Fantastic Diffusion Models and Where to Apply Them

Submitted in partial fulfillment for the requirements for
the degreee of
Doctor of Philosophy
in
Electrical & Computer Engineering

Xingyu Xu

B.S., Electronic Engineering, Tsinghua University

Carnegie Mellon University
Pittsburgh, PA

April 2026

Acknowledgments

The research in this thesis is supported in part by the grants ONR N00014-19-1-2404, NSF DMS-2134080 and ECCS-2126634, AFOSR FA9550-25-1-0060; as well as CIT Dean's Fellowship and the Axel Berny Presidential Graduate Fellowship at Carnegie Mellon University.

I am deeply indebted to my advisor, Professor Yuejie Chi. Before beginning my Ph.D. at Carnegie Mellon University, I could not have imagined such an extraordinary and rewarding journey, and it has been my great privilege to be her student. Her passion for scientific research, depth of insight, remarkable technical sharpness, and impressive sense of academic elegance have been a constant source of inspiration to me. Her guidance, however, has extended far beyond technical advice. She has taught me how to communicate ideas clearly, how to write with precision, how to pursue research with both rigor and taste, and has offered numerous invaluable insights on career development and life more broadly. I am profoundly grateful for her trust, patience, encouragement, and generous support throughout these years.

I would also like to express my sincere gratitude to my co-advisor, Professor Giulia Fanti. Her guidance and support during the final stage of my Ph.D. have meant a great deal to me. I deeply appreciate her kindness, helpfulness, and generosity with her time, as well as the care and thoughtfulness with which she offered her advice on this thesis. I am grateful for her encouragement and support during this important time of my graduate study.

I am very grateful to the other members of my thesis committee, Professor Jeffrey Fessler and Professor Guannan Qu. Professor Fessler made many sharp and insightful observations, and his advice greatly strengthened this thesis. Professor Qu provided valuable suggestions on the technical content and helped me find clearer and more effective ways to present the results. I sincerely appreciate their time, careful reading, and constructive feedback.

Special thanks go to Professor Gen Li, whose encouragement and help

have been crucial in shaping my career path. I have benefited greatly from his advice, support, and perspective on research. His beautiful and profound research has also been a tremendous source of learning and inspiration for me.

I have been fortunate to have wonderful groupmates, colleagues, and friends throughout my Ph.D. I would like to thank Amartya Banerjee, Shicong Cen, Harry Dong, Timofey Efimov, Lingjing Kong, Harlin Lee, Boyue Li, Yuanxin Li, Zhize Li, Laixi Shi, Tian Tong, Pedro Valdeira, He Wang, Zixin Wen, Jiin Woo, Tong Yang, Shuhua Yu, and Ziyi Zhang, among many others, for the many discussions, exchanges of ideas, and moments of support and friendship. I am also grateful to the many collaborators from whom I have learned throughout my Ph.D., including but not limited to Professor Cong Ma, Professor Christopher Musco, Professor Yandi Shen, and Professor Johannes Wiesel. Working with them has broadened my perspective and made my research journey much richer and more enjoyable.

I would also like to thank the faculty, staff, and students in the department and the broader academic community at Carnegie Mellon University. Having often learned most comfortably on my own before coming to CMU, I was pleasantly surprised by how inspiring and thoughtfully taught many courses here were. This led me to take, enjoy, and benefit from a range of courses that I had never expected to be interested in, and these experiences became an important part of my graduate education.

Finally, I owe my deepest gratitude to my family. Their unconditional love, patience, and support have sustained me through every stage of this journey. I am especially grateful to my wife, Wenlin, for her love, understanding, and companionship. This thesis would not have been possible without the encouragement and support of the people around me, and I am truly thankful to all of them.

Abstract

Diffusion models have become the state-of-the-art method for photorealistic image and video generation, yet their usefulness extends far beyond pure data generation. This thesis studies how diffusion models can be used to solve inverse problems in data science. Our central perspective is a plug-and-play Bayesian framework in which a diffusion model serves as a learned prior that can be injected into inverse problem solvers. We develop this framework first for unconstrained inverse problems and then extend it to settings with hard geometric constraints through the language of Riemannian manifolds.

For unconstrained inverse problems, we propose diffusion plug-and-play (DPnP), a general framework that alternates between a denoising diffusion step, which imposes the learned prior, and a proximal consistency step, which enforces measurement fidelity. This framework applies broadly and is supported by both empirical results and theoretical guarantees of correctness and robustness.

To provide deeper theoretical insight into the unconstrained framework, we study sparse recovery as a tractable yet practically relevant problem. In this setting, we show that a simple diffusion-based solver achieves denoising performance comparable to that of an oracle that knows the true sparsity pattern in advance, while running in polynomial time. This result clarifies how diffusion priors can effectively combine learned score-based prior with measurement information in inverse problems.

Motivated by inverse problems with hard structural constraints, we further develop manifold DPnP, which extends DPnP to Riemannian manifolds and enables posterior sampling under geometric constraints. To support this extension theoretically, we also prove the first polynomial iteration-complexity bound for Riemannian diffusion models, showing that the manifold diffusion step underlying manifold DPnP can be implemented efficiently and accurately.

Together, these results establish diffusion models as a mathematically

grounded and algorithmically flexible framework for inverse problems, bridging modern generative modeling, Bayesian inference, and geometric structure in a unified way.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Generative modeling aims to learn complex probability distributions from data and to draw new samples that faithfully reflect the variability present in the underlying population. In recent years, diffusion models [88, 107, 112, 114] have become one of the most successful approaches to this problem, with remarkable empirical performance across a wide range of domains, including image generation, video synthesis, speech modeling, and related tasks [28, 49, 50, 56, 95, 100, 102]. Their success stems not only from their sample quality, but also from the flexibility of the framework: diffusion models admit a clear probabilistic interpretation, connect naturally with stochastic analysis, and can be adapted to settings far beyond unconditional generation.

On a high level, diffusion models generate samples from a target distribution by operating on two stochastic processes:

- A *forward* process, which gradually injects noise to clean data from the target distribution, eventually transforming the data into pure noise.
- A *reverse* process which starts from pure noise, and denoises gradually to arrive at a new sample from a distribution close to the target distribution.

The forward process is typically implemented by a Brownian random walk or an Ornstein-Uhlenbeck process [89]. The construction of the reverse process, on the other hand, requires the theory of time-reversal for stochastic differential equations (SDE) [6, 48]. It is known that the reverse process can be implemented as long as the *score function*, i.e., the log-gradient of the the marginal density of the forward process, is known. In practice, the exact score function is often unavailable and is instead estimated with a neural network via score matching [54]. In this thesis, we

focus on the analysis of diffusion models given a pre-trained inexact estimate of the score function.

Beyond direct sampling from a target distribution on which the score function is trained, we further investigate the use of diffusion models as expressive representations of prior distributions in Bayesian sampling. This perspective is crucial for applying diffusion models to *inverse problems*, where the objective is to recover an unknown signal from noisy observations under an assumed prior on the signal. Such problems arise throughout imaging, scientific computing, and data science, including tasks such as denoising, deblurring, super-resolution, inpainting, tomographic reconstruction, and phase retrieval. They are typically ill-posed: the observations alone do not uniquely determine the unknown signal. Effective recovery therefore requires a prior that captures the structure of plausible solutions.

Diffusion models offer a data-driven alternative to hand-crafted regularization by encoding rich prior information through a learned score function. However, using such learned prior inside inverse problem solvers raises questions that are both algorithmic and theoretical: how should the diffusion prior be coupled with the likelihood, and under what conditions does the resulting method produce reliable solutions? This leads to the question:

How to inject the knowledge learned by a diffusion model into inverse problem solvers in a theoretically justified way?

While diffusion generative models are often formulated in a Euclildean space, many scientific domains are intrinsically *non-Euclidean*; examples include orientations on SO(3), directions on spheres, toroidal angles, articulated poses, and symmetric positive definite (SPD) matrices, which are naturally modeled on Riemannian manifolds [86, 92]. This motivates the study of *Riemannian diffusion models*, in which both the forward and reverse diffusion processes evolve intrinsically on a Riemannian manifold rather than in an ambient Euclidean space. This setting raises a fundamental question:

Can we sample data and solve constrained inverse problems efficiently with Riemannian diffusion models?

The goal of this thesis is to develop a mathematical framework for these questions. Broadly speaking, we view diffusion models as expressive Bayesian priors for inverse problems, use this perspective to design inverse problem solvers with provable

robustness guarantees, prove theoretical guarantees for such solvers, and extend this framework to settings with manifold constraints.

1.1 Solving general inverse problems with diffusion priors

Thanks to the expressive power of score-based diffusion models in generating complex and fine-grained images, they have emerged as a plausible candidate of an expressive prior in image reconstruction [23, 34, 113] via the lens of *Bayesian posterior sampling*. To accommodate diverse applications with various image characteristics and imaging modalities, it is desirable to develop *plug-and-play* methods that do not require training from scratch or end-to-end training for every new imaging task. Nonetheless, despite a flurry of recent efforts, existing algorithms either are computationally expensive [17, 126], inconsistent [23, 62, 83], or confined to linear inverse problems [17, 30]. Therefore, we are in need of a practical, *consistent and robust* algorithm that incorporates score-based diffusion models as an image prior with *general (possibly nonlinear)* forward models.

1.1.1 Our contribution

We develop a new algorithmic framework for solving general inverse problems with diffusion models in a plug-and-play manner. We also provide theoretical results toward understanding how diffusion-based inverse problem solvers use the learned diffusion prior.

A diffusion plug-and-play solver for nonlinear inverse problems. We develop a diffusion plug-and-play framework, denoted by DPnP, for posterior sampling in imaging inverse problems. The framework uses an unconditional score-based diffusion model as an expressive image prior and applies to general, potentially nonlinear, forward models.

