mathbf{1}_x]+\mathbb{E}_\zeta[\mathbf{1}_y])\right]&\le\mathbb{E}_{\xi_j}\left[\tau(\mathsf{M}_3+\mathsf{M}_4\|\xi_j\|^2)\sum_{k\in\mathcal{N}_j}\|x_k-y_k\|\right] \\
&\le\mathsf{C}_2\tau\sum_{k\in\mathcal{N}_j}\|x_k-y_k\|.
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_{\xi_j}\left[(\mathbb{E}_\zeta[\mathbf{1}_x]+\mathbb{E}_\zeta[\mathbf{1}_y])\cdot\|\xi_j\|\right]&\le\mathbb{E}_{\xi_j}\left[\tau(\mathsf{M}_3+\mathsf{M}_4\|\xi_j\|^2)\,\|\xi_j\|\sum_{k\in\mathcal{N}_j}\|x_k-y_k\|\right] \\
&\le\mathsf{C}_3\tau\sum_{k\in\mathcal{N}_j}\|x_k-y_k\|.
\end{aligned}
$$

Notice $\boldsymbol{H}_{jj}\ge 0$ and $\boldsymbol{H}_{ij}\le 0$ if $i\neq j$, so that pluging the above inequalities into (5.4.4), we have

$$
\begin{aligned}
\mathsf{LHS}\le&\ \|x_j-y_j\|-\tau\boldsymbol{H}_{jj}\left(1-\mathsf{C}_1\tau^{3/2}\right)\|x_j-y_j\|-\tau\sum_{i:i\neq j}\boldsymbol{H}_{ij}\|x_i-y_i\| \\
&+\left(\mathsf{C}_2\tau^2\mathsf{M}+\mathsf{C}_3\tau\sqrt{2\tau}\right)\sum_{k\in\mathcal{N}_j}\|x_k-y_k\|.
\end{aligned}
$$

Therefore, for some C, we proved that

$$
\begin{aligned}
\mathbb{E}^{n,j}\|x_j^{n,j+1}-y_j^{n,j+1}\|\le\|x_j^{n,j}-y_j^{n,j}\|&-\tau\sum_{i\in[\mathsf{b}]}\boldsymbol{H}_{ij}\|x_i^{n,j}-y_i^{n,j}\| \\
&+\mathsf{C}\tau^{3/2}\sum_{k\in\mathcal{N}_j}\|x_k^{n,j}-y_k^{n,j}\|.
\end{aligned}
\tag{5.4.5}
$$

II. Contraction in one cycle. By definition,

$$
x_j^{n,k}=\begin{cases}x_j^{n-1}, & \text{if } k\le j,\\ x_j^n, & \text{if } k>j.\end{cases}
$$

Similarly for $y_j^{n,k}$. Denote the error vector $e^n \in \mathbb{R}^{\mathsf{b}}$ s.t.

$$e_j^n := \|x_j^n - y_j^n\|.$$

Taking expectation on (5.4.5), and replacing $\|x_j^n - y_j^n\|$ by e_j^n, we have

$$\begin{aligned}\mathbb{E}[e_j^n] \le \mathbb{E}[e_j^{n-1}] - \tau \sum_{i:i<j} \boldsymbol{H}_{ij}\mathbb{E}[e_i^n] - \tau \sum_{i:i\ge j} \boldsymbol{H}_{ij}\mathbb{E}[e_i^{n-1}] \\ + \mathsf{C}\tau^{3/2} \sum_{k\in\mathcal{N}_j, k<j} \mathbb{E}[e_k^n] + \mathsf{C}\tau^{3/2} \sum_{k\in\mathcal{N}_j, k\ge j} \mathbb{E}[e_k^{n-1}].\end{aligned}$$

Denote the matrices $\boldsymbol{H}^{\mathrm{L}}, \boldsymbol{H}^{\mathrm{U}}, \boldsymbol{G}^{\mathrm{L}}, \boldsymbol{G}^{\mathrm{U}}$ as

$$\boldsymbol{H}_{ij}^{\mathrm{L}} = \boldsymbol{H}_{ij}\mathbf{1}_{i<j}, \quad \boldsymbol{H}_{ij}^{\mathrm{U}} = \boldsymbol{H}_{ij}\mathbf{1}_{i\ge j}, \quad \boldsymbol{G}_{ij}^{\mathrm{L}} = \mathsf{C}\mathbf{1}_{i\sim j}\mathbf{1}_{i<j}, \quad \boldsymbol{G}_{ij}^{\mathrm{U}} = \mathsf{C}\mathbf{1}_{i\sim j}\mathbf{1}_{i\ge j}. \tag{5.4.6}$$

Then the above inequality can be written in a matrix form

$$\left(I + \tau\boldsymbol{H}^{\mathrm{L}} - \tau^{3/2}\boldsymbol{G}^{\mathrm{L}}\right)\mathbb{E}[e^n] \le \left(I - \tau\boldsymbol{H}^{\mathrm{U}} + \tau^{3/2}\boldsymbol{G}^{\mathrm{U}}\right)\mathbb{E}[e^{n-1}].$$

Here $\le$ is defined in the elementwise sense. When $\tau < \|\boldsymbol{H}^{\mathrm{L}} - \tau^{1/2}\boldsymbol{G}^{\mathrm{L}}\|^{-1}$, we have the Neumann series identity

$$\left(I + \tau\boldsymbol{H}^{\mathrm{L}} - \tau^{3/2}\boldsymbol{G}^{\mathrm{L}}\right)^{-1} = \sum_{k=0}^{\infty}\left(-\tau\boldsymbol{H}^{\mathrm{L}} + \tau^{3/2}\boldsymbol{G}^{\mathrm{L}}\right)^k.$$

Note the right hand side consists only of (entrywise) positive matrices, so that

$$\mathbb{E}[e^n] \le \left(I + \tau\boldsymbol{H}^{\mathrm{L}} - \tau^{3/2}\boldsymbol{G}^{\mathrm{L}}\right)^{-1}\left(I - \tau\boldsymbol{H}^{\mathrm{U}} + \tau^{3/2}\boldsymbol{G}^{\mathrm{U}}\right)\mathbb{E}[e^{n-1}].$$

We prove in Lemma 5.2 that there exists $\rho, \tau_0 > 0$ s.t. if $\tau < \tau_0$, we have

$$\left\|\left(I + \tau\boldsymbol{H}^{\mathrm{L}} - \tau^{3/2}\boldsymbol{G}^{\mathrm{L}}\right)^{-1}\left(I - \tau\boldsymbol{H}^{\mathrm{U}} + \tau^{3/2}\boldsymbol{G}^{\mathrm{U}}\right)\right\|_2 \le 1 - \rho\tau.$$

So that viewing $\mathbb{E}[e^n]$ as a vector, we have

$$\|\mathbb{E}[e^n]\| \le (1-\rho\tau)\left\|\mathbb{E}[e^{n-1}]\right\| \;\Rightarrow\; \|\mathbb{E}[e^n]\| \le (1-\rho\tau)^n\left\|\mathbb{E}[e^0]\right\|.$$

This completes the proof. □

5.4.2 Proof of Theorem 5.3

Proof of Theorem 5.3. We first prove that π_ε is δ-localized with $\delta = \frac{2}{\lambda c}$. Notice

$$\nabla^2 \log \pi_\varepsilon(x) = -\lambda A^{\mathrm{T}} A - \nabla^2 \varphi_\varepsilon(x).$$

Since $A^{\mathrm{T}}A$ is c-diagonal block dominant, there exists a c-diagonal dominant matrix $M \in \mathbb{R}^{\mathsf{b}\times\mathsf{b}}$ s.t. $A^{\mathrm{T}}A(j,j) \succeq M_{jj} I_{d_j}$. By Lemma 5.3, $\nabla^2 \varphi_\varepsilon \succeq 0$, and thus

$$-\nabla^2_{jj} \log \pi_\varepsilon(x) \succeq \lambda M_{jj} I_{d_j}.$$

For $i \neq j$, note $\left\|A^{\mathrm{T}}A(i,j)\right\| \leq M_{ij}$, and by (5.4.17) in Lemma 5.3,

$$\begin{aligned}\left\|\nabla^2_{ij} \log \pi_\varepsilon(x)\right\| &\leq \lambda \left\|A^{\mathrm{T}}A(i,j)\right\| + \left\|\nabla^2_{ij}\varphi_\varepsilon(x)\right\| \\ &\leq M_{ij} + 4m\mu\varepsilon^{-1/2}\mathbf{1}_{i\sim j} + \mu\varepsilon^{-1/2}\mathbf{1}_{(i,j)\in\Gamma},\end{aligned}$$

where Γ is defined in (5.4.18). Therefore, denote $\tilde{M} \in \mathbb{R}^{\mathsf{b}\times\mathsf{b}}$ s.t.

$$\tilde{M}_{ij} = \lambda M_{ij} + 4m\mu\varepsilon^{-1/2}\mathbf{1}_{\mathsf{d_G}(i,j)=1} + \mu\varepsilon^{-1/2}\mathbf{1}_{(i,j)\in\Gamma},$$

and notice $\forall j \in [\mathsf{b}]$, since M is c-diagonal dominant, we have

$$\begin{aligned}\tilde{M}_{jj} - \sum_{i:i\neq j} \tilde{M}_{ij} &= \lambda\left(M_{jj} - \sum_{i:i\neq j} M_{ij}\right) - (16m+2)\mu\varepsilon^{-1/2} \\ &\geq \lambda c - (16m+2)\mu\varepsilon^{-1/2} \geq \lambda c/2.\end{aligned}$$

Here we use $\frac{\lambda}{\mu} \geq \frac{36m}{c\sqrt{\varepsilon}}$. By Theorem 3.2, π_ε is $\frac{2}{\lambda c}$-localized. So that by Theorem 3.3,

$$\max_j \mathsf{W}_1(\pi_j, \pi_{\varepsilon,j}) \leq \frac{2}{\lambda c} \max_i \mathbb{E}_\pi \left\|\nabla_i \log \pi_\varepsilon - \nabla_i \log \pi\right\|. \tag{5.4.7}$$

By (5.4.19) in Lemma 5.3, we have

$$\begin{aligned}\mathbb{E}_\pi \left\|\nabla_i \log \pi_\varepsilon - \nabla_i \log \pi\right\| &= \mathbb{E}_\pi \left\|\nabla_i \varphi_\varepsilon - \nabla_i \varphi_0\right\| \\ &\leq \mathbb{E}_\pi \sum_{t \text{ in } i} \mu \sum_{s:s\sim t} \left[1 - \frac{\|\boldsymbol{D}_s x\|}{\sqrt{\|\boldsymbol{D}_s x\|^2 + \varepsilon}}\right] \\ &\leq 4m^2\mu \max_s \mathbb{E}_\pi \left[1 - \frac{\|\boldsymbol{D}_s x\|}{\sqrt{\|\boldsymbol{D}_s x\|^2 + \varepsilon}}\right].\end{aligned} \tag{5.4.8}$$

Denote the function $I_s(t) := \mathbb{E}_\pi \left[1 - \frac{\|\boldsymbol{D}_s x\|}{\sqrt{\|\boldsymbol{D}_s x\|^2 + t^2}}\right]$. Then $I_s(0) = 0$ and

$$
\begin{aligned}
I_s'(t) = \mathbb{E}_\pi \left[\frac{\|\boldsymbol{D}_s x\| \, t}{\left(\|\boldsymbol{D}_s x\|^2 + t^2\right)^{3/2}}\right] &\leq \mathbb{E}_\pi \left[\frac{\|\boldsymbol{D}_s x\|^2 + t^2}{2\left(\|\boldsymbol{D}_s x\|^2 + t^2\right)^{3/2}}\right] \\
&\leq \frac{1}{2}\mathbb{E}_\pi \left[\left(\|\boldsymbol{D}_s x\|^2 + t^2\right)^{-1/2}\right] \leq \frac{1}{2}\mathbb{E}_\pi \|\boldsymbol{D}_s x\|^{-1}.
\end{aligned}
$$

By Lemma 5.4, there exists a dimension-independent constant C_π s.t.

$$
\max_s \mathbb{E}_\pi \|\boldsymbol{D}_s x\|^{-1} \leq C_\pi.
$$

This implies that $I_s'(t) \leq C_\pi/2 \Rightarrow I_s(t) \leq C_\pi t/2$. So that

$$
\max_s \mathbb{E}_\pi \left[1 - \frac{\|\boldsymbol{D}_s x\|}{\sqrt{\|\boldsymbol{D}_s x\|^2 + \varepsilon}}\right] \leq \frac{1}{2} C_\pi \varepsilon^{1/2}. \tag{5.4.9}
$$

Combine (5.4.7) (5.4.8) and (5.4.9), we have

$$
\max_j \mathsf{W}_1(\pi_j, \pi_{\varepsilon,j}) \leq \frac{2}{\lambda c} \cdot 4m^2\mu \cdot \frac{1}{2} C_\pi \varepsilon^{1/2} \leq \frac{1}{9} C_\pi m \varepsilon.
$$

Here we use $\frac{\lambda}{\mu} \geq \frac{36m}{c\sqrt{\varepsilon}}$. This completes the proof. □

5.4.3 Lemmas

Lemma 5.1. *Denote the function $a_j : \mathbb{R}^d \times \mathbb{R}^{d_j} \to \mathbb{R}$, s.t.*

$$
a_j(x, \xi) = \log \frac{\pi(z) Q_j(z_j, x_j \mid x_{-j})}{\pi(x) Q_j(x_j, z_j \mid x_{-j})}, \tag{5.4.10}
$$

where $z_j = x_j + \tau v_j(x) + \sqrt{2\tau}\xi$ and $z_{-j} = x_{-j}$. Under the assumptions in Theorem 5.1, it holds that for $0 < \tau \leq 1$, there exists M_i $(i = 1, 2, 3, 4)$ depending only on $\mathsf{M}, \mathsf{H}, \mathsf{L}$ and s, s.t.

$$
|a_j(x, \xi)| \leq \tau^{3/2}(\mathsf{M}_1 + \mathsf{M}_2 \|\xi\|^3). \tag{5.4.11}
$$

$$
\|\nabla_x a_j(x, \xi)\| \leq \tau(\mathsf{M}_3 + \mathsf{M}_4 \|\xi\|^2). \tag{5.4.12}
$$

Moreover, $\forall k \notin \mathcal{N}_j,\ \nabla_{x_k} a_j(x,\xi) = 0$. *As a result, we have*

$$|a_j(x,\xi) - a_j(y,\xi)| \le \tau(\mathsf{M}_3 + \mathsf{M}_4 \|\xi\|^2) \sum_{k\in\mathcal{N}_j} \|x_k - y_k\| . \tag{5.4.13}$$

Proof. By definition of a_j and note $z_j - x_j = \tau v_j(x) + \sqrt{2\tau}\xi$,

$$\begin{aligned} a_j(x,\xi) =&\ \log\pi(z) - \log\pi(x) - \frac{1}{4\tau}\left[\|x_j - z_j - \tau v_j(z)\|^2 - \|z_j - x_j - \tau v_j(x)\|^2\right] \\ =&\ \log\pi(z) - \log\pi(x) - \frac{\sqrt{2\tau}}{2}\langle v_j(z) + v_j(x), \xi\rangle - \frac{\tau}{4}\|v_j(z) + v_j(x)\|^2 . \end{aligned}$$