The key idea is to decompose posterior sampling into two modular steps. A *proximal consistency sampler* promotes consistency with the measurements using only the likelihood function of the forward model. A *denoising diffusion sampler*

enforces the prior constraint by sampling from the posterior distribution of an easier denoising problem, namely denoising under white Gaussian noise. This separation is the main structural feature of DPnP: the likelihood and the prior are handled by two separate samplers, each depending only on its own component of the model.

We show that the denoising diffusion sampler can be implemented using the standard diffusion-model pipeline, either through stochastic DDPM-type samplers or deterministic DDIM-type samplers, without any additional training. Both variants use the same unconditional score functions as ordinary diffusion-model generation.

We also establish theoretical guarantees for DPnP. In the idealized setting with exact unconditional score functions, we prove that DPnP converges to the desired posterior distribution of the unknown image given the measurements. We further provide non-asymptotic guarantees showing that the method degrades gracefully under sampling and score-estimation errors. To the best of our knowledge, this gives the first provably robust posterior sampling framework for nonlinear inverse problems using unconditional score-based diffusion priors.

Finally, we validate DPnP on both linear and nonlinear image reconstruction tasks, including super-resolution, phase retrieval, and quantized sensing. These experiments illustrate the promise of the proposed plug-and-play approach across a broad class of imaging inverse problems.

Explicit convergence analysis for a diffusion-based solver. In complement to the theoretical guarantee above, we provide a fully explicit convergence analysis under a sparse prior. This model is simple enough to allow sharp estimates, while retaining an important feature shared by many practical inverse problems: the unknown signal has low-dimensional structure, and the diffusion prior represents this structure across multiple noise scales.

Focusing on the denoising problem, we show that a simple diffusion-based heuristic solver can recover the underlying signal with oracle-level accuracy: under a moderately high signal-to-noise ratio condition, its accuracy matches what one would obtain if the sparse support of the signal were known in advance.

This result gives a transparent example of why diffusion-based solvers can succeed when the target signal has low-dimensional structure. It also illustrates a broader principle behind the algorithms studied in this thesis: diffusion priors are useful not merely as generative models, but as representations of prior information that can be

incorporated into inverse problem solvers with theoretical guarantees.

1.2 Riemannian diffusion models

While significant progresses on the convergence analysis of diffusion models defined on a Euclidean space have been made [7, 71, 72], there has been an increasing interest in effectively sampling from distributions supported on manifolds and providing theoretical guarantees [36, 38, 42, 74]. Although sampling on manifolds has been studied extensively [21], extending diffusion models to manifolds requires careful treatments to incorporate the manifold constraints into both the time-inhomogeneous forward and reverse processes, with selected attempts in De Bortoli et al. [27], Fishman et al. [35], Huang et al. [52], Liu et al. [78], Lou et al. [79].

Among these results, a notable development is De Bortoli et al. [27], who introduced *Riemannian Score-Based Generative Models* (RSGMs) with convergence guarantees in the Wasserstein distance. Specifically, they established a time-reversal diffusion process for geometric Brownian motion on manifolds, which can be similarly learned via score matching [54]. While groundbreaking, their convergence bound suffers from a few caveats: (1) it requires an exponentially small stepsize, leading to a possibly exponential iteration complexity in some of the manifold parameters; (2) it requires L_∞-accurate score estimates, which are impractical in deep learning; and (3) the data distribution is required to be smooth and strictly positive on compact manifolds. This calls for *polynomial iteration complexity* for manifold diffusion models using L_2-accurate score estimates under milder data assumptions.

1.2.1 Our contribution

Theoretical foundation of efficiency. We provide a discrete-time analysis of the RSGM sampler proposed in De Bortoli et al. [27], assuming L_2-accurate score estimates. Under mild geometric conditions of the manifold without assuming smooth or strictly positive data densities, we establish that *polynomial* stepsizes suffice for accurate sampling on manifolds in *total variation* (TV). This conveys a much more benign message about the efficiency of Riemannian diffusion models, compared with the iteration complexity in De Bortoli et al. [27] that scales exponentially with the dimension d, under relaxed assumptions on both the data distribution and the score

estimates.

Our proof highlights three ingredients: (i) high-probability Li-Yau gradient bounds for the manifold heat kernel together with early stopping to control $\|\nabla \log p_t\|$ without assuming positivity/smoothness of p_0; (ii) a localization scheme that "freezes" drifts across nearby tangent spaces but preserves continuous Brownian motion (BM), to separate the effects of discretization of scores and BM; and (iii) a quantitative estimates for Minakshisundaram–Pleijel parametrix that controls one-step deviations between the manifold heat flow and its discretized proxy. These components allow us to handle the discretization errors sharply to avoid exponential dependence.

Solving constrained inverse problems with manifold diffusion plug-and-play. Building on our theoretical foundation, we generalize the diffusion plug-and-play framework to Riemannian manifolds, thereby providing a principled approach to constrained inverse problems with diffusion priors. A key observation is that, in the Euclidean setting, our DPnP sampler admits an intrinsic interpretation through its connection with heat flow. More specifically, by establishing an equivalence between heat-flow smoothing and the denoising step implemented by DPnP, we uncover a formulation in terms of only the intrinsic properties of the support of the data prior, and is therefore amenable to extension beyond Euclidean space. This viewpoint leads naturally to a manifold version of DPnP. The resulting method yields a principled framework for posterior sampling under manifold constraints, which retains the same modularity, flexibility, and robustness that make DPnP attractive in the Euclidean setting. This lays the groundwork for diffusion-based algorithms for nonlinear inverse problems with manifold constraints.

1.3 Orgnization and notation

The rest of this thesis is organized as follows.

- Chapter 2 lays down the mathematical basis for diffusion models.
- Chapter 3 introduces the proposed algorithm DPnP and its theoretical guarantee and experimental evaluation [127].
- Chapter 4 establishes explicit convergence rates of diffusion-based inverse problem solvers under sparse priors.

- Chapter 5 introduces our convergence analysis of Riemannian diffusion models [128].
- Chapter 6 presents the extension of DPnP to Riemannian manifolds and its application in solving constrained inverse problems.

Notation. We introduce some key notation used throughout the thesis. Let p_x denote the probability distribution of x, and $p_x(\cdot|y)$ denotes the conditional distribution of x given y. We use $X \overset{(\mathrm{d})}{=} Y$ to denote random variables X and Y are equivalent in distribution. The matrix I_d denotes an identity matrix of dimension d. For two probability distributions with density $p(x)$ and $q(x)$, the total variation distance between them is

$$\mathsf{TV}(p, q) \coloneqq \int |p(x) - q(x)| \mathrm{d}x.$$

The χ^2-divergence of p to q is

$$\chi^2(p \,\|\, q) \coloneqq \int \frac{(p(x) - q(x))^2}{q(x)} \mathrm{d}x.$$

The Kullback-Leibler (KL) divergence of p to q is

$$\mathsf{KL}(p\|q) = \int p(x) \left(\log \frac{p(x)}{q(x)} \right) \mathrm{d}x.$$

We assume some familiarity with Riemannian geometry, and make use of standard notation. Please refer to Jost [58], Petersen [91] for a more in-depth treatment. In particular, we use $\alpha, \beta, \xi, \zeta$, etc., to index *coordinate representation* of tensors, and assume Einstein's summation convention. Let $(\mathcal{M}, g)$ be a connected, compact d-dimensional Riemannian manifold, with geodesic distance $\rho(\cdot, \cdot)$ and volume measure μ. We assume $\mu(\mathcal{M}) = 1$. The Levi–Civita connection is denoted by ∇, and the Laplace–Beltrami operator by

$$\Delta_{\mathcal{M}} f \coloneqq \nabla_\alpha \nabla^\alpha f.$$

We use $T_x\mathcal{M}$ for the tangent space at x and use $\exp_x : T_x\mathcal{M} \to \mathcal{M}$ for the exponential map and $\log_x$ for its local inverse on the normal neighborhood of x. The

geodesic diameter of $(\mathcal{M}, g)$ is defined as

$$\mathrm{Diam}(\mathcal{M}) := \sup_{x,y\in\mathcal{M}} \rho(x, y),$$

We further denote Rm as the Riemannian curvature tensor. Geodesic ball centered at x with radius r is denoted $B_x(r)$.

Chapter 2

Background on Diffusion Models

In this chapter, we introduce the preliminaries on diffusion-based generative models, which will serve as the foundation for our theory and algorithms. The key components consist of a *forward* process, which diffuses the data distribution $p^\star$ to the standard normal distribution by gradually injecting noise into the samples, and a *backward* process, which reverses the forward process so that it can transform the standard normal distribution to the data distribution $p^\star$.

2.1 Score-based generative models

Consider the forward Markov process in $\mathbb{R}^d$ that starts with a sample from the data distribution $p^\star$, and adds noise over the trajectory according to

$$x_0 \sim p^\star, \tag{2.1a}$$

$$x_k = \sqrt{1-\beta_k}\, x_{k-1} + \sqrt{\beta_k}\, w_k, \qquad 1 \le k \le T, \tag{2.1b}$$

where $\{w_k\}_{1\le k\le T}$'s are independent standard Gaussian vectors, i.e., $w_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, and $\{\beta_k \in (0,1)\}$ describes the noise-injection rates used in each step. Therefore, we can write x_k equivalently as

$$x_k \coloneqq \sqrt{\bar\alpha_k}\, x_0 + \sqrt{1-\bar\alpha_k}\, \varepsilon_k, \quad \varepsilon_k \sim \mathcal{N}(0, I_d), \quad k = 0, 1, \cdots, T. \tag{2.2}$$

Here, $(\bar\alpha_k)_{k=0,1,\cdots,T}$ is the *schedule* of diffusion given by

$$\alpha_k \coloneqq 1 - \beta_k, \qquad \bar{\alpha}_k \coloneqq \prod_{\ell=1}^{k} \alpha_\ell, \qquad 1 \le k \le T. \tag{2.3}$$

Clearly, it verifies that $1 \ge \bar{\alpha}_0 > \bar{\alpha}_1 > \cdots > \bar{\alpha}_T > 0$. As long as $\bar{\alpha}_T$ is vanishing, it is easy to observe that the distribution of x_T approaches $\mathcal{N}(0, I_d)$.

Score functions. As will be seen, to sample from $p^\star$, it turns out to be sufficient to learn the score functions of p_{x_t} at each step of the forward process [19, 114], defined as

$$s^\star_{t_k}(x) = \nabla \log p_{x_k}(x), \qquad k = 0, 1, \cdots, T, \quad t_k = \frac{1}{2} \log \frac{1}{\bar{\alpha}_k}. \tag{2.4}$$

We will see momentarily the rationale behind the definition of t_k.

Continuous-time perspective. To facilitate understanding, it will be convenient to formulate the diffusion process in continuous time. To distinguish from the discrete-time setting, we use capitalized letters like X to denote the continuous-time diffusion process. The continuous-time forward diffusion follows the Ornstein-Uhlenbeck (OU) process[1], defined by the Stochastic Differential Equation (SDE) [114]:

$$\mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}\, \mathrm{d}B_t, \quad t \ge 0, \quad X_0 \sim p^\star, \tag{2.5}$$

where $(B_t)_{t \ge 0}$ is the standard d-dimensional Brownian motion. It can be shown that [29, 32] the marginal distribution of X_t for $t \ge 0$ is

$$X_t \overset{(\mathrm{d})}{=} \mathrm{e}^{-t} X_0 + \sqrt{1 - \mathrm{e}^{-2t}}\varepsilon, \quad X_0 \sim p^\star, \ \varepsilon \sim \mathcal{N}(0, I_d). \tag{2.6}$$

It is then clear that the limiting distribution $X_\infty \sim \mathcal{N}(0, I_d)$ as $\tau \to \infty$, i.e., the OU process diffuses $X_0 \sim p^\star$ to the standard normal distribution. The score function of X_t is defined by

$$s_t(x) = \nabla \log p_{X_t}(x). \tag{2.7}$$

An enlightening property [122] of the score function is that it can be interpreted as the minimum mean-squared error (MMSE) estimate of ε given $X_t = x$, fueled by

[1] In the literature, other processes such as Variance-Exploding SDE (VE-SDE) are also used. The framework in this chapter also applies to these processes with straightforward modifications.

Tweedie's formula:

$$s_t(x) = -\frac{1}{\sqrt{1-\mathrm{e}^{-2t}}} \underbrace{\mathbb{E}_{X_0\sim p^\star,\,\varepsilon\sim\mathcal{N}(0,I_d)}(\varepsilon \,|\, \mathrm{e}^{-t}X_0 + \sqrt{1-\mathrm{e}^{-2t}}\varepsilon = x)}_{=:\varepsilon_t(x)} \tag{2.8}$$

Consequently, this makes it possible to estimate the score functions via learning to denoise [54], by estimating the denoising function ε, as typically done in practice [49].

2.2 The reverse process and sampling

To enable sampling, one needs to "reverse" the forward diffusion process. Fortunately, it is possible to leverage classical theory [5, 6] to reverse the SDE, and apply discretization to the time-reversal processes to collect samples. We shall describe two popular approaches below, corresponding to stochastic (i.e., DDPM-type [49]) and deterministic (i.e., DDIM-type [109]) samplers respectively following primarily the framework set forth in Song et al. [114].

Time-reversed SDEs and probability flow ODEs. Let us begin with the more general theory of *reversing* SDEs, which will be useful in future sections. Consider a SDE given by

$$\mathrm{d}M_t = \alpha M_t \mathrm{d}t + \sqrt{\beta}\mathrm{d}B_t, \quad t \geq 0, \quad M_0 \sim p_{M_0}, \tag{2.9}$$

where $\alpha \in \mathbb{R}$ and $\beta > 0$ are constants. For any positive time $\tau_\infty > 0$, define the reversed time parameter

$$\tau := \tau(t) = \tau_\infty - t. \tag{2.10}$$

We are now ready to describe the time-reversed processes.

1) The *time-reversed SDE* of (2.9) on the time interval $[0, \tau_\infty]$ is defined as

$$\mathrm{d}M^{\mathsf{rev}}_\tau = (-\alpha M^{\mathsf{rev}}_\tau + \beta\nabla \log p_{M_\tau}(M^{\mathsf{rev}}_\tau))\,\mathrm{d}t + \sqrt{\beta}\mathrm{d}\tilde{B}_t,\ t \in [0,\tau_\infty],\ M^{\mathsf{rev}}_{\tau_\infty} \sim p_{M_{\tau_\infty}}, \tag{2.11}$$

where $\tilde{B}$ is an independent copy of B, i.e., another Brownian motion. It is a classical result [6] that the reversed process M^{rev} shares the same path distribution as M, i.e., $(M^{\mathsf{rev}}_\tau)_{\tau\in[0,\tau_\infty]} \overset{(\mathrm{d})}{=} (M_\tau)_{\tau\in[0,\tau_\infty]}$. In other words, the joint distribution of $(M^{\mathsf{rev}}_{\tau_1}, M^{\mathsf{rev}}_{\tau_2}, \cdots, M^{\mathsf{rev}}_{\tau_k})$ for any $0 \leq \tau_1 \leq \tau_2 \leq \cdots \leq \tau_k \leq \tau_\infty$, for any integer $k \geq 1$,

coincides with that of $(M_{\tau_1}, M_{\tau_2}, \cdots, M_{\tau_k})$.

2) In place of the reversed SDE in (2.11), it is possible to consider the following probability flow ODE [5, 114]:

$$\mathrm{d}M_\tau^{\mathsf{rev}} = \left(-\alpha M_\tau^{\mathsf{rev}} + \frac{\beta}{2}\nabla \log p_{M_\tau}(M_\tau^{\mathsf{rev}})\right)\mathrm{d}t,\ t \in [0, \tau_\infty],\ M_{\tau_\infty}^{\mathsf{rev}} \sim p_{M_{\tau_\infty}}. \quad (2.12)$$

The reversed ODE satisfies a slightly weaker guarantee than that of the reversed SDE, which nevertheless suffices for most practical purposes [114]: $M_\tau^{\mathsf{rev}} \overset{(\mathrm{d})}{=} M_\tau, \quad \tau \in [0, \tau_\infty]$. Note that the reversed ODE only guarantees identical marginal distribution for each M_τ^{rev}, whereas the reversed SDE guarantees identical joint distribution.

Specializing the above to the OU process (2.5) with proper discretization then leads to popular samplers used for generation, as follows.

DDPM-type stochastic samplers. Specializing the time-reversed SDE (2.11) to the OU process gives

$$\mathrm{d}X_\tau^{\mathsf{rev}} = (X_\tau^{\mathsf{rev}} + 2s_\tau(X_\tau^{\mathsf{rev}}))\mathrm{d}t + \sqrt{2}\mathrm{d}\tilde{B}_t, \quad t \in [0, \tau_\infty], \quad X_{\tau_\infty}^{\mathsf{rev}} \sim p_{X_{\tau_\infty}}.$$

As $\tau_\infty \to \infty$, it can be seen from (2.6) that $p_{X_{\tau_\infty}}$ converges to $\mathcal{N}(0, I_d)$. Thus the solution of the above SDE can be approximated by initializing $X_{\tau_\infty}^{\mathsf{rev}} \sim \mathcal{N}(0, I_d)$ instead. The DDPM sampler [49] can be viewed as a discretization of this SDE [114].

DDIM-type deterministic samplers. On the other hand, the probability flow ODE (2.12) for the OU process reads as

$$\mathrm{d}X_\tau^{\mathsf{rev}} = (X_\tau^{\mathsf{rev}} + s_\tau(X_\tau^{\mathsf{rev}}))\mathrm{d}t, \quad t \in [0, \tau_\infty], \quad X_{\tau_\infty}^{\mathsf{rev}} \sim p_{X_{\tau_\infty}}. \quad (2.13)$$

Again, as $\tau_\infty \to \infty$, one may approximate the initialization with $X_{\tau_\infty}^{\mathsf{rev}} \sim \mathcal{N}(0, I_d)$. It is known that the popular DDIM sampler [109, 114] is a discretization of this ODE [131]. The ODE-based deterministic samplers allow more aggressive choice of discretization schedules, as well as fast ODE solvers [80], enabling significantly accelerated sampling process compared to the SDE-based stochastic samplers.

Chapter 3

DPnP: Robust Nonlinear Inverse Problem Solvers with Diffusion Priors

In this chapter, we introduce DPnP as a plug-and-play inverse problem solver that enjoys several important theoretical guarantees. This chapter is based on [127].

3.1 Posterior sampling for inverse problems

We are interested in solving (possibly nonlinear) inverse problems, where the aim is to infer an unknown image $x^\star \in \mathbb{R}^d$ from its measurements $y \in \mathbb{R}^m$,[1] given by

$$y = \mathcal{A}(x^\star) + \xi,$$

where $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^m$ is the measurement operator underneath the forward model, and xi denotes measurement noise. It has been well-understood that *prior information* of $x^\star$ plays an important role in solving inverse problems that are otherwise ill-posed, enabling successful reconstruction with much less measurements and higher accuracy. At the same time, it is desirable to understand and quantify the uncertainty in image reconstruction, especially when the available measurements are rather limited.