Note $z - x = (0,\dots,0, z_j - x_j, 0, \dots, 0)$. By Taylor expansion, we have

$$\begin{aligned} &\log\pi(z) - \log\pi(x) \\ =&\ v(x)^{\mathrm{T}}(z-x) + \frac{1}{2}\nabla v(x) : (z-x)^{\otimes 2} + \frac{1}{6}\nabla^2 v(\eta_1) : (z-x)^{\otimes 3} \\ =&\ v_j(x)^{\mathrm{T}}(z_j - x_j) + \frac{1}{2}\nabla_j v_j(x) : (z_j - x_j)^{\otimes 2} + \frac{1}{6}\nabla^2_{jj} v_j(\eta_1) : (z_j - x_j)^{\otimes 3} \\ =&\ \sqrt{2\tau} v_j(x)^{\mathrm{T}}\xi + \tau\|v_j(x)\|^2 + \tau\xi^{\mathrm{T}}\nabla_j v_j(x)\xi + \mathcal{O}(\tau^{3/2}). \end{aligned}$$

Similarly, we have

$$\begin{aligned} v_j(z) &= v_j(x) + \nabla_j v_j(x)(z_j - x_j) + \frac{1}{2}\nabla^2_{jj} v_j(\eta_2) : (z_j - x_j)^{\otimes 2} \\ &= v_j(x) + \sqrt{2\tau}\nabla_j v_j(x)\xi + \mathcal{O}(\tau). \end{aligned}$$

Therefore,

$$\frac{\sqrt{2\tau}}{2}\langle v_j(z) + v_j(x), \xi\rangle = \sqrt{2\tau} v_j(x)^{\mathrm{T}}\xi + \tau\xi^{\mathrm{T}}\nabla_j v_j(x)\xi + \mathcal{O}(\tau^{3/2}).$$

$$\frac{\tau}{4}\|v_j(z) + v_j(x)\|^2 = \tau\|v_j(x)\|^2 + \mathcal{O}(\tau^{3/2}).$$

Combine these equations, and notice $\mathcal{O}(\sqrt{\tau})$ and $\mathcal{O}(\tau)$ terms cancel out, we obtain

$$a(x,\xi) = \mathcal{O}(\tau^{3/2}).$$

And this term only depends on

$$v_j(x)^{\otimes 3},\ \xi^{\otimes 3},\ \nabla_j v_j(x),\ \nabla^2_{jj} v_j(\eta_1), \nabla^2_{jj} v_j(\eta_2).$$

Therefore, one can show that

$$|a(x,\xi)| \le \tau^{3/2}(\mathsf{M}_1 + \mathsf{M}_2 \|\xi\|^3).$$

This proves (5.4.11). For (5.4.12), notice $\nabla_x z = I + \tau \nabla v_j(x)$, and thus

$$\begin{aligned}
&\nabla_x a_j(x,\xi) \\
=\ & (I + \tau\nabla v_j(x))\, v(z) - v(x) - \frac{\sqrt{2\tau}}{2}\left[(I + \tau\nabla v_j(x))\,\nabla v_j(z) + \nabla v_j(x)\right]\xi \\
& - \frac{\tau}{2}\left[(I + \tau\nabla v_j(x))\,\nabla v_j(z) + \nabla v_j(x)\right](v_j(z) + v_j(x)) \\
=\ & v(z) - v(x) - \frac{1}{2}\left(\nabla v_j(z) + \nabla v_j(x)\right)\left(\tau v_j(x) + \sqrt{2\tau}\xi\right) \\
& + \tau\nabla v_j(x) v(z) - \frac{\sqrt{2\tau}}{2}\tau\nabla v_j(x)\nabla v_j(z)\xi \\
& - \frac{\tau}{2}\left[(I + \tau\nabla v_j(x))\,\nabla v_j(z) + \nabla v_j(x)\right] v_j(z) - \frac{\tau^2}{2}\nabla v_j(x)\nabla v_j(z) v_j(x).
\end{aligned}$$

Denote $y_t = (1-t)x + tz$ and $\boldsymbol{g}(t) = v(y_t)$, and we have

$$\boldsymbol{g}'(t) = (\nabla v(y_t))^{\mathrm{T}}(z - x) = \nabla v_j(y_t)(z_j - x_j).$$

Recall $z_j - x_j = \tau v_j(x) + \sqrt{2\tau}\xi$,

$$\begin{aligned}
& v(z) - v(x) - \frac{1}{2}\left(\nabla v_j(z) + \nabla v_j(x)\right)\left(\tau v_j(x) + \sqrt{2\tau}\xi\right) \\
=\ & \boldsymbol{g}(1) - \boldsymbol{g}(0) - \frac{1}{2}\left(\boldsymbol{g}'(1) + \boldsymbol{g}'(0)\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
& \left\| v(z) - v(x) - \frac{1}{2}\left(\nabla v_j(z) + \nabla v_j(x)\right)\left(\tau v_j(x) + \sqrt{2\tau}\xi\right)\right\| \\
=\ & \sup_{\|w\|=1}\left| w^{\mathrm{T}}\left[\boldsymbol{g}(1) - \boldsymbol{g}(0) - \frac{1}{2}\left(\boldsymbol{g}'(1) + \boldsymbol{g}'(0)\right)\right]\right| \\
\le\ & \sup_{\|w\|=1}\max_{t\in[0,1]}\frac{1}{12}\left|w^{\mathrm{T}}\boldsymbol{g}''(t)\right| \le \max_{t\in[0,1]}\frac{1}{12}\|\boldsymbol{g}''(t)\| \\
=\ & \frac{1}{12}\max_{t\in[0,1]}\left\|\nabla(\nabla_j v_j)(y_t) : (z_j - x_j)^{\otimes 2}\right\| \le \frac{\mathsf{L}}{12}\|z_j - x_j\|^2.
\end{aligned}$$

Here we use a classical numerical analysis result:

$$\left| f(1) - f(0) - \frac{1}{2}\left(f'(1) + f'(0)\right) \right| \leq \frac{1}{12} \max_{t\in[0,1]} |f''(t)| .$$

Finally, we have

$$\begin{aligned} \|\nabla_x a_j(x,\xi)\| &\leq \frac{\mathsf{L}}{12} \left\| \tau v_j(x) + \sqrt{2\tau}\xi \right\|^2 + \tau \left[\mathsf{HM} + \tau^{1/2}\mathsf{H}^2 \|\xi\| + \mathsf{HM} + \tau \mathsf{H}^2 \mathsf{M} \right] \\ &\leq \frac{\mathsf{L}}{6} \left(\tau^2 \mathsf{M}^2 + 2\tau \|\xi\|^2 \right) + \tau \left[2\mathsf{HM} + 2\mathsf{H}^2 \left(1 + \tau \|\xi\|^2 \right) + \tau \mathsf{H}^2 \mathsf{M} \right] . \end{aligned}$$

This proves (5.4.12). For the last claim, notice for $k \notin \mathcal{N}_j$,

$$\nabla_k v_j(x) = \nabla^2_{jk} \log \pi(x) \equiv 0.$$

Therefore, $\nabla_{x_k} z = \nabla_{x_k} x$, and we have

$$\nabla_{x_k} a_j(x,\xi) = v_k(z) - v_k(x) = \int_0^1 \nabla_j v_k(y_t)(z_j - x_j) \mathrm{d}t = 0.$$

Finally, since $a_j(\cdot,\xi)$ is only a function of $x_{\mathcal{N}_j}$, we have

$$\begin{aligned} |a_j(x,\xi) - a_j(y,\xi)| \leq\ & \|\nabla_x a_j(\eta,\xi)\| \left\| x_{\mathcal{N}_j} - y_{\mathcal{N}_j} \right\| \\ & \leq \tau(\mathsf{M}_3 + \mathsf{M}_4 \|\xi\|^2) \sum_{k\in\mathcal{N}_j} \|x_k - y_k\| . \end{aligned}$$

This completes the proof. □

Lemma 5.2. *Under the assumptions in Theorem 5.2, there exists $\rho, \tau_0 > 0$ s.t. if $\tau < \tau_0$, we have*

$$\left\| \left(I + \tau \boldsymbol{H}^{\mathrm{L}} - \tau^{3/2} \boldsymbol{G}^{\mathrm{L}} \right)^{-1} \left(I - \tau \boldsymbol{H}^{\mathrm{U}} + \tau^{3/2} \boldsymbol{G}^{\mathrm{U}} \right) \right\|_2 \leq 1 - \rho\tau.$$

where $\boldsymbol{H}^{\mathrm{L}}, \boldsymbol{H}^{\mathrm{U}}, \boldsymbol{G}^{\mathrm{L}}, \boldsymbol{G}^{\mathrm{U}}$ are defined in (5.4.6). Here ρ, τ_0 are independent of d.

Proof. Denote $\boldsymbol{G} = \boldsymbol{G}^{\mathrm{L}} + \boldsymbol{G}^{\mathrm{U}}$. By definition, $\forall v \in \mathbb{R}^{\mathsf{b}}$,

$$\|\boldsymbol{G} v\|^2 = \sum_j \left(\mathsf{C} \sum_{i:i\sim j} v_i \right)^2 \leq \mathsf{C}^2 \sum_j (1+\mathsf{s}) \sum_{i:i\sim j} v_i^2 \leq \mathsf{C}^2 (1+\mathsf{s})^2 \|v\|^2 .$$

So that $\|\boldsymbol{G}\|_2 \leq \mathsf{C}(1+\mathsf{s})$. Similarly, one can prove that $\|\boldsymbol{G}^{\mathrm{L}}\|_2, \|\boldsymbol{G}^{\mathrm{U}}\|_2 \leq \mathsf{C}(1+\mathsf{s})$.

With out loss of generality, we can take $\boldsymbol{H}_{ij} = 0$ if $i \not\sim j$. Similarly we have

$$\max\left\{\|\boldsymbol{H}\|_2,\ \left\|\boldsymbol{H}^{\mathrm{L}}\right\|_2, \left\|\boldsymbol{H}^{\mathrm{U}}\right\|_2\right\} \leq \max_{i,j} |\boldsymbol{H}_{ij}|\,(1+\mathsf{s}).$$

When $\tau \leq \frac{1}{2}\left\|\boldsymbol{H}^{\mathrm{L}} - \tau^{1/2}\boldsymbol{G}^{\mathrm{L}}\right\|^{-1}$, it holds that $\left\|\left(I + \tau\boldsymbol{H}^{\mathrm{L}} - \tau^{3/2}\boldsymbol{G}^{\mathrm{L}}\right)^{-1}\right\| \leq 2$, and

$$\begin{aligned}
&\left\|\left(I + \tau\boldsymbol{H}^{\mathrm{L}} - \tau^{3/2}\boldsymbol{G}^{\mathrm{L}}\right)^{-1} - \left(I - \tau\boldsymbol{H}^{\mathrm{L}} + \tau^{3/2}\boldsymbol{G}^{\mathrm{L}}\right)\right\| \\
\leq\ &\left\|\left(I + \tau\boldsymbol{H}^{\mathrm{L}} - \tau^{3/2}\boldsymbol{G}^{\mathrm{L}}\right)^{-1}\right\| \left\|I - \left(I + \tau\boldsymbol{H}^{\mathrm{L}} - \tau^{3/2}\boldsymbol{G}^{\mathrm{L}}\right)\left(I - \tau\boldsymbol{H}^{\mathrm{L}} + \tau^{3/2}\boldsymbol{G}^{\mathrm{L}}\right)\right\| \\
\leq\ &2\left\|\left(\tau\boldsymbol{H}^{\mathrm{L}} - \tau^{3/2}\boldsymbol{G}^{\mathrm{L}}\right)^2\right\| \\
\leq\ &2\tau^2\left[\left\|\boldsymbol{H}^{\mathrm{L}}\right\| + \tau^{1/2}\left\|\boldsymbol{G}^{\mathrm{L}}\right\|\right]^2.
\end{aligned}$$

So that

$$\begin{aligned}
&\left\|\left(I + \tau\boldsymbol{H}^{\mathrm{L}} - \tau^{3/2}\boldsymbol{G}^{\mathrm{L}}\right)^{-1}\left(I - \tau\boldsymbol{H}^{\mathrm{U}} + \tau^{3/2}\boldsymbol{G}^{\mathrm{U}}\right)\right\|_2 \\
\leq\ &\left\|\left(I - \tau\boldsymbol{H}^{\mathrm{L}} + \tau^{3/2}\boldsymbol{G}^{\mathrm{L}}\right)\left(I - \tau\boldsymbol{H}^{\mathrm{U}} + \tau^{3/2}\boldsymbol{G}^{\mathrm{U}}\right)\right\|_2 \\
&+ 2\tau^2\left[\left\|\boldsymbol{H}^{\mathrm{L}}\right\| + \tau^{1/2}\left\|\boldsymbol{G}^{\mathrm{L}}\right\|\right]^2\left\|I - \tau\boldsymbol{H}^{\mathrm{U}} + \tau^{3/2}\boldsymbol{G}^{\mathrm{U}}\right\|_2 \\
\leq\ &\left\|I - \tau\boldsymbol{H} + \tau^{3/2}\boldsymbol{G}\right\|_2 + \left\|\left(-\tau\boldsymbol{H}^{\mathrm{L}} + \tau^{3/2}\boldsymbol{G}^{\mathrm{L}}\right)\left(-\tau\boldsymbol{H}^{\mathrm{U}} + \tau^{3/2}\boldsymbol{G}^{\mathrm{U}}\right)\right\|_2 \\
&+ 2\tau^2\left[\left\|\boldsymbol{H}^{\mathrm{L}}\right\| + \tau^{1/2}\left\|\boldsymbol{G}^{\mathrm{L}}\right\|\right]^2\left[1 + \tau\left\|\boldsymbol{H}^{\mathrm{U}}\right\| + \tau^{3/2}\left\|\boldsymbol{G}^{\mathrm{U}}\right\|\right] \\
\leq\ &\|I - \tau\boldsymbol{H}\|_2 + \tau^{3/2}\|\boldsymbol{G}\|_2 + \mathcal{O}(\tau^2).
\end{aligned}$$

Note here $\|\boldsymbol{G}\|_2$ and $\mathcal{O}(\tau^2)$ are dimension independent. So that by taking

$$\tau \leq \frac{1}{2}\|\boldsymbol{H}^{\mathrm{L}} - \tau^{1/2}\boldsymbol{G}^{\mathrm{L}}\|^{-1},\ \tau < \|\boldsymbol{H}\|^{-1},\ \tau^{3/2}\|\boldsymbol{G}\|_2 + \mathcal{O}(\tau^2) \leq \frac{1}{2}\tau\lambda_{\boldsymbol{H}}, \quad (5.4.14)$$

We obtain that

$$\left\|\left(I + \tau\boldsymbol{H}^{\mathrm{L}} - \tau^{3/2}\boldsymbol{G}^{\mathrm{L}}\right)^{-1}\left(I - \tau\boldsymbol{H}^{\mathrm{U}} + \tau^{3/2}\boldsymbol{G}^{\mathrm{U}}\right)\right\|_2 \leq 1 - \frac{1}{2}\tau\lambda_{\boldsymbol{H}}.$$

Note the requirements on τ are independent of d. So that one can find some $\tau_0 > 0$ independent of d s.t. (5.4.14) holds if $\tau < \tau_0$. And we can take $\rho = \frac{1}{2}\lambda_{\boldsymbol{H}}$. □

Lemma 5.3. *The derivatives of* (5.3.3) *are given by*

$$\nabla\varphi_\varepsilon(x) = \mu\sum_{s=1}^{d}\frac{(D_s^{(v)}x)(D_s^{(v)})^{\mathrm{T}} + (D_s^{(h)}x)(D_s^{(h)})^{\mathrm{T}}}{\left((D_s^{(v)}x)^2 + (D_s^{(h)}x)^2 + \varepsilon\right)^{1/2}}. \tag{5.4.15}$$

$$\begin{aligned}&\nabla^2\varphi_\varepsilon(x)\\ &= \mu\sum_{s=1}^{d}\left((D_s^{(v)}x)^2 + (D_s^{(h)}x)^2 + \varepsilon\right)^{-3/2}\cdot\Big[\varepsilon\left((D_s^{(v)})^{\mathrm{T}}D_s^{(v)} + (D_s^{(h)})^{\mathrm{T}}D_s^{(h)}\right)\\ &\quad + \left((D_s^{(v)}x)D_s^{(v)} - (D_s^{(h)}x)D_s^{(h)}\right)^{\mathrm{T}}\left((D_s^{(v)}x)D_s^{(v)} - (D_s^{(h)}x)D_s^{(h)}\right)\Big].\end{aligned} \tag{5.4.16}$$

The following results hold: $\nabla^2\varphi_\varepsilon(x) \succeq 0$ *and for* $i \neq j$,

$$\left\|\nabla^2_{ij}\log\varphi_\varepsilon(x)\right\|_{\mathrm{F}} \le 4m\mu\varepsilon^{-1/2}\mathbf{1}_{i\sim j} + \mu\varepsilon^{-1/2}\mathbf{1}_{(i,j)\in\Gamma}. \tag{5.4.17}$$

Here Γ *denotes the pairs of blocks in 'antidiagonal' position, i.e.*

$$\Gamma := \{(i,j) : \mathsf{d}_{\mathsf{G}}(i,j) = 2, x_i + y_j = x_j + y_i\}, \tag{5.4.18}$$

where (x_i, y_i) *denotes the* x, y *indices of block* i*. The gradient estimate holds:*

$$\forall t\in[d],\quad |\partial_t\varphi_\varepsilon(x) - \partial_t\varphi_0(x)| \le \mu\sum_{s:s\sim t}\left[1 - \frac{\|\boldsymbol{D}_s x\|}{\sqrt{\|\boldsymbol{D}_s x\|^2 + \varepsilon}}\right]. \tag{5.4.19}$$

where we denote

$$\boldsymbol{D}_s x = (D_s^{(v)}x, D_s^{(h)}x))^{\mathrm{T}} \in \mathbb{R}^2. \tag{5.4.20}$$

Proof. (5.4.15) and (5.4.16) are obtained by direct computation. $\nabla^2\varphi_\varepsilon(x) \succeq 0$ directly follows from (5.4.16). For (5.4.17), first consider $\mathsf{d}_{\mathsf{G}}(i,j) = 1$. If i, j are vertical neighbors, only those s located in the boundary of the upper block

contribute to $\nabla^2_{ij}\varphi_\varepsilon$. Therefore,

$$
\begin{aligned}
&\left\|\nabla^2_{ij}\varphi_\varepsilon(x)\right\|_{\mathrm{F}} \\
\leq\ & \mu \sum_{s \text{ in boundary}} \left(\|\boldsymbol{D}_s x\|^2 + \varepsilon\right)^{-3/2} \cdot \Big[\varepsilon \|D^{(v)}_{s,i}\| \|D^{(v)}_{s,j}\| \\
&+ \left(|D^{(v)}_s x| \|D^{(v)}_{s,i}\| + |D^{(h)}_s x| \|D^{(h)}_{s,i}\|\right)\left(|D^{(v)}_s x| \|D^{(v)}_{s,j}\| + |D^{(h)}_s x| \|D^{(h)}_{s,j}\|\right)\Big] \\
\leq\ & \mu \sum_{s \text{ in boundary}} \left(\|\boldsymbol{D}_s x\|^2 + \varepsilon\right)^{-3/2} \cdot \left[2\varepsilon + 2\left(|D^{(v)}_s x| + |D^{(h)}_s x|\right)^2\right] \\
\leq\ & \mu \sum_{s \text{ in boundary}} \left(\|\boldsymbol{D}_s x\|^2 + \varepsilon\right)^{-3/2} \cdot \left(2\varepsilon + 4\|\boldsymbol{D}_s x\|^2\right) \\
\leq\ & \mu \sum_{s \text{ in boundary}} 4\varepsilon^{-1/2} = 4m\mu\varepsilon^{-1/2}.
\end{aligned}
$$

For $(i,j) \in \Gamma$, there is exactly one s that contributes to $\nabla^2_{ij}\varphi_\varepsilon(x)$, and the only non-zero term is the cross term of the vertical and horizontal differences. So that

$$
\begin{aligned}
\left\|\nabla^2_{ij}\varphi_\varepsilon(x)\right\|_{\mathrm{F}} &\leq \mu \left(\|\boldsymbol{D}_{s_*} x\|^2 + \varepsilon\right)^{-3/2} |D^{(v)}_s x| |D^{(h)}_s x| \|D^{(v)}_s\| \|D^{(h)}_s\| \\
&\leq \mu \left(\|\boldsymbol{D}_{s_*} x\|^2 + \varepsilon\right)^{-3/2} \frac{1}{2}\|\boldsymbol{D}_{s_*} x\|^2 \cdot 2 \leq \mu\varepsilon^{-1/2}.
\end{aligned}
$$

This proves (5.4.17). For (5.4.19), simply note that

$$
\begin{aligned}
&|\partial_t \varphi_\varepsilon(x) - \partial_t \varphi_0(x)| \\
=\ & \mu \sum_{s:s\sim t} \left|(D^{(v)}_s x) D^{(v)}_{s,t} + (D^{(h)}_s x) D^{(h)}_{s,t}\right| \cdot \left|\left(\|\boldsymbol{D}_s x\|^2 + \varepsilon\right)^{-1/2} - \|\boldsymbol{D}_s x\|^{-1}\right| \\
\leq\ & \mu \sum_{s:s\sim t} \|\boldsymbol{D}_s x\| \cdot \left|\left(\|\boldsymbol{D}_s x\|^2 + \varepsilon\right)^{-1/2} - \|\boldsymbol{D}_s x\|^{-1}\right| \\
=\ & \mu \sum_{s:s\sim t} \left[1 - \frac{\|\boldsymbol{D}_s x\|}{\sqrt{\|\boldsymbol{D}_s x\|^2 + \varepsilon}}\right].
\end{aligned}
$$

This completes the proof. □

Lemma 5.4. *There exists a dimension-independent constant C_π s.t.*

$$
\max_s \mathbb{E}_{x\sim\pi} \|\boldsymbol{D}_s x\|_2^{-1} \leq C_\pi.
$$

Proof. Fix any $s \in [d]$. For simplicity, denote $x_s^{(v)}, x_s^{(h)}$ s.t.

$$\boldsymbol{D}_s x = (D_s^{(v)} x, D_s^{(h)} x) = (x_s - x_s^{(v)}, x_s - x_s^{(h)}).$$

Now introduce the change of variable $z = Tx$ for some linear map determined via

$$z_s^{(v)} = x_s^{(v)} - x_s, \quad z_s^{(h)} = x_s^{(h)} - x_s,$$

and $z_{s-} = x_{s-} \in \mathbb{R}^{d-2}$ for the other coordinates. Accordingly, $\pi(x)$ is transformed into another distribution $\tilde{\pi}(z) = \pi(T^{-1}z)$ (note $\det(T) = 1$). Also note that T^{-1} admits an explicit form, i.e.

$$x_s^{(v)} = z_s^{(v)} + z_s, \quad x_s^{(h)} = z_s^{(h)} + z_s, \quad x_{s-} = z_{s-}.$$

Denote $z_{s+} = (z_s^{(v)}, z_s^{(h)})$ for convenience. Consider the factorization

$$\tilde{\pi}(z) = \tilde{\pi}(z_{s+}, z_{s-}) = \tilde{\pi}(z_{s+}|z_{s-})\tilde{\pi}(z_{s-}),$$

where $\tilde{\pi}(z_{s-})$ denotes the marginal of $\tilde{\pi}$ on z_{s-}. Notice

$$\log \tilde{\pi}(z_{s+}|z_{s-}) = -\frac{\lambda}{2} \left\| y - AT^{-1}z \right\|_2^2 - \mu \sum_{s=1}^{d} \left\| \boldsymbol{D}_s (T^{-1}z) \right\|_2 - \log \tilde{\pi}(z_{s-}) + \text{const.}$$

Fix z_{s-} for the moment. Notice $\frac{\lambda}{2} \left\| y - AT^{-1}z \right\|_2^2$ is L-smooth for some dimension-independent $L > 0$, since only a dimension-independent number of coordinates of $AT^{-1}z$ depend on z_{s-}. Therefore, fix any z_{s+}^0, so that

$$\frac{\lambda}{2} \left\| y - AT^{-1}(z_{s+}, z_{s-}) \right\|_2^2 - \frac{\lambda}{2} \left\| y - AT^{-1}(z_{s+}^0, z_{s-}) \right\|_2^2 - v^0 \cdot (z_{s+} - z_{s+}^0) \\ \leq \frac{L}{2} \left\| z_{s+} - z_{s+}^0 \right\|_2^2,$$

where v^0 is the gradient w.r.t. z_{s+} of $\frac{\lambda}{2} \left\| y - AT^{-1}z \right\|_2^2$ at z_{s+}^0. Notice also

$$\mu \sum_{s=1}^{d} \left\| \boldsymbol{D}_s (T^{-1}(z_{s+}, z_{s-})) \right\|_2 \leq \mu \sum_{s=1}^{d} \left\| \boldsymbol{D}_s (T^{-1}(z_{s+}^0, z_{s-})) \right\|_2 + 8\mu \left\| z_{s+} - z_{s+}^0 \right\|_2,$$

since changing $z_s^{(v)}$ or $z_s^{(h)}$ affects 4 finite difference terms in the summation. Combining the above controls, when $z_{s+} \in B_1(z_{s+}^0) := \{z_{s+} : \left\| z_{s+} - z_{s+}^0 \right\|_2 \leq 1\}$, it

holds that

$$\begin{aligned}&\log\tilde{\pi}(z_{s+}|z_{s-})-\log\tilde{\pi}(z_{s+}^0|z_{s-})+v^0\cdot(z_{s+}-z_{s+}^0)\\ \geq\ &-\frac{L}{2}\left\|z_{s+}-z_{s+}^0\right\|_2^2-8\mu\left\|z_{s+}-z_{s+}^0\right\|_2\geq-\frac{L}{2}-8\mu.\end{aligned}$$

Therefore,

$$\begin{aligned}1=&\int\tilde{\pi}(z_{s+}|z_{s-})\mathrm{d}z_{s+}\\ \geq&\int_{B_1(z_{s+}^0)}\exp\left(\log\tilde{\pi}(z_{s+}|z_{s-})\right)\mathrm{d}z_{s+}\\ \geq&\,\tilde{\pi}(z_{s+}^0|z_{s-})\exp\left(-\frac{L}{2}-8\mu\right)\int_{B_1(z_{s+}^0)}\exp\left(-v^0\cdot(z_{s+}-z_{s+}^0)\right)\mathrm{d}z_{s+}\\ \geq&\,\tilde{\pi}(z_{s+}^0|z_{s-})\exp\left(-\frac{L}{2}-8\mu\right)|B_1|,\end{aligned}$$

where we use the Jensen's inequality and the symmetry of $B_1(z_{s-}^0)$. Therefore,

$$\tilde{\pi}(z_{s+}^0|z_{s-})\leq|B_1|^{-1}\exp\left(\frac{L}{2}+8\mu\right).$$

This holds for arbitrary z_{s+}^0, so that the marginal distribution of $\tilde{\pi}(z_{s+})$ satisfies

$$\tilde{\pi}(z_{s+})=\int\tilde{\pi}(z_{s+}|z_{s-})\tilde{\pi}(z_{s-})\mathrm{d}z_{s-}\leq|B_1|^{-1}\exp\left(\frac{L}{2}+8\mu\right)=:C'.$$

Note C' is dimension-independent. Finally, notice

$$\begin{aligned}\mathbb{E}_{x\sim\pi}\left\|\boldsymbol{D}_s x\right\|_2^{-1}=&\int_{\mathbb{R}^2}\frac{\tilde{\pi}(z_{s+})}{\left\|z_{s+}\right\|_2}\mathrm{d}z_s^{(v)}\mathrm{d}z_s^{(h)}\\ \leq&\,1+\int_{\|z_{s+}\|_2\leq1}\frac{C'}{\left\|z_{s+}\right\|_2}\mathrm{d}z_s^{(v)}\mathrm{d}z_s^{(h)}\\ =&\,1+\int_0^{2\pi}\int_0^1\frac{C'}{r}\cdot r\mathrm{d}r\mathrm{d}\theta\\ =&\,1+2\pi C'=:C_\pi.\end{aligned}$$

Thus, C_π is dimension-independent. This completes the proof. □

Chapter 6

Localized Diffusion Models

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). In this chapter, we consider exploiting the locality structure to circumvent the CoD in diffusion models. 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.

6.1 Localized diffusion models

6.1.1 Review on 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 π_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 initialized with the target distribution π_0, i.e.,

$$\mathrm{d}X_t = -X_t\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \quad X_0 \sim \pi_0. \tag{6.1.1}$$

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 [102]

$$\mathrm{d}Y_t = (Y_t + 2\nabla \log \pi_{T-t}(Y_t))\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \quad Y_0 \sim \pi_T, \tag{6.1.2}$$

where we denote $\pi_t = \mathsf{Law}(X_t)$. The target distribution π_0 can then be sampled by first sampling $Y_0 \sim \pi_T$ and then evolving Y_t according to (6.1.2) to obtain a sample $Y_T \sim \pi_0$.

To implement the above scheme, several approximations are needed:

1. Score estimation. The score function $s(x,t) := \nabla \log \pi_t(x)$ is not accessible, and needs to be estimated from the data via the denoising score matching scheme [114, 101, 61]

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

$$\mathcal{L}(s_\theta) := \int_0^T \mathbb{E}_{x_0\sim\pi_0}\left[\mathbb{E}_{x_t\sim\pi_{t|0}(x_t|x_0)}\left[\left\|s_\theta(x_t,t) - \nabla_{x_t}\log\pi_{t|0}(x_t|x_0)\right\|^2\right]\right]\mathrm{d}t. \tag{6.1.3}$$

   In the sampling process, the true score $\nabla \log \pi_{T-t}(Y_t)$ in (6.1.2) is approximated by the estimated score $\widehat{s}(Y_t, T-t)$.