Posterior sampling. In this work, we focus on the Bayesian setting where the

[1] For simplicity, we limit our presentation to the real-valued case; our framework generalizes to the complex-valued case in a straightforward manner.

prior information of $x^\star$ is provided in the form of some prior distribution $p^\star(\cdot)$, i.e.,

$$x^\star \sim p^\star(x), \tag{3.1}$$

The *posterior distribution* given measurements y is defined as

$$p^\star(x|y) \propto p^\star(x)\, p(y|x^\star = x) = p^\star(x)\, \mathrm{e}^{\mathcal{L}(x;y)}. \tag{3.2}$$

Here, $\mathcal{L}(\cdot; y)$ is the log-likelihood function of the measurements. For example, when the noise $\xi \sim \mathcal{N}(0, \sigma^2 I_m)$ is standard Gaussian, it follows that

$$\mathcal{L}(x; y) = -\frac{1}{2\sigma^2}\|y - \mathcal{A}(x)\|^2 - \frac{m}{2}\log(2\pi\sigma^2).$$

Notwithstanding, our framework allows flexible choices of the forward model and the noise distributions. In addition, while this formulation is derived from probabilistic interpretations, it also subsumes the "reward-guided" or "loss-guided" setting [111], where $\mathcal{L}$ can be viewed as a reward function or a negative loss function, both of which characterize preference over structural properties of $x^\star$. In all these settings, it will be useful to bear in mind the intuition that the higher value of $\mathcal{L}$ corresponds to better consistency with the measurements, higher rewards, etc.

Assumption on the forward model. Throughout the chapter, for simplicity, we make the following mild assumption on $\mathcal{L}$, which is applicable to many applications of interest.

Assumption 1. *We assume $\mathcal{L}(\cdot\,; y)$ is differentiable almost everywhere, and*

$$\sup_{x \in \mathbb{R}^d} \mathcal{L}(x; y) < \infty.$$

Goal. Our goal is to sample $\hat{x}$ from the posterior distribution

$$\hat{x} \sim p^\star(\cdot \,|\, y)$$

given estimates $\hat{s}_t(x)$ (resp. $\hat{\varepsilon}_t(x)$) of the *unconditional* score functions $s_t^\star(x)$ (resp. the noise function $\varepsilon_t^\star(x)$) in (2.4), assuming knowledge of the likelihood function $\mathcal{L}(\cdot; y)$.

3.2 Key ingredient: score-based denoising posterior sampling

We begin with one of the most fundamental inverse problems: denoising under white Gaussian noise. We demonstrate how to solve this problem via stochastic (i.e., DDPM-type) and deterministic (i.e., DDIM-type) denoising diffusion samplers using the same set of unconditional score functions trained for generation. As shall be elucidated shortly, the denoising diffusion samplers turn out to be an important building block in our algorithm for general inverse problems.

Image denoising under white Gaussian noise. Suppose that we have access to a noisy version of $x^\star \sim p^\star$ contaminated by white Gaussian noise, given by

$$x_{\mathsf{noisy}} = x^\star + \xi, \quad \xi \sim \mathcal{N}(0, \eta^2 I_d), \tag{3.3}$$

where $\eta > 0$ is the noise intensity *assumed to be known*. Our goal is to sample from $p^\star(\cdot \,|\, x_{\mathsf{noisy}})$ given the score estimates $\hat{s}_t(x)$ (resp. the noise estimates $\hat{\varepsilon}_t(x)$). We will develop our score-based denoising posterior sampler, termed DDS, with two variants, DDS-DDPM and DDS-DDIM, which can be viewed as analogues of the well-known DDPM and DDIM samplers in unconditional score-based sampling respectively. Before proceeding, it is worth highlighting that the two variants will be derived from different forward diffusion processes, since we observe the resulting variants empirically lead to more competitive performance.

A stochastic DDPM-type sampler via heat flow. We begin with a stochastic DDPM-type sampler for denoising, termed DDS-DDPM. We divide our development into the following steps.

- *Step 1: introducing the heat flow.* Let us introduce a *heat flow* with initial distribution $p^\star$, defined by the following SDE:

$$\mathrm{d}Y_t = \mathrm{d}B_t, \quad t \geq 0, \quad Y_0 \sim p^\star, \tag{3.4}$$

  where $(B_t)_{t\geq 0}$ is the standard d-dimensional Brownian motion. The solution of (3.4) is simply

$$Y_t = Y_0 + B_t, \quad t \geq 0. \tag{3.5}$$

Since $B_t \sim \mathcal{N}(0, tI_d)$, it readily follows that $B_{\eta^2} \overset{(\mathrm{d})}{=} \xi$, which together with $Y_0 \sim p^\star$ yield the important observation that $x_{\mathsf{noisy}} = x^\star + \xi$ can be viewed as an endpoint of the heat flow, in the sense that

$$x_{\mathsf{noisy}} = x^\star + \xi \overset{(\mathrm{d})}{=} Y_{\eta^2}.$$

- *Step 2: reversing the heat flow.* Following the framework in Section 2.1 and Section 2.2, the next step boils down to reverse the heat flow (3.4). The time-reversal of the heat flow SDE (3.4) is (cf. (2.11)) given by

$$\mathrm{d}Y^{\mathsf{rev}}_{\eta^2-t} = \nabla \log p_{Y_{\eta^2-t}}(Y^{\mathsf{rev}}_{\eta^2-\tau})\mathrm{d}t + \mathrm{d}\tilde{B}_t, \quad t \in [0, \eta^2], \quad Y^{\mathsf{rev}}_{\eta^2} \sim p_{Y_{\eta^2}}, \tag{3.6}$$

  where $(\tilde{B}_t)_{t\geq 0}$ is an independent copy of $(B_t)_{t\geq 0}$. As introduced earlier, the virtue of the time-reversed SDE (3.6) is that it produces a process $(Y^{\mathsf{rev}}_t)_{0\leq t\leq \eta^2}$ with the same *path* distribution as $(Y_t)_{0\leq t\leq \eta^2}$, i.e.,

$$(Y^{\mathsf{rev}}_t)_{t\in[0,\eta^2]} \overset{(\mathrm{d})}{=} (Y_t)_{t\in[0,\eta^2]}.$$

  In particular, the joint distribution of $(Y^{\mathsf{rev}}_0, Y^{\mathsf{rev}}_{\eta^2})$ is the same as that of $(Y_0, Y_{\eta^2}) \overset{(\mathrm{d})}{=} (x^\star, x_{\mathsf{noisy}})$. This implies that the conditional distribution $p^\star(\cdot \,|\, x_{\mathsf{noisy}})$ is the same as $p_{Y^{\mathsf{rev}}_0}(\cdot \,|\, Y^{\mathsf{rev}}_{\eta^2} = x_{\mathsf{noisy}})$. Surprisingly, the latter admits a simple interpretation: $p_{Y^{\mathsf{rev}}_0}(\cdot \,|\, Y^{\mathsf{rev}}_{\eta^2} = x_{\mathsf{noisy}})$ is the distribution of Y^{rev}_0 when we initialize (3.6) with $Y^{\mathsf{rev}}_{\eta^2} = x_{\mathsf{noisy}}$! Therefore, sampling the posterior $p^\star(\cdot \,|\, x_{\mathsf{noisy}})$ amounts to solving the following simple SDE:

$$\mathrm{d}Y^{\mathsf{rev}}_{\eta^2-t} = \nabla \log p_{Y_{\eta^2-t}}(Y^{\mathsf{rev}}_{\eta^2-t})\mathrm{d}t + \mathrm{d}\tilde{B}_t, \quad t \in [0, \eta^2], \quad Y^{\mathsf{rev}}_{\eta^2} = x_{\mathsf{noisy}}. \tag{3.7}$$

- *Step 3: connecting the score functions.* It is now immediate to arrive at our proposed stochastic sampler DDS-DDPM by discretization of this SDE (3.7), which requires knowledge of the score functions $\nabla \log p_{Y_t}(\cdot)$. A key observation is that they can in fact be computed from the score function s_t (cf. (2.7)), thanks to the following lemma.

Lemma 1 (Score function of Y_t). *For $t \geq 0$, we have*

$$\nabla \log p_{Y_t}(x) = \frac{1}{\sqrt{1+t}} s_{\frac{1}{2}\log(1+t)}\left(\frac{x}{\sqrt{1+t}}\right).$$

Details on the procedure of discretization can be found in Appendix A.2. The resulting sampler, DDS-DDPM, is summarized in Algorithm 1.

Algorithm 1 Denoising Diffusion Sampler (stochastic) DDS-DDPM($x_{\text{noisy}}, \hat{s}, \eta$)

Input: noisy data $x_{\text{noisy}} \in \mathbb{R}^d$, score estimates $\hat{s} := \{\hat{s}_{t_\ell}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d,\ \ell = 1, \ldots, T\}$, where $t_\ell = \log(1/\bar{\alpha}_\ell)$ as in (2.3)–(2.4), and noise level $\eta > 0$.
Scheduling: Compute the diffusion schedule $(\tau_\ell)_{0 \leq \ell \leq T'}$ by

$$\tau_\ell = \bar{\alpha}_\ell^{-1} - 1, \quad 0 \leq \ell \leq T',$$

where

$$T' := \max\left\{t : 0 \leq \ell \leq T,\ \bar{\alpha}_\ell > \frac{1}{\eta^2+1}\right\}.$$

Initialization: Set $\hat{x}_{T'} = x_{\text{noisy}}$.
Diffusion: **for** $\ell = T', T'-1, \ldots, 1$ **do**

$$\hat{x}_{\ell-1} = \hat{x}_\ell - 2(\sqrt{\tau_\ell} - \sqrt{\tau_{\ell-1}})\,\hat{\varepsilon}_{t_\ell} + \sqrt{\tau_\ell - \tau_{\ell-1}}\, w_\ell, \quad w_\ell \sim \mathcal{N}(0, I_d).$$

where

$$\hat{\varepsilon}_{t_\ell} := -\frac{1}{\sqrt{1-\bar{\alpha}_\ell}} \hat{s}_{t_\ell}\left(\sqrt{\bar{\alpha}_\ell}\,\hat{x}_\ell\right).$$

Output: $\hat{x}_0$.

A deterministic DDIM-type sampler via OU process. We next develop a deterministic DDIM-type sampler for denoising, termed DDS-DDIM, presented in Algorithm 2.