2. Approximation of π_T. The initial distribution π_T in the reverse process is intractable. But since the OU process converges exponentially to $\pi_\infty = \mathsf{N}(0,I)$, we can approximate π_T by $\mathsf{N}(0,I)$ in (6.1.2), 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 $Y_{T-\underline{t}} \sim \pi_{\underline{t}}$, which should be close to π_0 when $\underline{t}$ is small.

4. Time discretization. The Euler-Maruyama scheme is used to discretize (6.1.2). Pick time steps $0 = t_0 < t_1 < \cdots < t_N = T - \underline{t}$, and evolve $n = 0, 1, \ldots, N-1$

by

$$Y_{t_{n+1}} = Y_{t_n} + (Y_{t_n} + 2\widehat{s}(Y_{t_n}, T - t_n))\,\Delta t_n + \sqrt{2\Delta t_n}\xi_n, \tag{6.1.4}$$

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

$$\pi_{t|0}(x_t|x_0) = \mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I), \quad \alpha_t := \mathrm{e}^{-t}, \quad \sigma_t := \sqrt{1 - \mathrm{e}^{-2t}}. \tag{6.1.5}$$

So that $\nabla_{x_t} \log \pi_{t|0}(x_t|x_0) = -\sigma_t^{-2}(x_t - \alpha_t x_0)$, and $\pi_{t|0}(x_t|x_0)$ 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 (6.1.3) can be written as

$$\mathcal{L}(s_\theta) = \int_{\underline{t}}^{T} \mathbb{E}_{x_0 \sim \pi_0} \mathbb{E}_{\epsilon_t \sim \mathsf{N}(0,I)} \left[\left\| s_\theta(\alpha_t x_0 + \sigma_t \epsilon_t, t) + \sigma_t^{-1} \epsilon_t \right\|^2 \right] \mathrm{d}t, \tag{6.1.6}$$

where we involve the early stopping truncation. The above loss provides a convenient form for implementation [61].

6.1.2 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 (6.1.5) of the OU process implies that π_t has an explicit density

$$\pi_t(x_t) = \int \mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I) \pi_0(x_0) \mathrm{d}x_0.$$

π_t can be viewed as an interpolation between π_0 and $\pi_\infty = \mathsf{N}(0, I)$. Suppose π_0 is a localized distribution on an undirected graph G. It is obvious that π_∞ is localized, but their interpolation π_t may not remain strictly localized. However, π_t is still approximately localized, as proved in the following theorem.

Theorem 6.1. *Suppose π_0 has dependency graph G and is log-concave and smooth, i.e. $\exists 0 < m \le M < \infty$ s.t. $mI \preceq -\nabla^2 \log \pi(x) \preceq MI$. Then for any $t \in (0, T]$, π_t*

is approximately localized on G. *Specifically,*

$$\left\|\nabla_{ij}^2 \log \pi_t\right\|_\infty \le \frac{\alpha_t^2}{\sigma_t^2\left(m\sigma_t^2+\alpha_t^2\right)}\left(1-\frac{m\sigma_t^2+\alpha_t^2}{M\sigma_t^2+\alpha_t^2}\right)^{\mathsf{d}_{\mathsf{G}}(i,j)}. \tag{6.1.7}$$

Here $\alpha_t = \mathrm{e}^{-t}$ *and* $\sigma_t = \sqrt{1-\mathrm{e}^{-2t}}$ *(cf. (6.1.5)).*

Remark 6.1. (1) While Theorem 6.1 assumes log-concavity to apply Theorem 3.5, the exponential decay of correlations is ubiquitous for distributions with locality structure [68, 92, 33] and does not inherently depend on log-concavity. The assumption is adopted here for simplicity and to derive an explicit bound.
(2) Here we assume that π_0 is localized for technical convenience. In practice, many distributions of interest, such as image distributions, are typically only approximately localized. We believe that our results can be extended to this setting. Due to the additional technical challenges, we leave this extension to future work.

The proof is based on the observation that

$$\nabla_{ij}^2 \log \pi_t(x_t) = \alpha_t^2\sigma_t^{-4}\mathsf{Cov}_{\pi_{0|t}(x_0|x_t)}\left(x_{0,i}, x_{0,j}\right).$$

So that the result directly follows Theorem 3.5. Detailed proofs are delayed to Section 6.4.1.

6.1.3 Localized hypothesis space

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

$$\mathscr{H}_r = \left\{ s_\theta : \mathbb{R}^{d+1} \to \mathbb{R}^d \mid s_{\theta,j}(x,t) = u_{\theta,j}(x_{\mathcal{N}_j^r}, t), u_{\theta,j} \in \mathscr{U}_j, j \in [\mathsf{b}] \right\}, \tag{6.1.8}$$

where r denotes the localization radius, $\mathcal{N}_j^r$ is the extended neighborhood (2.1.2), 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 π_t in light of Theorem 6.1.

Define the *effective dimension* of s_θ as

$$d_{\text{eff}} := \max_j d_{j,r}, \quad d_{j,r} := \sum_{i\in\mathcal{N}_j^r} d_i. \tag{6.1.9}$$

Since $s_\theta(\cdot,t)$ can be viewed as a collection of functions $\{u_{\theta,j}(\cdot,t) : \mathbb{R}^{d_{j,r}} \to \mathbb{R}^{d_j}\}_{j\in[\mathsf{b}]}$,

it is essentially a function of d_{eff} variables. For local graph, $d_{\text{eff}} \ll d$, so that intuitively estimating s_θ in $\mathscr{H}_r$ does not suffer from the CoD.

ReLU neural network In practice, $\mathscr{H}_r$ can be realized by a neural network (NN) with locality constraints. Here we consider the widely used ReLU NN class. We note that our method and analysis result apply to other types of NNs as well. Following [84], 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 $(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B})$:

$$\begin{aligned} \mathsf{NN}(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B}) &= \{u_\theta : \mathbb{R}^{\mathsf{w}_0} \to \mathbb{R}^{\mathsf{w}_\mathsf{L}} \mid \theta \in \Theta(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B})\}, \\ \Theta(\mathsf{L}, \mathsf{W}, \mathsf{S}, \mathsf{B}) &= \left\{\theta = \{W_l, b_l\}_{l=1}^{\mathsf{L}} \mid W_l \in \mathbb{R}^{\mathsf{w}_l \times \mathsf{w}_{l-1}}, b_l \in \mathbb{R}^{\mathsf{w}_l}, \|\theta\|_0 \le \mathsf{S}, \|\theta\|_\infty \le \mathsf{B}\right\}, \\ u_\theta(x) &= W_\mathsf{L}\sigma(W_{\mathsf{L}-1}\sigma(\cdots\sigma(W_1 x + b_1)\cdots) + b_{\mathsf{L}-1}) + b_\mathsf{L}, \end{aligned} \tag{6.1.10}$$

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:

$$\mathscr{U}_j = \mathsf{NN}(\mathsf{L}^j, \mathsf{W}^j, \mathsf{S}^j, \mathsf{B}^j), \quad \text{where } \mathsf{w}_0^j = d_{j,r} + 1,\ \mathsf{w}_\mathsf{L}^j = d_j. \tag{6.1.11}$$

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

6.1.4 Localized denoising score matching

Given the hypothesis space $\mathscr{H}_r$ (6.1.8) with localized NN score $\mathscr{U}_j$ (6.1.11), we can learn the localized score function by minimizing the denoising score matching loss (6.1.6). Given i.i.d. sample $\{X^{(i)}\}_{i=1}^N$ from π_0, the population loss (6.1.6) is approximated by the empirical loss, i.e.,

$$\widehat{s} = \arg\min_{s_\theta \in \mathscr{H}_r} \widehat{\mathcal{L}}_N(s_\theta), \tag{6.1.12}$$

with

$$\widehat{\mathcal{L}}_N(s_\theta) = \frac{1}{N}\sum_{i=1}^{N}\int_{\underline{t}}^{T}\mathbb{E}_{\epsilon_t\sim\mathsf{N}(0,I)}\left[\left\|s_\theta(\alpha_t X^{(i)}+\sigma_t\epsilon_t,t)+\sigma_t^{-1}\epsilon_t\right\|^2\right]\mathrm{d}t. \tag{6.1.13}$$

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

$$\widehat{\mathcal{L}}_{j,N}(u_{\theta,j}) = \frac{1}{N}\sum_{i=1}^{N}\int_{\underline{t}}^{T}\mathbb{E}_{\epsilon_t\sim\mathsf{N}(0,I)}\left[\left\|u_{\theta,j}(\alpha_t X^{(i)}_{\mathcal{N}_j^r}+\sigma_t\epsilon_{t,\mathcal{N}_j^r},t)+\sigma_t^{-1}\epsilon_{t,j}\right\|^2\right]\mathrm{d}t. \tag{6.1.14}$$

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

$$\widehat{u}_j = \underset{u_{\theta,j}\in\mathscr{U}_j}{\arg\min}\,\widehat{\mathcal{L}}_{j,N}(u_{\theta,j}). \tag{6.1.15}$$

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

6.2 Analysis of localized diffusion models

6.2.1 Error decomposition

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

$$\mathrm{d}\widehat{Y}_t = \left(\widehat{Y}_t + 2\widehat{s}(\widehat{Y}_t, T-t)\right)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \quad \widehat{Y}_0\sim\mathsf{N}(0,I). \tag{6.2.1}$$

And we take the early stopped distribution $\widehat{\mu}_{T-\underline{t}} = \mathsf{Law}(\widehat{Y}_{T-\underline{t}})$ as the approximation of π_0. It suffices to consider the error between $\widehat{\mu}_{T-\underline{t}}$ and $\pi_{\underline{t}}$, as it is easier to control the early stopping error, i.e., the distance between $\pi_{\underline{t}}$ and π_0. The following error decomposition is standard [22].

Proposition 6.1. *Under Novikov's condition [22]:*

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

where $\mathsf{Q}=\mathsf{Law}(Y_{[0,T-\underline{t}]})$ *denotes the path measure of the reverse process* (6.1.2)*. It holds that*

$$\mathsf{KL}(\pi_{\underline{t}}\|\widehat{\mu}_{T-\underline{t}})\le \mathrm{e}^{-2T}\mathsf{KL}(\pi_0\|\mathsf{N}(0,I))+\int_{\underline{t}}^{T}\mathbb{E}_{x_t\sim\pi_t}\left[\left\|\widehat{s}(x_t,t)-s(x_t,t)\right\|^2\right]\mathrm{d}t. \quad (6.2.2)$$

Proofs are delayed to Section 6.4.2. We note that the first term on the right hand side can be replaced by $\mathrm{e}^{-2(T-\underline{t})}\mathsf{KL}(\pi_{\underline{t}}\|\mathsf{N}(0,I))$ when π_0 is singular w.r.t. $\mathsf{N}(0,I)$, so that it always decays exponentially in T regardless of π_0. Thus it suffices to control the second term; i.e. the score approximation error.

6.2.2 Localized score function

As discussed in Section 6.1.2, strict locality is not preserved in the forward OU process, so that in general, the true score $s\notin\mathscr{H}_r$. 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 $\mathscr{H}_r$ (6.1.8), 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 (6.2.2), we consider the optimal approximation in the $L^2(\pi_t)$ sense, i.e.,

$$\begin{aligned}s^*&=\mathop{\arg\min}_{s_\theta\in\mathscr{H}_r^*}\int_{\underline{t}}^{T}\int\left\|s_\theta(x,t)-s(x,t)\right\|^2\pi_t(x)\mathrm{d}x\mathrm{d}t\\&\Leftrightarrow\ \forall j\in[\mathsf{b}],\quad s_j^*(x,t)=u_j^*(x_{\mathcal{N}_j^r},t),\\ \text{where}\quad u_j^*&=\mathop{\arg\min}_{u_{\theta,j}\in\mathscr{U}_j}\int_{\underline{t}}^{T}\int\left\|u_{\theta,j}(x_{\mathcal{N}_j^r},t)-s_j(x,t)\right\|^2\pi_t(x)\mathrm{d}x\mathrm{d}t.\end{aligned}$$

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

the optimizer is

$$\begin{aligned} u_j^*(x_{\mathcal{N}_j^r}, t) &= \mathbb{E}_{x' \sim \pi_t}\left[s_j(x', t) \middle| x'_{\mathcal{N}_j^r} = x_{\mathcal{N}_j^r} \right] \\ &= \frac{1}{\pi_t(x_{\mathcal{N}_j^r})} \int \nabla_j \log \pi_t(x_{\mathcal{N}_j^r}, x_{-\mathcal{N}_j^r}) \pi_t(x_{\mathcal{N}_j^r}, x_{-\mathcal{N}_j^r}) \mathrm{d}x_{-\mathcal{N}_j^r}. \end{aligned} \tag{6.2.3}$$

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

Due to the approximate locality (Theorem 6.1), one can expect that the approximation error decays exponentially with the radius r. Precisely, we prove the following theorem.

Theorem 6.2. *Let π_0 satisfy the conditions in Theorem 6.1, and its dependency graph is (s, ν)-local. Consider the hypothesis space $\mathscr{H}_r^*$ (6.1.8) 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*

$$\int_{\underline{t}}^{T} \left\| s_j^*(x,t) - s_j(x,t) \right\|_{L^2(\pi_t)}^2 \mathrm{d}t \le C d_j (r+1)^{\nu} \mathrm{e}^{-c(r+1)}, \tag{6.2.4}$$

where C and c are some dimensional independent constants depending on m, M, s, ν:

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

Note (6.2.4) is independent of $\underline{t}, T$. Moreover, for any $s_\theta \in \mathscr{H}_r^*$, the Pythagorean equality holds

$$\left\| s_{\theta,j}(x,t) - s_j(x,t) \right\|_{L^2(\pi_t)}^2 = \left\| s_{\theta,j}(x,t) - s_j^*(x,t) \right\|_{L^2(\pi_t)}^2 + \left\| s_j^*(x,t) - s_j(x,t) \right\|_{L^2(\pi_t)}^2. \tag{6.2.5}$$

The proof can be found in Section 6.4.3. (6.2.4) provides an upper bound for the hypothesis error of using a localized score function to approximate the true score function. Note although the true score $s_j(x,t)$ is a d-dimensional function, the bound is *independent* of the ambient dimension d. 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 $j \in [\mathsf{b}]$ in (6.2.4) gives the total approximation error

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

which scales linearly with the dimension d.

6.2.3 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 (6.2.3). Since the localized scores are low-dimensional functions, the sample complexity should be independent of d.

Equivalent to diffusion models for marginals A key observation is that the localized denoising score matching loss (6.1.14) is *equivalent* to the j-th component loss of the score function when we use standard diffusion model to approximate the *marginal distribution* $\pi_0(x_{\mathcal{N}_j^r})$. To be precise, denote its population version as

$$\mathcal{L}_j(u_{\theta,j}) = \mathbb{E}_{x_0\sim\pi_0}\int_{\underline{t}}^{T}\mathbb{E}_{\epsilon_t\sim\mathsf{N}(0,I)}\left[\left\|u_{\theta,j}(\alpha_t x_{0,\mathcal{N}_j^r}+\sigma_t\epsilon_{t,\mathcal{N}_j^r},t)+\sigma_t^{-1}\epsilon_{t,j}\right\|^2\right]\mathrm{d}t. \tag{6.2.6}$$

The following proposition shows the equivalence.

Proposition 6.2. *The following equalities hold:*

$$\begin{aligned}
&\mathcal{L}_j(u_{\theta,j})\\
&=\mathbb{E}_{x_{0,\mathcal{N}_j^r}\sim\pi_0}\int_{\underline{t}}^{T}\mathbb{E}_{\epsilon_t\sim\mathsf{N}(0,I)}\left[\left\|u_{\theta,j}(\alpha_t x_{0,\mathcal{N}_j^r}+\sigma_t\epsilon_{t,\mathcal{N}_j^r},t)+\sigma_t^{-1}\epsilon_{t,j}\right\|^2\right]\mathrm{d}t\\
&=\mathbb{E}_{x_{0,\mathcal{N}_j^r}\sim\pi_0}\int_{\underline{t}}^{T}\mathbb{E}_{x_{t,\mathcal{N}_j^r}\sim\pi_{t|0}(x_{t,\mathcal{N}_j^r}|x_{0,\mathcal{N}_j^r})}\left[\left\|u_{\theta,j}(x_{t,\mathcal{N}_j^r},t)-\nabla_j\log\pi_{t|0}(x_{t,\mathcal{N}_j^r}|x_{0,\mathcal{N}_j^r})\right\|^2\right]\mathrm{d}t\\
&=\int_{\underline{t}}^{T}\mathbb{E}_{x_{t,\mathcal{N}_j^r}\sim\pi_t}\left[\left\|u_{\theta,j}(x_{t,\mathcal{N}_j^r},t)-u_j^*(x_{t,\mathcal{N}_j^r},t)\right\|^2\right]\mathrm{d}t+\text{const.}
\end{aligned}$$

Here u_j^* is the optimal localized approximation of the score function (6.2.3), and the constant depends only on π_0.

The proof can be found in Section 6.4.4. Proposition 6.2 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. [84], one immediately obtains that the sample complexity of the localized score matching is essentially independent of the ambient dimension d.

A complete error analysis We provide a concrete result below. Following [84], we assume a further boundedness constraint on the hypothesis space $\mathscr{H}_r$ (6.1.8):

$$\mathscr{H}_r^N = \left\{ s \in \mathscr{H}_r \;\middle|\; \forall j \in [\mathsf{b}],\ \|s_j(\cdot,t)\|_\infty \lesssim \frac{\log^2 N}{\sigma_t} \right\}. \tag{6.2.7}$$

The constraint is natural as the score function scales with σ_t^{-1}; see [84] for more discussions. We also assume the following technical regularity conditions on the target distribution.

Assumption 6.1. The target distribution π_0 satisfies the following conditions:

1. (Boundedness) π_0 is supported on $[-M, M]^d$, and its density is upper and lower bounded by some constants C_π, C_π^{-1} respectively.

2. (γ-smoothness) For any $j \in [\mathsf{b}]$, its marginal density

$$\pi_0(x_{\mathcal{N}_j^r}) \in \mathcal{B}_R(B_{a,b}^\gamma([-M, M]^{d_{j,r}})).$$

   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) $\pi_0(x_{\mathcal{N}_j^r})|_\Omega \in \mathcal{B}_1(C^\infty(\Omega))$, 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 (6.1.9).

Remark 6.3. [84] only considers the standard domain $[-1, 1]^d$. It can be simply extended to $[-M, M]^d$ by scaling argument. Denote $\pi^M := M^d \pi_0(M\cdot)$, then π^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 [84] for more discussions on the regularity conditions. The following theorem provides an overall error analysis by combining Proposition 6.1, Theorem 6.2 and Theorem 4.3 in [84]. We comment that [119] points out a flaw in the proof in [84], but the issue is fixed in [119].

Theorem 6.3. *Let π_0 satisfy Assumption 6.1 and the conditions in Theorem 6.2. Given sample size N, let $\mathscr{H}_r^N$ be the bounded hypothesis space (6.2.7) with $\mathscr{U}_j =$*

$\mathsf{NN}(\mathsf{L}^j, \mathsf{W}^j, \mathsf{S}^j, \mathsf{B}^j)$ (6.1.11). *Denote* $n_j = N^{-d_j/(2\gamma+d_j)}$, *and choose the hyperparameters as*

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

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* (6.1.13) *in* $\mathscr{H}_r^N$. *Denote* $\widehat{\mu}_{T-t}$ *as the sampled distribution using learned score* $\widehat{s}$. *Then it holds that*

$$\begin{aligned}\mathbb{E}_{\{X^{(i)}\}_{i=1}^N}[\mathsf{KL}(\pi_{\underline{t}}\|\widehat{\mu}_{T-\underline{t}})] \le\ & \mathrm{e}^{-2T}\mathsf{KL}(\pi_0\|\mathsf{N}(0,I)) \\ & + Cd(r+1)^\nu \mathrm{e}^{-c(r+1)} + C'bN^{-\frac{2\gamma}{d_{\mathrm{eff}}+2\gamma}}\log^{16} N.\end{aligned} \tag{6.2.8}$$

Here d_{eff} *is the effective dimension* (6.1.9), C, c *are dimensional independent constants in Theorem 6.2, and* C' *is a dimensional independent constant.*

The proof can be found in Section 6.4.5. There are three sources of error in (6.2.8):

(1) Approximation error of π_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 $\frac{2\gamma}{d_{\mathrm{eff}}+2\gamma}$.

Remark 6.4. (1) Compared to the vanilla method, the localized diffusion models achieve a much faster statistical rate $\frac{2\gamma}{d_{\mathrm{eff}}+2\gamma} \gg \frac{2\gamma}{d+2\gamma}$, and thus potentially mitigate the curse of dimensionality.

(2) (6.2.8) 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 $d_{\mathrm{eff}} \asymp r^\nu$ (see Definition 2.1), 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:

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

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 $\pi_{\underline{t}}$ by convention. In fact, the early-stopping error can be controlled straightforwardly in

Wasserstein distance. For instance, by Lemma 3 in [22], it holds that $\mathsf{W}_2^2(\pi_0, \pi_{\underline{t}}) \lesssim d\underline{t}$. So that the overall error

$$\begin{aligned}\mathbb{E}_{\{X^{(i)}\}_{i=1}^N}[\mathsf{W}_2^2(\pi_0, \widehat{\mu}_{T-\underline{t}})] &\lesssim \mathsf{W}_2^2(\pi_0, \pi_{\underline{t}}) + \mathbb{E}_{\{X^{(i)}\}_{i=1}^N}[\mathsf{W}_2^2(\pi_{\underline{t}}, \widehat{\mu}_{T-\underline{t}})] \\ &\lesssim dN^{-k} + \mathbb{E}_{\{X^{(i)}\}_{i=1}^N}[\mathsf{KL}(\pi_{\underline{t}} \| \widehat{\mu}_{T-\underline{t}})].\end{aligned}$$

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

6.3 Numerical experiments

6.3.1 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.

Approximate locality

Consider localized Gaussian distribution

$$\pi_0 = \mathsf{N}(0, C_0),$$

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 $s(x, t) = \nabla \log \pi_t(x) = -P_t x$, where

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

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.

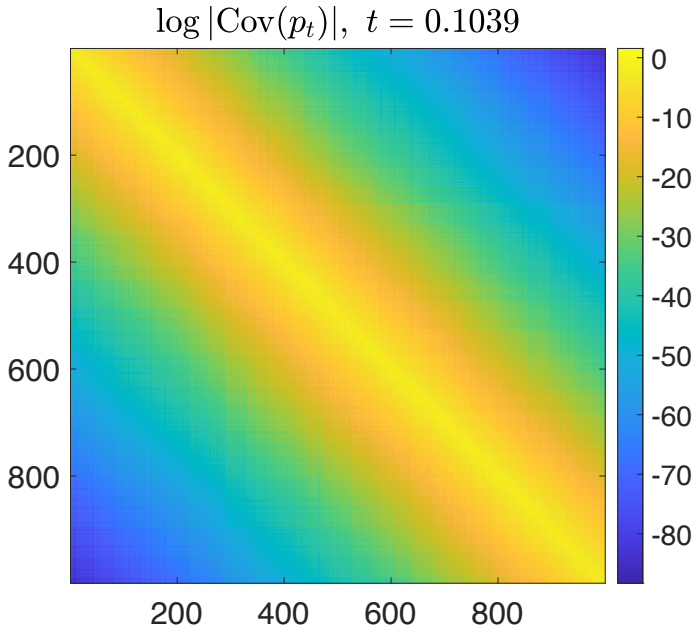

Figure 6.1: Locality in diffusion models. The precision matrix P_t at $t = 0.1039$, plotted in $\log |P_t|$ scale. We can see precise exponential decay of $P_t(i, j)$ in $|i - j|$.

First, we show in the top-left plot in Figure 6.1 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 middle plot in Figure 6.2). 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,

$$r_{\mathrm{loc}}(t) := \max \left\{ 1 \le r < d \ : \ \frac{1}{d-r} \sum_{1 \le i \le d-r} |P_t(i, i+r)| \ge \epsilon \cdot \frac{1}{d} \mathrm{tr}(P_t) \right\}. \quad (6.3.1)$$

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 6.2.

From Figure 6.2, we can see that the effective localization radius $r_{\mathrm{loc}}(t)$ 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 6.1, where the bound

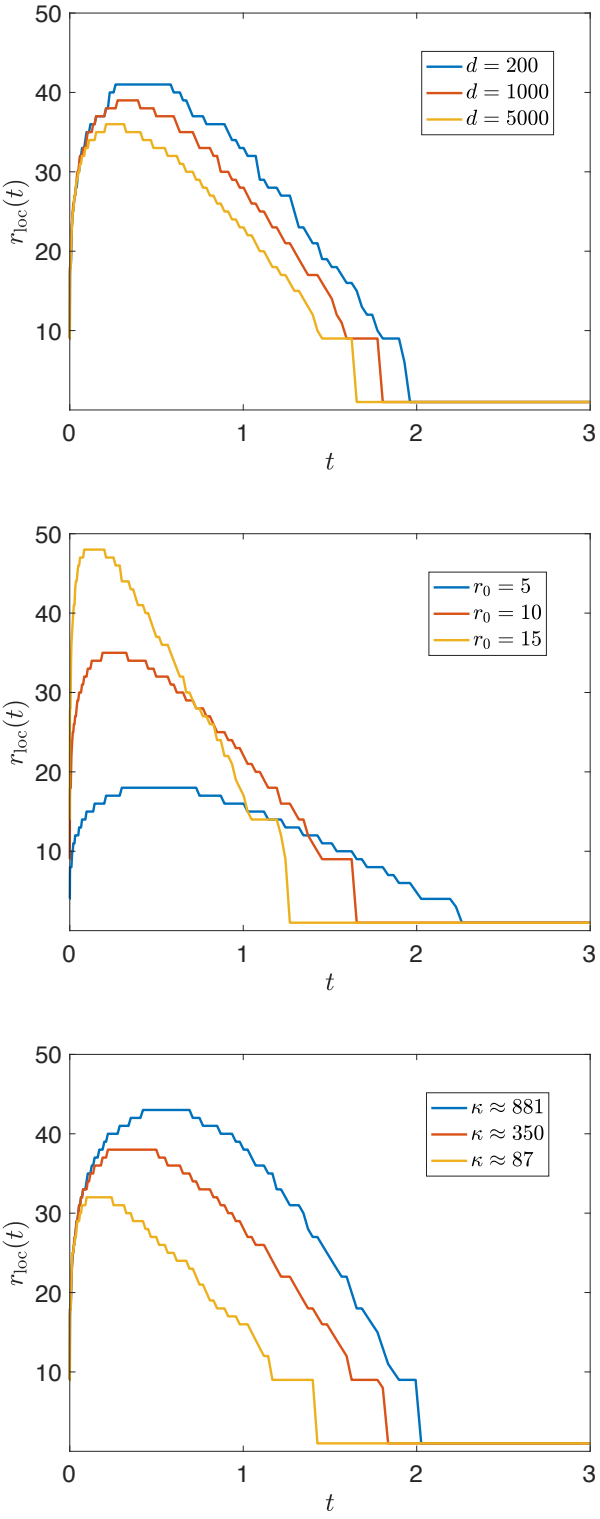

Figure 6.2: Localized diffusion model: effective localization radius $r_{\mathrm{loc}}(t)$ (6.3.1) under different problem dimension d, precision matrix bandwidth r_0 and condition number κ. Left: $r_{\mathrm{loc}}(t)$ with different dimensions. Here $r_0 = 10$ and the condition numbers are similar ($\kappa \approx 193, 191, 197$). Middle: $r_{\mathrm{loc}}(t)$ with different bandwidths. Here $d = 1,000$ and condition numbers $\kappa \approx 163, 146, 132$. Right: $r_{\mathrm{loc}}(t)$ with different condition numbers. Here $d = 1,000$ and $r_0 = 10$.

of $\|\nabla^2_{ij} \log \pi_t\|$ 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.

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

$$\pi_0(x) = \mathsf{N}(x_1; 0, 1) \prod_{n=1}^{d-1} \mathsf{N}(x_{n+1}; \alpha_h x_n, \sigma_h^2). \tag{6.3.2}$$

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 (6.1.6), which admits an explicit solution

$$\widehat{P}_t = (\alpha_t^2 \widehat{C}_0 + \sigma_t^2 I)^{-1}, \tag{6.3.3}$$

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

$$Y_{t_{n+1}} = Y_{t_n} + \Delta t_n \left(I - 2\widehat{P}_{T-t_n}\right) Y_{t_n} + \sqrt{2\Delta t_n}\xi_n. \tag{6.3.4}$$

Here $\widehat{P}_t$ is the estimated optimal precision matrix (6.3.3), $\xi_n \sim \mathsf{N}(0, I), Y_0 \sim \mathsf{N}(0, I)$, and $\Delta t_n = t_{n+1} - t_n$ is the time step. We use the linear variance schedule $\beta_n = (\beta_N - \beta_1)\frac{n-1}{N-1} + \beta_1$ ($1 \le n \le N$) [61], 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 (6.3.4) is

$$\begin{aligned} Y^{\text{loc},r}_{t_{n+1}} &= Y^{\text{loc},r}_{t_n} + \Delta t_n \left(I - 2\widehat{P}^{\text{loc},r}_{T-t_n}\right) Y^{\text{loc},r}_{t_n} + \sqrt{2\Delta t_n}\xi_n, \\ \widehat{P}^{\text{loc},r}_{T-t_n}(i,j) &:= \widehat{P}_{T-t_n}(i,j)\mathbf{1}_{|i-j|\le r}. \end{aligned} \tag{6.3.5}$$

We will use (6.3.5) to sample the target distribution with different localization radii r, and compare it to the reference sampling process (6.3.4). Although the localized score $\widehat{s}^{\mathrm{loc},r}(t,x) = -\widehat{P}_t^{\mathrm{loc},r} x$ in (6.3.5) is not the minimizer of $\widehat{\mathcal{L}}_N(s_\theta)$ (6.1.13), it is very close to the minimizer, and it still yields a good approximation, see Figure 6.3.