- *Step 1: introducing a posterior-initialized OU process.* To sample from the posterior distribution $p^\star(\cdot|x_{\text{noisy}})$, we first introduce a random variable w which has (unconditional) distribution

$$p_w(x) := p^\star(x^\star = x \,|\, x^\star + \xi = x_{\text{noisy}}), \tag{3.8}$$

in the same form of the desired posterior distribution $p^\star(\cdot|x_{\mathsf{noisy}})$. Here, since the noisy observation x_{noisy} is given, we regard it as fixed.[2] We further introduce $z = w - x_{\mathsf{noisy}}$, which is a "centered" version of w, whose distribution is

$$p_z(x) \coloneqq p_w(x + x_{\mathsf{noisy}}) = p^\star(x^\star = x + x_{\mathsf{noisy}} \,|\, x^\star + \xi = x_{\mathsf{noisy}}).$$

The OU process with initial distribution p_z is defined by the SDE:

$$\mathrm{d}Z_t = -Z_t \mathrm{d}t + \mathrm{d}B_t, \quad t \geq 0, \quad Z_0 \sim p_z, \tag{3.9}$$

where $(B_t)_{t\geq 0}$ is the standard d-dimensional Brownian motion. As in (2.6), the marginal distribution of Z_τ is given by

$$Z_t \overset{(\mathrm{d})}{=} \mathrm{e}^{-t} Z_0 + \sqrt{1 - \mathrm{e}^{-2t}}\varepsilon, \quad Z_0 \sim p_z,\ \varepsilon \sim \mathcal{N}(0, I_d), \quad \tau \geq 0. \tag{3.10}$$

- *Step 2: reversing the OU process.* Following the framework in Section 2.1 and Section 2.2, reversing the OU process (3.9) will enable us to generate samples $z \sim p_z$. Then we can set $w = z + x_{\mathsf{noisy}}$, which, by definition, has distribution p_w defined in (3.8), and is a sample from the desired posterior distribution $p^\star(\cdot|x_{\mathsf{noisy}})$. We are thus led to solve the time-reversed probability flow ODE (cf. (2.12)) of (3.9), given by

$$\mathrm{d}Z^{\mathsf{rev}}_\tau = \big(Z^{\mathsf{rev}}_\tau + \nabla \log p_{Z_\tau}(Z^{\mathsf{rev}}_\tau)\big)\mathrm{d}t, \quad t \in [0, \tau_\infty], \quad Z^{\mathsf{rev}}_{\tau_\infty} \sim \mathcal{N}(0, I_d), \quad \tau = \tau_\infty - t. \tag{3.11}$$

- *Step 3: connecting the score functions.* We are now one step away from our proposed deterministic sampler DDS-DDIM, which is derived by discretization of the ODE (3.11). We need to know the score functions $\nabla \log p_{Z_\tau}(\cdot)$, which again can be computed from the score function s_t (cf. (2.7)), as documented by the following lemma.

  Lemma 2 (Score function of Z_t)**.** *For $t \geq 0$, we have*

$$\nabla \log p_{Z_t}(x) = -\frac{\mathrm{e}^{2t} x}{\eta^2 + \mathrm{e}^{2t} - 1} + \frac{\mathrm{e}^{t-\tilde{t}}\eta^2}{\eta^2 + \mathrm{e}^{2t} - 1} s_{\tilde{t}}\left(\mathrm{e}^{-\tilde{t}} x_{\mathsf{noisy}} + \frac{\mathrm{e}^{t-\tilde{t}}\eta^2 x}{\eta^2 + \mathrm{e}^{2t} - 1}\right), \tag{3.12}$$

[2] Technically, this can done by conditioning on x_{noisy} throughout our discussion of DDS-DDIM.

where

$$\tilde{t} := \tilde{t}(t) = \frac{1}{2}\log\left(\frac{\eta^2(\mathrm{e}^{2t}-1)}{\eta^2+\mathrm{e}^{2t}-1}+1\right). \tag{3.13}$$

After plugging this into (3.11) and solving the ODE for Z_τ^{rev}, we see that $Z_0^{\mathsf{rev}} + x_{\mathsf{noisy}}$ is the desired sample from the posterior distribution $p^\star(\cdot|x_{\mathsf{noisy}})$, as argued before. Numerically, the ODE (3.11) is solved by discretization with an exponential integrator [131], resulting in the sampler DDS-DDIM as summarized in Algorithm 2.

Algorithm 2 Denoising Diffusion Sampler (deterministic) DDS-DDIM$(x_{\mathsf{noisy}}, \hat{s}, \eta)$

Input: noisy data $x_{\mathsf{noisy}} \in \mathbb{R}^d$, score estimates $\hat{s} := \{\hat{s}_{t_\ell}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d,\ \ell = 1,\dots,T\}$ and noise level $\eta > 0$.
Scheduling: Compute the diffusion schedule $(\bar{u}_\ell)_{0\le\ell\le T'}$ by

$$\bar{u}_\ell = \frac{(\eta^2+1)\bar{\alpha}_\ell - 1}{\eta^2 + \bar{\alpha}_\ell - 1}, \quad 0 \le k \le T',$$

where

$$T' := \max\left\{\ell : 0 \le \ell \le T,\ \bar{\alpha}_\ell > \frac{1}{\eta^2+1}\right\}.$$

Initialization: Draw $z_{T'} \sim \mathcal{N}(0, I_d)$.
Diffusion: **for** $\ell = T', T'-1, \dots, 1$ **do**

$$z_{\ell-1} = \frac{\sqrt{(\eta^2-1)\bar{u}_{\ell-1}+1}}{\sqrt{(\eta^2-1)\bar{u}_\ell+1}} z_\ell + \sqrt{(\eta^2-1)\bar{u}_{\ell-1}+1}\cdot\big(h(\eta,\bar{u}_{\ell-1}) - h(\eta,\bar{u}_\ell)\big)\hat{\varepsilon}_{t_\ell},$$

where

$$h(\eta,u) := -\arctan\frac{\eta}{\sqrt{u^{-1}-1}},$$
$$\hat{\varepsilon}_{t_\ell} := -\frac{1}{\sqrt{1-\bar{\alpha}_\ell}}\hat{s}_{t_\ell}\left(\sqrt{\bar{\alpha}_\ell}x_{\mathsf{noisy}} + \frac{\eta^2\sqrt{\bar{u}_\ell\bar{\alpha}_\ell}z_\ell}{(\eta^2-1)\bar{u}_\ell+1}\right).$$

Output: $x_{\mathsf{noisy}} + z_0$.

Algorithm 3 Diffusion Plug-and-Play (DPnP)

Input: Measurements $y \in \mathbb{R}^m$, log-likelihood function $\mathcal{L}(\cdot; y)$ of the forward model, score estimates $\{\hat{s}_t(\cdot)\}$, annealing schedule $(\eta_k)_{0 \le k \le K}$.
Initialization: Sample $\hat{x}_0 \sim \mathcal{N}(0, \frac{\eta_0}{4} I_d)$
Alternating sampling: **for** $k = 0, 1, 2, \ldots, K-1$ **do**

(1) *Proximal consistency sampler:* Sample

$$\hat{x}_{k+\frac{1}{2}} \propto \exp\left(\mathcal{L}(\cdot\,;\, y) - \frac{1}{2\eta_k^2}\|\cdot - \hat{x}_k\|^2\right)$$

using subroutine $\mathsf{PCS}(\hat{x}_k, y, \mathcal{L}, \eta_k)$.

(2) *Denoising diffusion sampler:* Sample

$$\hat{x}_{k+1} \sim \exp\left(\log p^\star(x) - \frac{1}{2\eta_k^2}\|x - \hat{x}_{k+\frac{1}{2}}\|^2\right)$$

using subroutine $\mathsf{DDS\text{-}DDPM}(\hat{x}_{k+\frac{1}{2}}, \hat{s}, \eta_k)$ or $\mathsf{DDS\text{-}DDIM}(\hat{x}_{k+\frac{1}{2}}, \hat{s}, \eta_k)$.

Output: $\hat{x}_K$.

3.3 Our algorithm: diffusion plug-and-play

Now we turn to the general setting where the measurement operator $\mathcal{A}$ is arbitrary. From the factorization of posterior distribution in (3.2), one intuitively understands that a posterior sampler must balance two sources of information: (i) the *data prior*, corresponding to the first factor $p^\star(x)$, which imposes that the posterior sampler should be less likely to sample at those points where $p^\star(x)$ is small; (ii) the *measurement consistency*, corresponding to the second factor $\mathrm{e}^{\mathcal{L}(x;y)}$, which imposes that $\mathcal{A}(x) \approx y$.

A prelude: proximal gradient method. Informally speaking, bridging the perspective of optimization and sampling [125], posterior sampling can be viewed as a "soft" solution to the following optimization problem:

$$\max_{x \in \mathbb{R}^d} \; \mathcal{L}(x; y) + \log p^\star(x). \tag{3.14}$$

Instead of producing the point estimate that maximizes $\mathcal{L}(x; y) + \log p^\star(x)$, posterior sampling produces samples from the posterior distribution $p^\star(x)\mathrm{e}^{\mathcal{L}(x;y)} = \mathrm{e}^{\mathcal{L}(x;y)+\log p^\star(x)}$ instead, which allows characterizing the underlying uncertainty. For

better understanding, it is useful to bear in mind the special case when the image prior $p^\star$ is supported on some low-dimensional manifold[3] $\mathcal{M}$. In this setting, we notice that $\log p^\star(x) = -\infty$ for $x \notin \mathcal{M}$, hence the optimization problem (3.14) is implicitly constrained in $x \in \mathcal{M}$.

Recall the well-known proximal gradient method [90] for solving (3.14), where one initializes a random $\hat{x}_0 \in \mathbb{R}^d$ and uses the following update rule

$$\hat{x}_{k+1} = \mathrm{Prox}_{\mathcal{M},\eta_k}\left(\hat{x}_k + \eta_k \nabla_{\hat{x}_k} \mathcal{L}(\hat{x}_k; y)\right), \quad k = 0, 1, \ldots.$$

Here, $\eta_k > 0$ is the stepsize at the k-th iteration, and $\mathrm{Prox}_{\mathcal{M},\eta} : \mathbb{R}^d \to \mathbb{R}^d$ is the proximal operator defined by

$$\mathrm{Prox}_{\mathcal{M},\eta}(x) \coloneqq \operatorname*{argmin}_{x' \in \mathbb{R}^d} \; -\log p^\star(x') + \frac{1}{2\eta^2}\|x' - x\|^2. \tag{3.15}$$

Intuitively, one may view $\mathrm{Prox}_{\mathcal{M},\eta}(x)$ as some kind of denoising to make x more consistent with its structural property [120]. The proximal gradient method alternatively applies two operations:

(i) *Gradient step to enforce the measurement consistency.* The gradient step tries to boost consistency $y \approx \mathcal{A}(x)$ via moving along the direction to increase $\mathcal{L}(\cdot; y)$.

(ii) *Proximal mapping to enforce the data prior.* The proximal step moves the iterate towards those points that increase $\log p^\star(x)$. In particular, when $p^\star$ is supported on a low-dimensional manifold $\mathcal{M}$, the proximal map forces x to reside in $\mathcal{M}$.

Diffusion plug-and-play (DPnP). Although the proximal gradient method does not apply to the posterior sampling problem directly, we borrow its splitting principle from a sampling perspective. Namely, we alternately incorporate the prior information and the measurement likelihood, in the same spirit of Bouman and Buzzard [11], Lee et al. [67], Vono et al. [123]. Our algorithm, dubbed diffusion plug-and-play (DPnP), alternates between two samplers, the denoising diffusion sampler (DDS) and the proximal consistency sampler (PCS), which can be viewed as

[3]This assumption, known as the manifold hypothesis, is commonly adopted as a flexible structural characterization of high-dimensional data. We mention it here to facilitate the understanding of our design, which will not be imposed in our algorithm or analysis for this chapter.

the substitutes for the proximal operator and the gradient step respectively. Given the iterate $\hat{x}_k$ and *annealing* parameter η_k at the k-th iteration, DPnP proceeds with the following two steps:

(i) *Proximal consistency sampler to enforce the measurement consistency.* DPnP draws a sample $\hat{x}_{k+\frac{1}{2}}$ from the distribution proportional to

$$\exp\left(\mathcal{L}(x\,;\,y) - \frac{1}{2\eta_k^2}\|x - \hat{x}_k\|^2\right)$$

to promote the image to be consistent with the measurements. This step, which we denote as the *proximal consistency sampler*, can be achieved by small modifications of standard algorithms such as Metropolis-Adjusted Langevin Algorithm (MALA) [98] given in Algorithm 4.

(ii) *Denoising diffusion sampler to enforce the data prior.* DPnP next draws a sample $\hat{x}_{k+1}$ from the distribution proportional to

$$\begin{aligned}
&\exp\left(-\big(-\log p^\star(x) + \frac{1}{2\eta_k^2}\|x - \hat{x}_{k+\frac{1}{2}}\|^2\big)\right) \\
&\quad\propto p^\star(x)\mathrm{e}^{-\frac{1}{2\eta_k^2}\|x-\hat{x}_{k+\frac{1}{2}}\|^2} \\
&\quad\propto p^\star(x^\star = x \,|\, x^\star + \eta_k w = \hat{x}_{k+\frac{1}{2}})
\end{aligned} \tag{3.16}$$

to promote the image to be consistent with the prior, where $w \sim \mathcal{N}(0, I_d)$. The last step, which follows from the Bayes' rule, makes it clear that this step can be precisely achieved by the denoising diffusion sampler (developed in Section 3.2) using solely the unconditional score function, with two options given in Algorithm 1 and Algorithm 2.

Combining both steps lead to the proposed DPnP method described in Algorithm 3.

3.4 Theoretical guarantee

In this section, we present asymptotic and non-asymptotic performance guarantees of DPnP.

Algorithm 4 Proximal Consistency Sampler $\mathsf{PCS}(x, y, \mathcal{L}, \eta)$ (adapted from Metropolis-Adjusted Langevin Algorithm [98])

Input: starting point $x \in \mathbb{R}^d$, measurements $y \in \mathbb{R}^m$, log-likelihood function of the forward model $\mathcal{L}(\cdot; y)$, proximal parameter $\eta > 0$.
Hyperparameter: Langevin stepsize γ, and the number of iterations N.
Initialization: $z_0 = x$.
Update: **for** $n = 0, 1, \cdots, N-1$ **do**

(1) **One step of discretized Langevin**: Set $r = \mathrm{e}^{-\gamma/\eta^2}$, and

$$z_{n+\frac{1}{2}} = rz_n + (1-r)x + \eta^2(1-r)\nabla_{z_n}\mathcal{L}(z_n; y) + \eta\sqrt{1-r^2}w_n, \quad w_n \sim \mathcal{N}(0, I_d).$$

This is equivalent to drawing $z_{n+\frac{1}{2}}$ from a distribution with density $Q(\cdot; z_n)$, where

$$Q(z'; z) \propto \exp\left(-\frac{\left\|z' - (rz + (1-r)x + \eta^2(1-r)\nabla_z\mathcal{L}(z; y))\right\|^2}{2(1-r^2)}\right).$$

(2) **Metropolis adjustment**: Compute

$$q = \frac{\exp\left(\mathcal{L}(z_{n+\frac{1}{2}}; y) - \frac{1}{2\eta^2}\|z_{n+\frac{1}{2}} - x\|^2\right)}{\exp\left(\mathcal{L}(z_n; y) - \frac{1}{2\eta^2}\|z_n - x\|^2\right)} \cdot \frac{Q(z_n; z_{n+\frac{1}{2}})}{Q(z_{n+\frac{1}{2}}; z_n)},$$

and set

$$z_{n+1} = \begin{cases} z_{n+\frac{1}{2}}, & \text{with probability } \min(1, q), \\ z_n & \text{with probability } 1 - \min(1, q). \end{cases}$$

Output: z_N.

3.4.1 Asymptotic consistency

We first collect the asymptotic correctness of our subroutines PCS and DDS in the following two lemmas. The correctness of PCS is actually well-known, see e.g., Tierney [117, Corollary 2].

Lemma 3 (Correctness of PCS)**.** *Under Assumption 1, with notation in Algorithm 4, in the continuous-time limit:*

$$\gamma \to 0, \quad N \to \infty,$$

the algorithm PCS *outputs samples with distribution* $\propto \exp(\mathcal{L}(\cdot; y) + \frac{1}{2\eta}\|\cdot - x\|^2)$.

The next lemma guarantees the correctness of DDS with exact unconditional score functions.

Lemma 4 (Correctness of DDS)**.** *Assume the score function estimation $\hat{s}_t$ is accurate, i.e. $\hat{s}_t = s_t^\star$. In the continuous-time limit:*

$$T \to \infty, \quad \bar{\alpha}_T \to 0, \quad \frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t} \to 1, \ \textit{unifomly in } t,$$

both DDS-DDIM *and* DDS-DDPM *output samples x obeying the denoising posterior distribution $p^\star(x^\star = x \,|\, x^\star + \eta\varepsilon = x_{\mathsf{noisy}})$, $\varepsilon \sim \mathcal{N}(0, I_d)$.*

We are now ready to state our main result, which concerns the asymptotic correctness of DPnP.

Theorem 1 (Asymptotic consistency of DPnP)**.** *Under the settings of Lemma 4 and Lemma 3, the following holds. Let $\varepsilon_1 > \varepsilon_2 > \cdots$ be a decreasing sequence of positive numbers satisfying $\lim_{l\to\infty} \varepsilon_l = 0$, and $0 = k_0 < k_1 < k_2 < \cdots$ be an increasing sequence of integers. Set the annealing schedule as follows:*

$$\eta_k = \varepsilon_l, \quad \textit{for } k_{l-1} \le k < k_l, \ l = 1, 2, \cdots$$

Let $\min_{l'=1,2,\cdots} |k_{l'} - k_{l'-1}| \to \infty$, the output $\hat{x}_{k_l}$ of DPnP *converges in distribution to the posterior distribution $p^\star(\cdot|y)$ for $l \to \infty$.*

In words, Theorem 1 establishes the asymptotic consistency of DPnP under fairly mild assumptions on the forward model (cf. Assumption 1): as long as the sampled distributions of DDS and PCS are exact, then running DPnP with a slowly diminishing annealing schedule of $\{\eta_k\}$ will output samples approaching the desired posterior distribution $p^\star(\cdot|y)$ when the number of iterations K goes to infinity.

3.4.2 Non-asymptotic error analysis

We now step away from the idealized setting when the sampled distributions of DDS and PCS are exact. In practice, there are many sources of errors that can influence the sampled distributions of DDS and PCS: non-diminishing γ and finite number of sampling steps N in PCS, and score estimation error and finite number of discretization steps T in DDS. In effect, these non-idealities will make the subroutines PCS and DDS *inexact*. In other words, the distribution they generate will slightly

deviate from the distribution they ought to sample from. In this chapter, we model such deviations by the *total variation* distance from the distribution generated by PCS (resp. DDS) to the ideal distribution proportional to $\exp(\mathcal{L}(x;y) - \frac{1}{2\eta_k^2}\|x - \hat{x}_k\|^2)$ (resp. $p^\star(x^\star = x | x^\star + \eta_k \varepsilon = \hat{x}_{k+\frac{1}{2}})$) uniformly over all iterations. Analyzing these total variations errors is out of the scope of this thesis, and we point the interested readers to parallel lines of works, e.g., Chewi et al. [22], Li et al. [72], Mangoubi and Vishnoi [82], among many others. In our analysis, we will assume a black-box bound for the total variation errors of PCS and DDS, which can be combined with existing analyses of the respective samplers to bound the iteration complexity of DPnP.