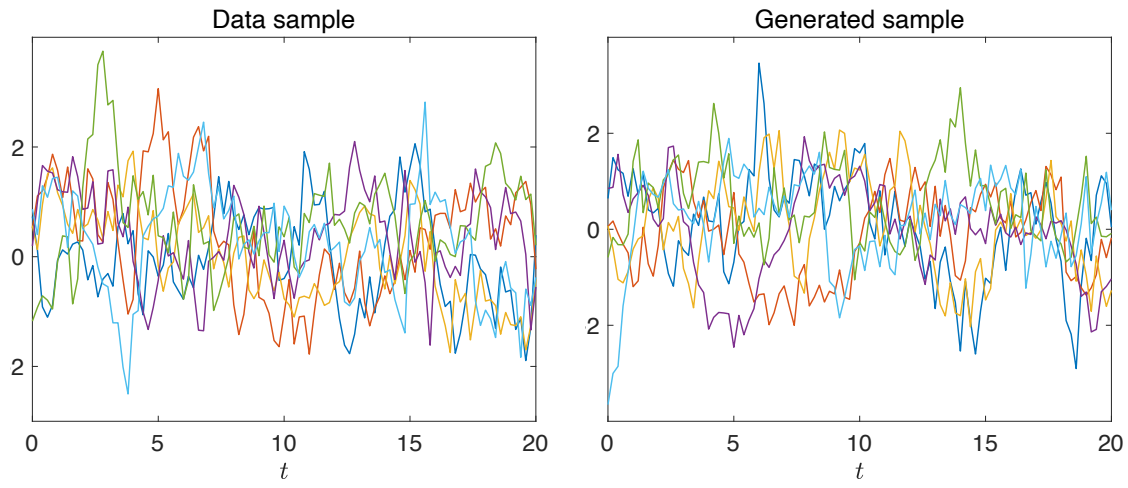

Figure 6.3: Localized diffusion model: sampling OU process. Left: Trajectories directly sampled from OU process. Right: Sampled trajectories using the localized sampling process (6.3.5) with localization radius $r = 12$.

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 6.4. We measure the relative ℓ^2-error of the sample covariance

$$\mathrm{err} := \frac{\|\widehat{C} - C\|_2}{\|C\|_2}, \tag{6.3.6}$$

where $C = P_0^{-1}$ is the true covariance, $\widehat{C}$ is the sample covariance of samples from (6.3.4) or (6.3.5), and $\|\cdot\|_2$ is the matrix 2-norm. The reference error is computed using the sample covariance of the non-localized backward process (6.3.4). 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 Figure 6.5, we plot the entrywise error of the sample covariance (normalized

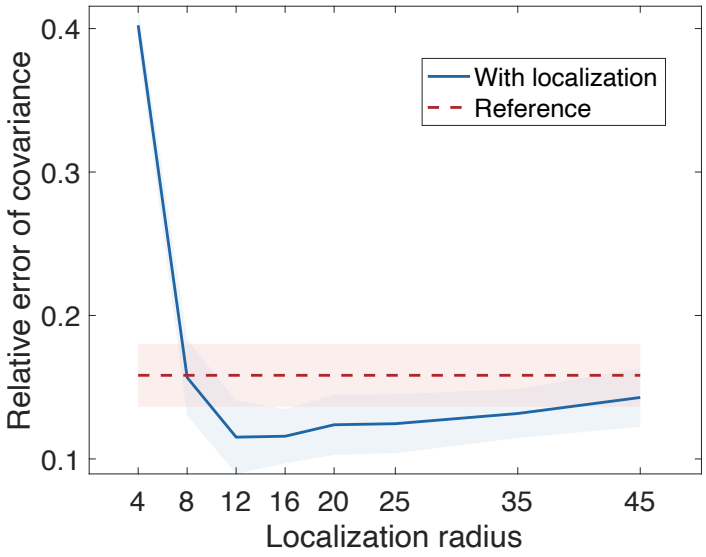

Figure 6.4: Localized diffusion model: error tradeoff in sampling OU process. Relative ℓ^2-error (6.3.6) of the sample covariance for different localization radii r; the reference error is from the non-localized sampling process (6.3.4). The shaded area denotes the 1σ region.

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 [63]. 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 6.3.

6.3.2 Cox-Ingersoll-Ross model

We consider the Cox-Ingersoll-Ross (CIR) model [30, 31]

$$\mathrm{d}X = 2a(b - X)\mathrm{d}t + \sigma\sqrt{X}\mathrm{d}W_t, \tag{6.3.7}$$

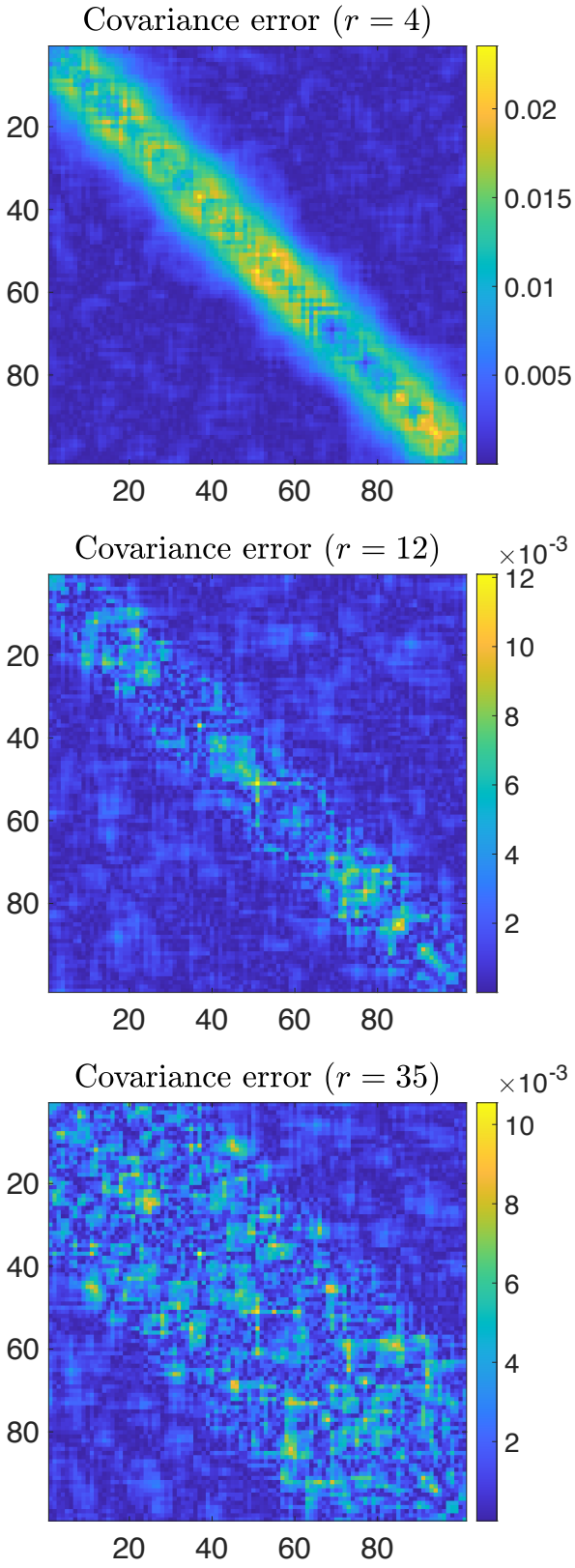

Figure 6.5: Localized diffusion model: entrywise sample covariance error in sampling OU process with different localization radius $r \in \{4, 12, 35\}$.

where W_t is standard one-dimensional Brownian motion. The CIR model (6.3.7) possesses a closed form solution

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

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 (6.3.7) 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 6.6 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.

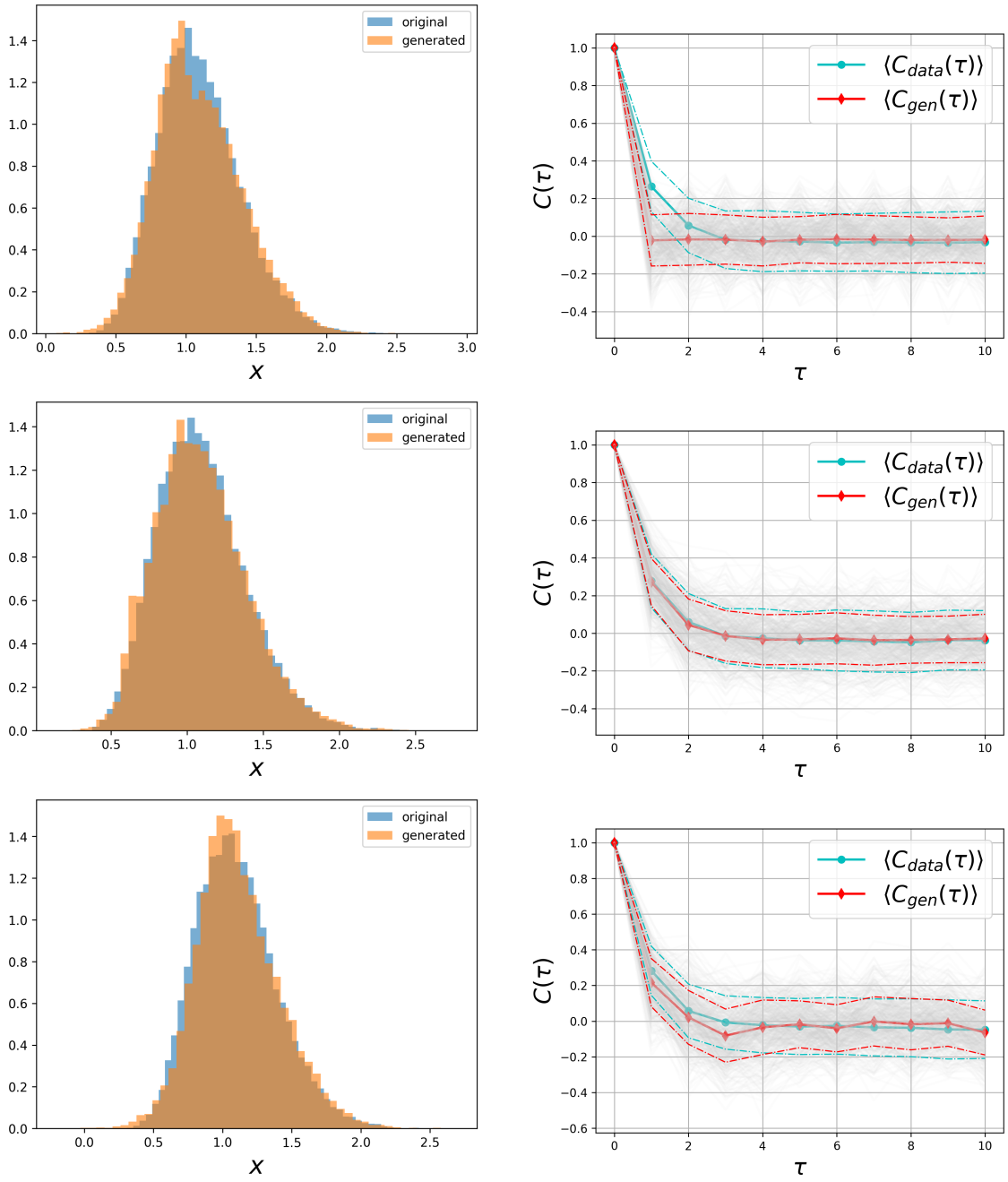

Figure 6.6: Localized diffusion model: sampling CIR model. Comparison of the data obtained from the original CIR model (6.3.7) and from the diffusion model for localization radii $r = 0$ (left), $r = 2$ (middle) and $r = 20$ (right). Top: Empirical histogram. Bottom: Auto-correlation function, averaged 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 $i = 2r+1, \ldots, N-2r-1$ from the localized state $(x_r)_i = [x_{i-r}, \ldots, x_i, \ldots, x_{i+r}] \in \mathbb{R}^{2r+1}$. 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. $i \leq r$ or $i \geq d-r$. 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 $i = 1, \ldots, r$ and $i = N-r, \ldots, N$, 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.

6.4 Proofs

6.4.1 Proof of Theorem 6.1

Proof of Theorem 6.1. Recall

$$\pi_t(x_t) = \int \mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I)\pi_0(x_0)\mathrm{d}x_0.$$

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

$$\begin{aligned}
\nabla^2 \log \pi_t(x_t) &= \frac{\nabla^2 \pi_t(x_t)}{\pi_t(x_t)} - \frac{\nabla \pi_t(x_t)}{\pi_t(x_t)} \frac{\nabla \pi_t(x_t)^{\mathrm{T}}}{\pi_t(x_t)} \\
&= \frac{1}{\pi_t(x_t)} \int \left(-\frac{x_t - \alpha_t x_0}{\sigma_t^2}\right) \left(-\frac{x_t - \alpha_t x_0}{\sigma_t^2}\right)^{\mathrm{T}} \mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I)\pi_0(x_0)\mathrm{d}x_0 \\
&\quad - \frac{1}{\pi_t(x_t)} \int \left(-\frac{x_t - \alpha_t x_0}{\sigma_t^2}\right) \mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I)\pi_0(x_0)\mathrm{d}x_0 \\
&\quad \cdot \frac{1}{\pi_t(x_t)} \int \left(-\frac{x_t - \alpha_t x_0}{\sigma_t^2}\right)^{\mathrm{T}} \mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I)\pi_0(x_0)\mathrm{d}x_0 \\
&= \sigma_t^{-4} \mathbb{E}_{\pi_{0|t}(x_0|x_t)} \left(x_t - \alpha_t x_0\right) \left(x_t - \alpha_t x_0\right)^{\mathrm{T}} \\
&\quad - \sigma_t^{-4} \mathbb{E}_{\pi_{0|t}(x_0|x_t)} \left(x_t - \alpha_t x_0\right) \mathbb{E}_{\pi_{0|t}(x_0|x_t)} \left(x_t - \alpha_t x_0\right)^{\mathrm{T}} \\
&= \sigma_t^{-4} \mathsf{Cov}_{\pi_{0|t}(x_0|x_t)} \left(x_t - \alpha_t x_0, x_t - \alpha_t x_0\right) \\
&= \alpha_t^2 \sigma_t^{-4} \mathsf{Cov}_{\pi_{0|t}(x_0|x_t)} \left(x_0, x_0\right),
\end{aligned}$$

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

$$\nabla^2_{ij} \log \pi_t(x_t) = \alpha_t^2 \sigma_t^{-4} \mathsf{Cov}_{\pi_{0|t}(x_0|x_t)} (x_{0,i}, x_{0,j}) . \tag{6.4.1}$$

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

$$\log \pi_{0|t}(x_0|x_t) = -\log \pi_t(x_t) + \log \pi_0(x_0) - \frac{1}{2\sigma_t^2} \|x_t - \alpha_t x_0\|^2 - \frac{d}{2} \log(2\pi\sigma_t^2).$$

Fix x_t, and denote for simplicity $p(x) = \pi_{0|t}(x|x_t)$. Then

$$\nabla^2 \log p(x) = \nabla^2 \log \pi_0(x) - \frac{\alpha_t^2}{\sigma_t^2} I.$$

Note by assumption, $\nabla^2_{ij} \log \pi_0 = 0$ if $i \notin \mathcal{N}_j$, and $mI \preceq -\nabla^2 \log \pi_0 \preceq MI$. So that

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

$$\left(m + \frac{\alpha_t^2}{\sigma_t^2}\right) I \preceq -\nabla^2 \log \pi_0 \preceq \left(M + \frac{\alpha_t^2}{\sigma_t^2}\right) I. \tag{6.4.2}$$

By Theorem 3.5, for any Lipschitz functions f, g, we have

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

Recall (6.4.1), and by definition of the matrix norm,

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

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

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

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

6.4.2 Proof of Proposition 6.1

Proof of Proposition 6.1. Denote the path measures for the reverse process (6.1.2) and the sampling process (6.2.1) 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

$$\mathsf{KL}(\pi_{\underline{t}} \| \widehat{\mu}_{T-\underline{t}}) = \mathsf{KL}(\mathsf{Q}_{T-\underline{t}} \| \widehat{\mathsf{Q}}_{T-\underline{t}}) \leq \mathsf{KL}(\mathsf{Q}_{[0,T-\underline{t}]} \| \widehat{\mathsf{Q}}_{[0,T-\underline{t}]}).$$

By the Girsanov theorem [3], we have

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

By the convergence of the OU process [3], we have

$$\mathsf{KL}(\pi_T \| \mathsf{N}(0, I)) \leq \mathrm{e}^{-2T} \mathsf{KL}(\pi_0 \| \mathsf{N}(0, I)).$$

The conclusion follows by combining the above relations. □

6.4.3 Proof of Theorem 6.2

Proof of Theorem 6.2. Note the optimal solution is given by (6.2.3), i.e.,

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

By (6.4.2), π_t is $\left(m + \frac{\alpha_t^2}{\sigma_t^2}\right)$-strongly log-concave, so that the conditional distribution $\pi_t(x_{-\mathcal{N}_j^r} | x_{\mathcal{N}_j^r})$ is also $\left(m + \frac{\alpha_t^2}{\sigma_t^2}\right)$-strongly log-concave. By the Poincaré inequality [3],

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

Here $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. By Theorem 6.1, it holds that

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

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

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

Therefore,

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

The last step uses Lemma 6.1. By the Abel transformation and the sparsity assumption (2.1.3),

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

By Lemma 3.6, it holds that $\sum_{k\in\mathbb{Z}_+} k^n x^k \le n! x(1-x)^{-n-1}$, 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

$$\begin{aligned}
&\int_{\underline{t}}^{T} \left\|s_j^*(x,t) - s_j(x,t)\right\|_{L^2(\pi_t)}^2 \mathrm{d}t \\
&\le \int_{0}^{T} \left\|s_j^*(x,t) - s_j(x,t)\right\|_{L^2(\pi_t)}^2 \mathrm{d}t \\
&\le d_j \max\{1, m^{-1}\} \log\kappa \cdot 2\mathsf{s}\kappa^{-1}(1-\kappa^{-1})^{2r} \cdot (r+1)^{\nu}\nu!(1-\kappa^{-1})^2\kappa^{2(\nu+1)} \\
&= C d_j (r+1)^{\nu}(1-\kappa^{-1})^{2(r+1)}.
\end{aligned}$$

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

The second claim follows from the property of conditional expectation:

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

This completes the proof. □

Lemma 6.1. *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

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

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. □

6.4.4 Proof of Proposition 6.2

Proof of Proposition 6.2. The first equality directly follows from the definition (6.2.6). 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

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

It holds that

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

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

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

This verifies the second inequality.

For the third inequality, we first claim that

$$u_j^*(x_{t,\mathcal{N}_j^r}, t) = \nabla_j \log \pi_t(x_{t,\mathcal{N}_j^r}). \tag{6.4.3}$$

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 $\pi(y,z) = \pi_{t,0}(x_{t,\mathcal{N}_j^r}, x_{0,\mathcal{N}_j^r})$ 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 (6.4.3). Note that

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

Since

$$\pi_t(x_t) = \int \mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I)\pi_0(x_0)\mathrm{d}x_0.$$

$$\Rightarrow\ \nabla_j \pi_t(x_t) = \int \left(-\sigma_t^{-2}(x_{t,j} - \alpha_t x_{0,j})\right)\mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I)\pi_0(x_0)\mathrm{d}x_0.$$

So that

$$\begin{aligned}
u_j^*(x_{t,\mathcal{N}_j^r}, t) &= \frac{\int \left(-\sigma_t^{-2}(x_{t,j} - \alpha_t x_{0,j})\right)\mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I)\pi_0(x_0)\mathrm{d}x_0\mathrm{d}x_{t,-\mathcal{N}_j^r}}{\int \mathsf{N}(x_t; \alpha_t x_0, \sigma_t^2 I)\pi_0(x_0)\mathrm{d}x_0\mathrm{d}x_{t,-\mathcal{N}_j^r}} \\
&= \frac{\int \left(-\sigma_t^{-2}(x_{t,j} - \alpha_t x_{0,j})\right)\mathsf{N}(x_{t,\mathcal{N}_j^r}; \alpha_t x_{0,\mathcal{N}_j^r}, \sigma_t^2 I)\pi_0(x_{0,\mathcal{N}_j^r})\mathrm{d}x_{0,\mathcal{N}_j^r}}{\int \mathsf{N}(x_{t,\mathcal{N}_j^r}; \alpha_t x_{0,\mathcal{N}_j^r}, \sigma_t^2 I)\pi_0(x_{0,\mathcal{N}_j^r})\mathrm{d}x_{0,\mathcal{N}_j^r}}.
\end{aligned}$$

On the other hand,

$$
\begin{aligned}
&\nabla_j \log \pi_t(x_{t,\mathcal{N}_j^r}) \\
=\ & \frac{\nabla_j \pi_t(x_{t,\mathcal{N}_j^r})}{\pi_t(x_{t,\mathcal{N}_j^r})} = \frac{\int \nabla_j \mathsf{N}(x_{t,\mathcal{N}_j^r}; \alpha_t x_{0,\mathcal{N}_j^r}, \sigma_t^2 I)\pi_0(x_{0,\mathcal{N}_j^r})\mathrm{d}x_{0,\mathcal{N}_j^r}}{\int \mathsf{N}(x_{t,\mathcal{N}_j^r}; \alpha_t x_{0,\mathcal{N}_j^r}, \sigma_t^2 I)\pi_0(x_{0,\mathcal{N}_j^r})\mathrm{d}x_{0,\mathcal{N}_j^r}} \\
=\ & \frac{\int \left(-\sigma_t^{-2}(x_{t,j} - \alpha_t x_{0,j})\right) \mathsf{N}(x_{t,\mathcal{N}_j^r}; \alpha_t x_{0,\mathcal{N}_j^r}, \sigma_t^2 I)\pi_0(x_{0,\mathcal{N}_j^r})\mathrm{d}x_{0,\mathcal{N}_j^r}}{\int \mathsf{N}(x_{t,\mathcal{N}_j^r}; \alpha_t x_{0,\mathcal{N}_j^r}, \sigma_t^2 I)\pi_0(x_{0,\mathcal{N}_j^r})\mathrm{d}x_{0,\mathcal{N}_j^r}} \\
=\ & u_j^*(x_{t,\mathcal{N}_j^r}, t).
\end{aligned}
$$

This completes the proof. □

6.4.5 Proof of Theorem 6.3

Proof of Theorem 6.3. By the Pythagorean equality (6.2.5),

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

Combining Proposition 6.1 and Theorem 6.2, we obtain

$$
\begin{aligned}
\mathsf{KL}(\pi_{\underline{t}}\|\widehat{\mu}_{T-\underline{t}}) &\le \mathrm{e}^{-2T}\mathsf{KL}(\pi_0\|\mathsf{N}(0,I)) + \int_{\underline{t}}^{T} \mathbb{E}_{x_t\sim\pi_t}\left[\|\widehat{s}(x_t,t) - s(x_t,t)\|^2\right]\mathrm{d}t \\
&= \mathrm{e}^{-2T}\mathsf{KL}(\pi_0\|\mathsf{N}(0,I)) + \int_{\underline{t}}^{T} \mathbb{E}_{x_t\sim\pi_t}\left[\|s^*(x_t,t) - s(x_t,t)\|^2\right]\mathrm{d}t + \mathcal{R} \\
&\le \mathrm{e}^{-2T}\mathsf{KL}(\pi_0\|\mathsf{N}(0,I)) + Cd(r+1)^{\nu}\mathrm{e}^{-c(r+1)} + \mathcal{R},
\end{aligned}
$$

where we denote

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

By Proposition 6.2, $\mathcal{R}_j$ is the j-th component loss of the score function when we use a standard diffusion model to approximate the marginal distribution $\pi_0(x_{N_j^r})$. Note one can use the same constructive solution as in [84] for the marginal target

$\pi_0(x_{\mathcal{N}_j^r})$ 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 [84]:

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

where $n_j = N^{-d_j/(2\gamma+d_j)}$. Note n, N in our paper correspond to N, n in [84] 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 [84]:

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

Therefore,

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

This completes the proof. □

Chapter 7

Conclusions and Future Work

In this thesis, we presented a comprehensive study of the theory and numerical methods for localized sampling in high dimensions by leveraging locality structure. Our main contributions are twofold:

1. Marginal Stein's method. We developed a new analysis method for localized distributions, deriving a dimension-independent marginal transport inequality and rigorous proofs of exponential correlation decay. These results clarify how sparse dependencies control the propagation of errors and form the foundation for localization methods in sampling.

2. Localized sampling algorithms. Motivated by the theoretical insights, we formulated a general framework to reduce global samplers into a collection of low-dimensional, neighborhood-based samplers. Our study of MALA-within-Gibbs and localized diffusion models demonstrates that localization can greatly reduce both computational cost and sample complexity without sacrificing accuracy, making them potentially more efficient for large-scale applications.

By combining rigorous theory with practical algorithm design, this thesis laid the groundwork for a new class of sampling methods that can overcome the curse of dimensionality. Beyond these studies, the marginal Stein's method offers a new tool for analyzing more sampling or variational algorithms for problems with locality structures. The localization framework also brings new methods for scalable inference in graphical models, spatiotemporal processes, and large-scale Bayesian inverse problems.

Building on the results of this thesis, there are several future research directions:

- Adaptive localization. Design sampling algorithms that adaptively learn the

locality structure. This broadens the applications of localization to problems with unknown or dynamic dependencies.

- Multiscale generalization. Combine localization with global dimension reduction techniques.
- Applications to deep generative models. Understand how locality structures in the data distribution or the neural network architecture can be exploited to improve the learning and generation performance of deep generative models.

We believe that the localization framework will continue to inspire new theoretical insights and practical algorithms for high-dimensional sampling problems, and we are excited to see how it develops in the future.

Bibliography

[1] M. B. Averintsev, *Description of Markovian random fields by Gibbsian conditional probabilities*, Theory Probab. Appl., 17 (1972), pp. 20–33.

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

[3] D. Bakry, I. Gentil, and M. Ledoux, *Analysis and Geometry of Markov Diffusion Operators*, vol. 348 of Fundamental Principles of Mathematical Sciences, Springer, Cham, 2014.

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

[5] J. Benton, V. D. Bortoli, A. Doucet, and G. Deligiannidis, *Nearly d-linear convergence bounds for diffusion models via stochastic localization*, in 12th Int. Conf. Learn. Represent., 2024.

[6] M. Benzi, *Localization in Matrix Computations: Theory and Applications*, Springer International Publishing, Cham, 2016, pp. 211–317.

[7] M. Benzi, P. Boito, and N. Razouk, *Decay properties of spectral projectors with applications to electronic structure*, SIAM Rev., 55 (2013), pp. 3–64.

[8] J. Besag, *Spatial interaction and the statistical analysis of lattice systems*, J. R. Stat. Soc., 36 (1974), pp. 192–236.

[9] P. Bickel and M. Lindner, *Approximating the inverse of banded matrices by banded matrices with applications to probability and statistics*, Theory Probab. Appl., 56 (2012), pp. 1–20.

[10] K. Binder, D. M. Ceperley, J.-P. Hansen, M. Kalos, D. Landau, D. Levesque, H. Mueller-Krumbhaar, D. Stauffer, and J.-J.

Weis, *Monte Carlo Methods in Statistical Physics*, vol. 7, Springer Science & Business Media, 2012.

[11] K. Binder and D. W. Heermann, *Monte Carlo Simulation in Statistical Physics: An Introduction*, Graduate Texts in Physics, Springer, Cham, sixth ed., 2019.

[12] B. Bollobás, *Modern Graph Theory*, vol. 184 of Graduate Texts in Mathematics, Springer-Verlag, New York, 1998.

[13] G. E. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*, John Wiley & Sons, 2011.

[14] G. a. A. Braga, P. C. Lima, and M. L. O'Carroll, *Exponential decay of truncated correlation functions via the generating function: a direct method*, Rev. Math. Phys., 10 (1998), pp. 429–438.

[15] M. Campanino, D. Ioffe, and Y. Velenik, *Ornstein-Zernike theory for finite range Ising models above T_c*, Probab. Theory Related Fields, 125 (2003), pp. 305–349.

[16] H. Chen, H. Lee, and J. Lu, *Improved analysis of score-based generative modeling: user-friendly bounds under minimal smoothness assumptions*, in Proc. 40th Int. Conf. Mach. Learn., vol. 202 of Proc. Mach. Learn. Res., PMLR, 2023, pp. 4735–4763.

[17] L. H. Y. Chen, *Poisson approximation for dependent trials*, Ann. Probab., 3 (1975), pp. 534–545.

[18] M. Chen, K. Huang, T. Zhao, and M. Wang, *Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data*, in Proc. 40th Int. Conf. Mach. Learn., vol. 202 of Proc. Mach. Learn. Res., PMLR, 2023, pp. 4672–4712.

[19] 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).

[20] Y. Chen, X. Cheng, J. Niles-Weed, and J. Weare, *Convergence of unadjusted langevin in high dimensions: delocalization of bias*, arXiv preprint arXiv:2408.13115, (2024).

[21] Y. Chen, T. T. Georgiou, and M. Pavon, *Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger bridge*, SIAM Rev., 63 (2021), pp. 249–313.

[22] Y. Chen, D. Z. Huang, J. Huang, S. Reich, and A. M. Stuart, *Sampling via gradient flows in the space of probability measures*, arXiv preprint arXiv:2310.03597, (2023).

[23] V. Chernozhukov, D. Chetverikov, and K. Kato, *Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors*, Ann. Statist., 41 (2013), pp. 2786–2819.

[24] P. Clifford, *Markov random fields in statistics*, Disorder in physical systems: A volume in honour of John M. Hammersley, (1990), pp. 19–32.

[25] P. Clifford and J. Hammersley, *Markov fields on finite graphs and lattices*, Unpublished manuscript, (1971).

[26] P. Concus, G. H. Golub, and G. Meurant, *Block preconditioning for the conjugate gradient method*, SIAM J. Sci. Stat. Comput., 6 (1985), pp. 220–252.