Theorem 2 (Non-asymptotic robustness of DPnP)**.** *With the notation in* DPnP *(Algorithm 3), set $\eta_k \equiv \eta > 0$. Under Assumption 1, there exists $\lambda := \lambda(p^\star, \mathcal{L}, \eta) \in (0,1)$, such that the following holds. Define a stationary distribution π_η by*

$$\pi_\eta(x) \propto p^\star(x) q_\eta(x),$$

where q_η is defined by

$$q_\eta(x) := \mathrm{e}^{\mathcal{L}(\cdot;y)} * p_{\eta\varepsilon}(x) = \frac{1}{(2\pi)^{d/2}\eta^d} \int \mathrm{e}^{\mathcal{L}(x';y) - \frac{1}{2\eta^2}\|x - x'\|^2} \mathrm{d}x', \quad \varepsilon \sim \mathcal{N}(0, I_d), \tag{3.17}$$

where $*$ denotes convolution. If PCS *has error at most $\varepsilon_{\mathsf{PCS}}$ in total variation and* DDS *has error at most $\varepsilon_{\mathsf{DDS}}$ in total variation per iteration, then for any accuracy goal $\varepsilon_{\mathsf{acc}} > 0$, with $K \asymp \frac{\log(1/\varepsilon_{\mathsf{acc}})}{1-\lambda}$, we have*

$$\mathsf{TV}(p_{\hat{x}_K}, \pi_\eta) \lesssim \varepsilon_{\mathsf{acc}} \sqrt{\chi^2(p_{\hat{x}_1} \,\|\, \pi_\eta)} + \frac{1}{1-\lambda}(\varepsilon_{\mathsf{DDS}} + \varepsilon_{\mathsf{PCS}}) \log\left(\frac{1}{\varepsilon_{\mathsf{acc}}}\right). \tag{3.18}$$

Before interpreting Theorem 2, we observe that $q_0(x) = \mathrm{e}^{\mathcal{L}(x;y)}$, thus $\pi_0(x) \propto p^\star(x)\mathrm{e}^{\mathcal{L}(x;y)}$ coincides with the desired posterior distribution $p^\star(\cdot|y)$. Thus Theorem 2 tells us that, assuming a constant annealing schedule $\eta_k = \eta$, the output of DPnP converges in total variation to the distribution π_η, which is a distorted version of the desired posterior distribution up to level η, with sufficiently many iterations.

A few remarks are in order.

Non-diminishing η. It can be seen from Theorem 2 that even with a nonzero η, DPnP already enforces the data prior strictly. On the other hand, the measurement

consistency is distorted by an order of η. This is usually tolerable, since the measurements are themselves contaminated by noise, thus when η is smaller than the noise level, the distortion would be tolerable. In practice, it is beneficial to choose an annealing schedule of $\{\eta_k\}$, which will be elaborated in Section 3.5.

Spectral gap and worst-case convergence rate. The term $1-\lambda$ is known as the *spectral gap* of the associated Markov chain of DPnP. In many situations, it can be shown that $1-\lambda \gtrsim \frac{\eta}{\mathrm{poly}(d)}$, see e.g. Vono et al. [124] for the case of log-concave $p^\star$ and negative quadratic $\mathcal{L}$. In such cases, the factor $\frac{1}{1-\lambda}$ in the right hand side of (3.18) can also be improved significantly [4]. However, under the minimal assumption in this chapter and without any additional assumption on $p^\star$ and $\mathcal{L}$, $1-\lambda$ can be exponentially small in the worst case, thus our result does not contradict the worst-case lower bound in Gupta et al. [44].

Provable robustness. Theorem 2 indicates the performance of DPnP degenerates gracefully in the presence of sampling errors. To the best of our knowledge, this is the first provably consistent and robust posterior sampling method for nonlinear inverse problems using score-based diffusion priors.

3.5 Numerical experiments

We provide numerical evidence to corroborate the promise of DPnP in solving both linear and nonlinear image reconstruction tasks. We denote DPnP with the subroutines DDS-DDPM and DDS-DDIM as DPnP-DDPM and DPnP-DDIM respectively.

3.5.1 Synthetic data

To demonstrate the correctness of our algorithm, we run DPS [23] and DPnP on a simple linear inverse problem under two-dimensional Gaussian mixture prior. Here, the unconditional distribution p_0 is a 2-dimensional random vector generated by a GMM:

$$p_0 = 0.6 \cdot \mathcal{N}\big([-3,-1]^\top, 0.75I\big) + 0.4 \cdot \mathcal{N}\big([1,1]^\top, 0.75I\big).$$

Our observation model is a simple rank-one linear measurement:

$$y = \langle a, x\rangle, \quad a = [1,-1]^\top.$$

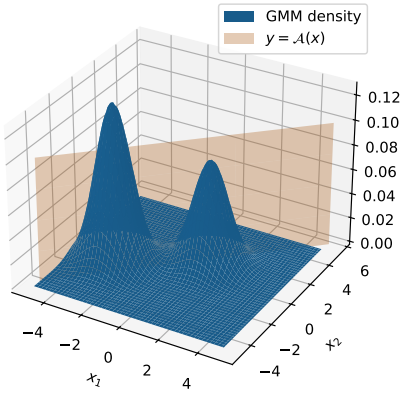

Figure 3.1: Illustration of the simple inverse problem for a Gaussian mixture model.

Assume we observed $y = -0.5$ and wish to sample from the posterior distribution $p(x|y)$, depicted in Figure 3.1. We run Monte Carlo simulations to evaluate the performance of DPS and DPnP: each algorithm is run for $N_{\mathsf{MC}} = 10000$ times, obtaining N_{MC} outputs per algorithm. We then estimate the probability density of the outputs of each algorithm via a kernel estimator. We compare these densities with the true conditional probability density $p(x|y)$, which can be computed analytically for GMM. The result is depicted in Figure 3.2. It can be seen that DPnP is able to fully approach the true posterior distribution, while DPS fails in capturing the weights of differents modes.

3.5.2 Inverse problems

We consider the following linear and nonlinear inverse problems in our experiments.

Phase retrieval. We consider phase retrieval with a coded mask, which is a classical inverse problem [16]. For a 256×256 image x (for each color channel) in our experiments, we first generate a random mask $M \in \mathbb{R}^{256\times256}$ (which is shared across color channels), then apply Fourier transform $\mathcal{F}$ to $M \odot x$, where $\odot$ denotes the Hadamard (entrywise) product, and finally preserve only the magnitudes of the Fourier transform. Formally, the forward measurement operator is $\mathcal{A}(x) = \mathrm{mag}(\mathcal{F}(M \odot x))$, where $\mathrm{mag}(\cdot)$ computes the entrywise magnitude of a matrix with complex entries. The measurement noise is again set to be white Gaussian, with

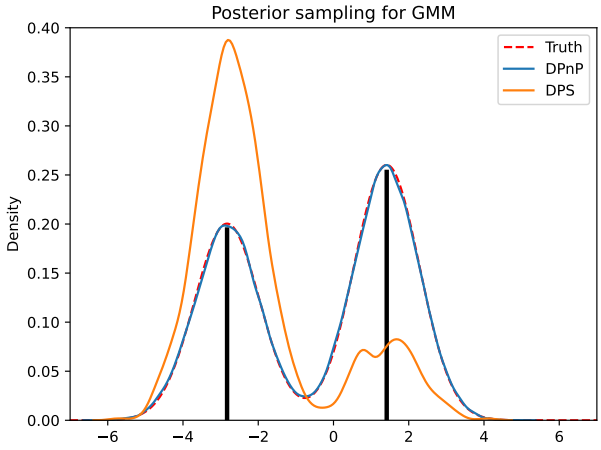

Figure 3.2: Output distribution of different posterior sampling algorithms. DPnP is able to recover the true posterior distribution.

variance 0.2.

Quantized sensing. Quantized sensing refers to the task of reconstructing an image from its low-bit quantized version. Here, the forward measurement operator is a one-bit per channel, dithered quantization operator. More precisely, it applies entrywise the following stochastic function Q with dithering level $\theta > 0$:

$$Q(\mathsf{pixel}) = \begin{cases} 1, & \text{with probability } \frac{e^{\mathsf{pixel}/\theta}}{1+e^{\mathsf{pixel}/\theta}} \\ -1, & \text{with probability } \frac{1}{1+e^{\mathsf{pixel}/\theta}}, \end{cases}$$

where $\mathsf{pixel} \in [-1, 1]$ is the value of each pixel in each channel. The measurements in quantized sensing are therefore one-bit-per-channel images. The dithering level θ is set to 0.4 in our experiments.

Super resolution. The forward model for super-resolution is the bicubic downsampling operator [63], which is a linear operator (in fact, a block Hankel matrix). We use a downsampling ratio of 4 in all our experiments. The measurement noise is set to be white Gaussian, with variance 0.2.

3.5.3 Experimental setups

We compare DPnP with the state-of-the-art DPS algorithm [23] and LGD-MC algorithm [111] on the FFHQ validation dataset [60] and the ImageNet validation dataset [101]. We use the same pre-trained score functions as in [23],[4] and all images are normalized to fit into the range $[-1, 1]$.