[27] G. Conforti, A. Durmus, and M. G. Silveri, *KL convergence guarantees for score diffusion models under minimal data assumptions*, SIAM J. Math. Data Sci., 7 (2025), pp. 86–109.

[28] P. G. Constantine, E. Dow, and Q. Wang, *Active subspace methods in theory and practice: applications to kriging surfaces*, SIAM J. Sci. Comput., 36 (2014), pp. A1500–A1524.

[29] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White, *MCMC methods for functions: modifying old algorithms to make them faster*, Statist. Sci., 28 (2013), pp. 424–446.

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

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

[32] T. Cui, K. J. H. Law, and Y. M. Marzouk, *Dimension-independent likelihood-informed MCMC*, J. Comput. Phys., 304 (2016), pp. 109–137.

[33] T. Cui, S. Liu, and X. T. Tong, *Stein's method for marginals on large graphical models*, arXiv preprint arXiv:2410.11771, (2025).

[34] T. Cui, J. Martin, Y. M. Marzouk, A. Solonen, and A. Spantini, *Likelihood-informed dimension reduction for nonlinear inverse problems*, Inverse Problems, 30 (2014), pp. 114015–114042.

[35] T. Cui and X. T. Tong, *A unified performance analysis of likelihood-informed subspace methods*, Bernoulli, 28 (2022), pp. 2788–2815.

[36] M. Cuturi, *Sinkhorn distances: lightspeed computation of optimal transport*, in Adv. Neural Inf. Process. Syst., vol. 26, Curran Associates, Inc., 2013.

[37] A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand, *Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets*, J. Amer. Statist. Assoc., 111 (2016), pp. 800–812.

[38] S. Demko, W. F. Moss, and P. W. Smith, *Decay rates for inverses of band matrices*, Math. Comp., 43 (1984), pp. 491–499.

[39] P. L. Dobruschin, *The description of a random field by means of conditional probabilities and conditions of its regularity*, Theory Probab. Appl., 13 (1968), pp. 197–224.

[40] R. L. Dobrushin and S. B. Shlosman, *Completely Analytical Gibbs Fields*, Birkhäuser Boston, Boston, MA, 1985, pp. 371–403.

[41] H. Duminil-Copin, S. Goswami, and A. Raoufi, *Exponential decay of truncated correlations for the Ising model in any dimension for all but the critical temperature*, Comm. Math. Phys., 374 (2020), pp. 891–921.

[42] R. Durrett, *Probability: Theory and Examples*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, fifth ed., 2019.

[43] F. J. Dyson, *The radiation theories of Tomonaga, Schwinger, and Feynman*, Phys. Rev., 75 (1949), pp. 486–502.

[44] J. Fan, W. Wang, and Y. Zhong, *An ℓ_∞ eigenvector perturbation bound and its application*, J. Mach. Learn. Res., 18 (2018), pp. 1–42.

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

[46] R. Flock, S. Liu, Y. Dong, and X. T. Tong, *Local MALA-within-Gibbs for Bayesian image deblurring with total variation prior*, SIAM J. Sci. Comput., 47 (2025), pp. A2127–A2153.

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

[48] A. Gelman and D. B. Rubin, *Inference from iterative simulation using multiple sequences*, Statist. Sci., 7 (1992), pp. 457–472.

[49] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Mach. Intell., PAMI-6 (1984), pp. 721–741.

[50] R. Ghanem, D. Higdon, and H. Owhadi, eds., *Handbook of Uncertainty Quantification. Vols. 1, 2, 3*, Springer, Cham, 2017.

[51] M. Girolami and B. Calderhead, *Riemann manifold Langevin and Hamiltonian Monte Carlo methods*, J. R. Stat. Soc. Ser. B, 73 (2011), pp. 123–214.

[52] P.-L. Giscard, K. Lui, S. J. Thwaite, and D. Jaksch, *An exact formulation of the time-ordered exponential using path-sums*, J. Math. Phys., 56 (2015), pp. 053503–053520.

[53] S. Goedecker, *Linear scaling electronic structure methods*, Rev. Mod. Phys., 71 (1999), pp. 1085–1123.

[54] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, in Adv. Neural Inf. Process. Syst., vol. 27, Curran Associates, Inc., 2014.

[55] G. A. Gottwald, F. Li, Y. Marzouk, and S. Reich, *Stable generative modelling using Schrödinger bridges*, Philos. Trans. R. Soc. A, 383 (2025), p. 20240332.

[56] G. A. Gottwald, S. Liu, Y. Marzouk, S. Reich, and X. T. Tong, *Localized diffusion models for high dimensional distributions generation*, arXiv preprint arXiv:2505.04417, (2025).

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

[58] T. M. Hamill, J. S. Whitaker, and C. Snyder, *Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter*, Mon. Weather Rev., 129 (2001), pp. 2776–2790.

[59] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer, New York, second ed., 2009.

[60] E. Hernández-Lemus, *Random fields in physics, biology and data science*, Front. Phys., 9 (2021).

[61] J. Ho, A. Jain, and P. Abbeel, *Denoising diffusion probabilistic models*, in Adv. Neural Inf. Process. Syst., vol. 33, Curran Associates, Inc., 2020, pp. 6840–6851.

[62] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, second ed., 2013.

[63] P. L. Houtekamer and H. L. Mitchell, *A sequential ensemble Kalman filter for atmospheric data assimilation*, Mon. Weather Rev., 129 (2001), p. 123.

[64] Y. Hu and W. Wang, *Network-adjusted covariates for community detection*, Biometrika, 111 (2024), pp. 1221–1240.

[65] E. Ising, *Contribution to the theory of ferromagnetism*, Z. Phys, 31 (1925), pp. 253–258.

[66] R. Kindermann and J. L. Snell, *Markov Random Fields and Their Applications*, vol. 1 of Contemporary Mathematics, American Mathematical Society, Providence, RI, 1980.

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

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

[69] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA, 2009.

[70] R. Kotecký and D. Preiss, *Cluster expansion for abstract polymer models*, Comm. Math. Phys., 103 (1986), pp. 491–498.

[71] H. Künsch, *Decay of correlations under Dobrushin's uniqueness condition and its applications*, Comm. Math. Phys., 84 (1982), pp. 207–222.

[72] L. D. Landau, E. M. Lifshitz, E. M. Lifshitz, and L. Pitaevskii, *Statistical Physics: Theory of the Condensed State*, vol. 9, Butterworth-Heinemann, 1980.

[73] O. E. Lanford and D. Ruelle, *Observables at infinity and states with short range correlations in statistical mechanics*, Comm. Math. Phys., 13 (1969), pp. 194–215.

[74] K. Law, A. Stuart, and K. Zygalakis, *Data Assimilation: A Mathematical Introduction*, vol. 62 of Texts in Applied Mathematics, Springer, Cham, 2015.

[75] H. Lee, J. Lu, and Y. Tan, *Convergence for score-based generative modeling with polynomial complexity*, in Adv. Neural Inf. Process. Syst., vol. 35, Curran Associates, Inc., 2022, pp. 22870–22882.

[76] Q. Liu, *Stein variational gradient descent as gradient flow*, in Adv. Neural Inf. Process. Syst., vol. 30, Curran Associates, Inc., 2017.

[77] Q. Liu and D. Wang, *Stein variational gradient descent: a general purpose Bayesian inference algorithm*, in Adv. Neural Inf. Process. Syst., vol. 29, Curran Associates, Inc., 2016.

[78] T. Marshall and G. Roberts, *An adaptive approach to Langevin MCMC*, Stat. Comput., 22 (2012), pp. 1041–1057.

[79] Y. M. Marzouk, H. N. Najm, and L. A. Rahn, *Stochastic spectral methods for efficient Bayesian solution of inverse problems*, J. Comput. Phys., 224 (2007), pp. 560–586.

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

[81] 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.

[82] K. Murphy, *Machine Learning: A Probabilistic Perspective*, Adaptive Computation and Machine Learning series, MIT Press, 2012.

[83] B. Ø ksendal, *Stochastic Differential Equations: An Introduction with Applications*, Universitext, Springer-Verlag, Berlin, sixth ed., 2003.

[84] K. Oko, S. Akiyama, and T. Suzuki, *Diffusion models are minimax optimal distribution estimators*, in Proc. 40th Int. Conf. Mach. Learn., vol. 202 of Proc. Mach. Learn. Res., PMLR, 2023, pp. 26517–26582.

[85] J. P. Oliveira, J. M. Bioucas-Dias, and M. A. Figueiredo, *Adaptive total variation image deblurring: a majorization–minimization approach*, Signal Processing, 89 (2009), pp. 1683–1693.

[86] F. Otto and C. Villani, *Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality*, J. Funct. Anal., 173 (2000), pp. 361–400.

[87] N. Paragios, Y. Chen, and O. Faugeras, eds., *Handbook of Mathematical Models in Computer Vision*, Springer, New York, 2006.

[88] G. Peyré and M. Cuturi, *Computational Optimal Transport: With Applications to Data Science*, Foundations and trends in machine learning, Now Publishers, 2019.

[89] N. S. Pillai, A. M. Stuart, and A. H. Thiéry, *Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions*, Ann. Appl. Probab., 22 (2012), pp. 2320–2356.

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

[91] C. J. Preston, *Gibbs States on Countable Sets*, Cambridge Tracts in Mathematics, Cambridge University Press, 1974.

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

[93] S. Reich and C. Cotter, *Probabilistic Forecasting and Bayesian Data Assimilation*, Cambridge University Press, New York, 2015.

[94] G. O. Roberts and J. S. Rosenthal, *Optimal scaling of discrete approximations to Langevin diffusions*, J. R. Stat. Soc. Ser. B, 60 (1998), pp. 255–268.

[95] G. O. Roberts and R. L. Tweedie, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, 2 (1996), pp. 341–363.

[96] A. Röllin, *Stein's method in high dimensions with applications*, Ann. Inst. Henri Poincaré Probab. Stat., 49 (2013), pp. 529–549.

[97] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, *High-resolution image synthesis with latent diffusion models*, in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., June 2022, pp. 10684–10695.

[98] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D: Nonlinear Phenomena, 60 (1992), pp. 259–268.

[99] F. Schäfer, M. Katzfuss, and H. Owhadi, *Sparse Cholesky factorization by Kullback–Leibler minimization*, SIAM J. Sci. Comput., 43 (2021), pp. A2019–A2046.

[100] K. Shah, S. Chen, and A. Klivans, *Learning mixtures of Gaussians using the DDPM objective*, in Adv. Neural Inf. Process. Syst., vol. 36, Curran Associates, Inc., 2023, pp. 19636–19649.

[101] Y. Song and S. Ermon, *Generative modeling by estimating gradients of the data distribution*, in Adv. Neural Inf. Process. Syst., vol. 32, Curran Associates, Inc., 2019.

[102] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, *Score-based generative modeling through stochastic differential equations*, in Int. Conf. Learn. Represent., 2021.

[103] A. Spantini, D. Bigoni, and Y. Marzouk, *Inference via low-dimensional couplings*, J. Mach. Learn. Res., 19 (2018), pp. 1–71.

[104] C. Stein, *A bound for the error in the normal approximation to the distribution of a sum of dependent random variables*, in Proc. 6th Berkeley Symp. on Math. Statist. and Prob., Univ. California Press, Berkeley, CA, 1972, pp. 583–602.

[105] ——, *Approximate Computation of Expectations*, vol. 7 of Institute of Mathematical Statistics Lecture Notes—Monograph Series, Institute of Mathematical Statistics, Hayward, CA, 1986.

[106] A. M. Stuart, *Inverse problems: a Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.

[107] T. J. Sullivan, *Introduction to Uncertainty Quantification*, vol. 63 of Texts in Applied Mathematics, Springer, Cham, 2015.

[108] M. A. T. Figueiredo, J. B. Dias, J. P. Oliveira, and R. D. Nowak, *On total variation denoising: a new majorization-minimization algorithm and an experimental comparison with wavalet denoising*, in IEEE Int. Conf. Image Process., 2006, pp. 2633–2636.

[109] R. Tang and Y. Yang, *Adaptivity of diffusion models to manifold structures*, in Proc. 27th Int. Conf. Artif. Intell. Stat., vol. 238 of Proc. Mach. Learn. Res., PMLR, 2024, pp. 1648–1656.

[110] 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.

[111] X. T. Tong, W. Wang, and Y. Wang, *Uniform error bound for PCA matrix denoising*, Bernoulli, 31 (2025), pp. 2251–2275.

[112] A. V. Vecchia, *Estimation and model identification for continuous spatial processes*, J. R. Stat. Soc., 50 (1988), pp. 297–312.

[113] C. Villani, *Optimal Transport: Old and New*, vol. 338 of Fundamental Principles of Mathematical Sciences, Springer-Verlag, Berlin, 2009.

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

[115] D. Wang, Z. Zeng, and Q. Liu, *Stein variational message passing for continuous graphical models*, in Proc. 35th Int. Conf. Mach. Learn., vol. 80 of Proc. Mach. Learn. Res., PMLR, 2018, pp. 5219–5227.

[116] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature, 393 (1998), pp. 440–442.

[117] A. Wibisono, Y. Wu, and K. Y. Yang, *Optimal score estimation via empirical Bayes smoothing*, in Proc. 37th Conf. Learn. Theory, vol. 247 of Proc. Mach. Learn. Res., PMLR, 2024, pp. 4958–4991.

[118] G. Winkler, *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction*, vol. 27 of Applications of Mathematics (New York), Springer-Verlag, Berlin, second ed., 2003.

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

[120] O. Zahm, T. Cui, K. Law, A. Spantini, and Y. Marzouk, *Certified dimension reduction in nonlinear Bayesian inverse problems*, Math. Comp., 91 (2022), pp. 1789–1835.

[121] J. Zhuo, C. Liu, J. Shi, J. Zhu, N. Chen, and B. Zhang, *Message passing Stein variational gradient descent*, in Proc. 35th Int. Conf. Mach. Learn., vol. 80 of Proc. Mach. Learn. Res., PMLR, 2018, pp. 6018–6027.

List of Publications

[A1] **S. Liu**, S. Reich, and X. T. Tong, *Dropout ensemble Kalman inversion for high dimensional inverse problems*, SIAM J. Numer. Anal., 63 (2025), pp. 685–715.

[A2] R. Flock, **S. Liu**, Y. Dong, and X. T. Tong, *Local MALA-within-Gibbs for Bayesian image deblurring with total variation prior*, SIAM J. Sci. Comput., 47 (2025), pp. A2127–A2153.

[A3] N. Liu, **S. Liu**, X. T. Tong, and L. Jiang, *Estimate of Koopman modes and eigenvalues with Kalman filter*, arXiv preprint arXiv:2410.02815, (2024).

[A4] T. Cui, **S. Liu**, and X. T. Tong, *Stein's method for marginals on large graphical models*, arXiv preprint arXiv:2410.11771, (2025).

[A5] G. A. Gottwald, **S. Liu**, Y. Marzouk, S. Reich, and X. T. Tong, *Localized diffusion models for high dimensional distributions generation*, arXiv preprint arXiv:2505.04417, (2025).