Annealing schedule. For DPnP, we use a heuristic strategy to choose the annealing schedule η_k in DPnP (Algorithm 3). As seen in the theoretical analysis (Theorem 2), if we set all the $\eta_k \equiv \eta$ for some constant $\eta > 0$, then DPnP converges to a distribution π_η, which can be regarded as a version of the posterior distribution $p^\star(\cdot|y)$ distorted by an order of $O(\eta)$. The smaller η is, the more accurate the final distribution will be. On the other hand, it was also seen that in many cases, the spectral gap is $\Omega(\eta)$, hence the convergence time is $O(\frac{1}{\eta})$. Therefore, smaller η would make it take longer to converge.[5]

To strike a balance between the accuracy and the convergence rate, we adapt an gradually decreasing schedule for η_k, similar to Bouman and Buzzard [11]. In the first few iterations, we set η_k to be a large constant. After this initial phase, we decrease η_k slowly, eventually to η_N which is chosen to be a small constant. An example of such an annealing schedule is

$$\begin{aligned} &\eta_0 = \eta_1 = \cdots = \eta_{K_0}, \quad \eta_0 > 0 \text{ is a large constant,} \\ &\eta_k = (\eta_K/\eta_0)^{\frac{k-K_0}{K-K_0}}\, \eta_0, \quad K_0 < k \le K, \quad \eta_K > 0 \text{ a small constant,} \end{aligned}$$

where $K_0 < K$ is the length of the initial phase, which can be chosen as, e.g., $K_0 = K/5$. For all the numerical experiments, we set $\eta_0 = 0.4$, $\eta_N = 0.15$, $K_0 = 4$, $K = 20$. The annealing schedule $\{\eta_k\}$ of DPnP is fixed across *all* tasks, while DPS and LGD-MC are fine-tuned with reasonable effort for best performance. All experiments are run on a single Nvidia L40 GPU.

Initialization. In Algorithm 3, the initial guess $\hat{x}_0$ is set to be a properly scaled Gaussian random vector. Notwithstanding, from Theorem 2 it can be infered that

[4] https://github.com/DPS2022/diffusion-posterior-sampling

[5] Strictly speaking, while the number of iterations required to converge increases as η gets smaller, the computational complexity of PCS and DDS per iteration will decrease. However, in experiments, the decrease is not strong enough to offset the increase in the total number of iterations, thus the overall computational complexity still increases as η becomes smaller.

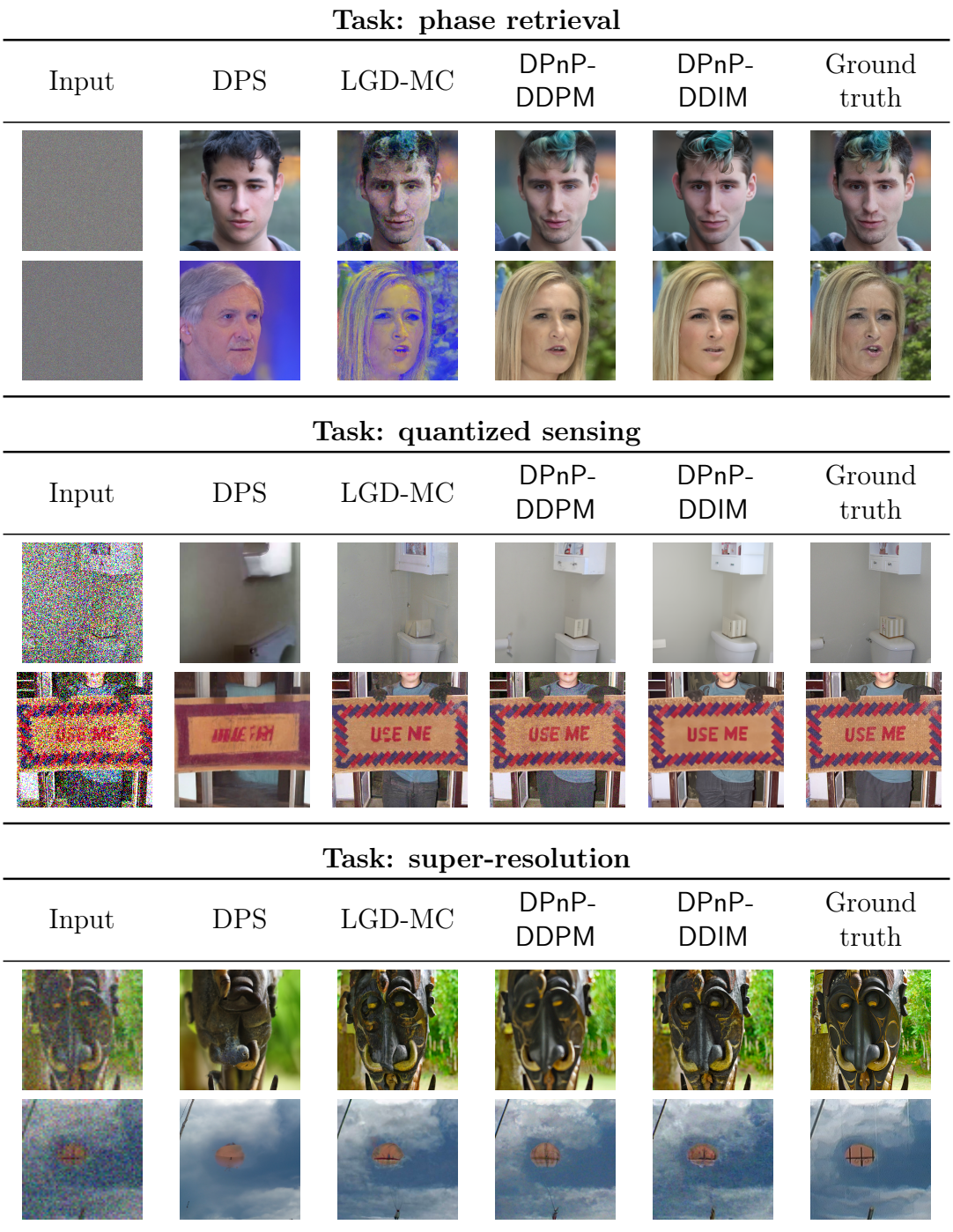

Figure 3.3: Samples of different algorithms for phase retrieval, quantized sensing, and super resolution, where DPnP generate images of higher quality and recover fine details of the image more faithfully than the state-of-the-art DPS [23] and LGD-MC [111] algorithms.

using a heuristic posterior sampler as the initializer could decrease $\chi^2(p_{\hat{x}_1} \,\|\, \pi_\eta)$, hence potentially improve the convergence speed of DPnP. By using existing algorithms like DPS or LGD-MC as initializers, DPnP can improve upon the results of existing algorithms towards the correct posterior distribution efficiently and provably. In our experiments, we find it helpful to initialize DPnP with LGD-MC, which accelerates the algorithm significantly.

3.5.4 Results

Visual results. The samples generated by different algorithms are shown in Figure 3.3. It can be seen from these results that, DPnP is capable of solving both linear and nonlinear problems, and, in comparison with state-of-the-art algorithms, performs better in recovering fine and crisper details.

Performance metric. We report the performance metric of DPnP in terms of LPIPS and PSNR — which are two of the most relevant metrics for inverse problems — on the FFHQ and ImageNet datasets in 3.1 and 3.2, respectively. Since DPnP-DDIM has similar performance with DPnP-DDPM but admits much faster implementation, only DPnP-DDIM is evaluated. It can be seen that DPnP-DDIM performs strongly on both datasets, albeit taking about 1.5x more computation time.

Table 3.1: Evaluation of solving inverse problems on FFHQ 256×256 validation dataset (1k samples).

	Super-resolution (4x, linear)		Phase retrieval (nonlinear)		Quantized sensing (nonlinear)		Time per sample
Algorithm	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	
DPnP-DDIM (ours)	**0.301**	**24.2**	**0.376**	**22.4**	**0.293**	**24.2**	∼ 90s
DPS [23]	0.331	23.1	0.490	17.4	0.367	21.7	∼ 60s
LGD-MC ($n = 5$) [111]	0.318	23.9	0.522	16.4	0.317	23.9	∼ 60s

3.6 Related works

Given its interdisciplinary nature, our work sits at the intersection of generative modeling, computational imaging, optimization and sampling. Here, we discuss some works that are most related to ours.

Table 3.2: Evaluation of solving inverse problems on ImageNet 256×256 validation dataset (1k samples).

	Super-resolution (4x, linear)		Phase retrieval (nonlinear)		Quantized sensing (nonlinear)		Time per sample
Algorithm	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	
DPnP-DDIM (ours)	**0.416**	**21.6**	**0.562**	**13.4**	**0.363**	**23.0**	$\sim$ 240s
DPS [23]	0.473	20.2	0.677	**13.4**	0.542	18.7	$\sim$ 150s
LGD-MC ($n = 5$) [111]	**0.416**	20.9	0.592	12.8	0.384	22.3	$\sim$ 150s

Algorithmic unrolling and plug-and-play image reconstruction. Composite optimization algorithms, which aim to minimize the sum of a measurement fidelity term and a regularization term promoting desirable solution structures, have been the backbone of inverse problem solvers. To unleash the power of deep learning, Gregor and LeCun [41] advocates the perspective of algorithmic unrolling, which turns an iterative algorithm into concatenations of linear and nonlinear layers like in a neural network. Venkatakrishnan et al. [120] recognized that the proximal mapping step in many composite optimization algorithms can be regarded as a denoiser or denoising operator with respect to the given prior, and proposed to "plug in" alternative denoisers, in particular state-of-the-art deep learning denoisers, leading to a class of popular algorithms known as plug-and-play methods [15]; see Monga et al. [84] for a review.

Regularization by denoising and score matching. Vincent [122] pointed out a connection between score matching and image denoising, which is a consequence of the Tweedie's formula [31]. The regularization by denoising (RED) framework [99] follows the plug-and-play framework to minimize a regularized objective function, where the regularizer is defined based on the plug-in image denoiser; Reehorst and Schniter [96] later clarified that the RED framework can be interpreted as score matching by denoising using the Tweedie's formula. Kawar et al. [61] developed a stochastic image denoiser for posterior sampling of image denoising using annealed Langevin dynamics. Fang et al. [33] provided a framework to learn exact proximal operators for inverse problems.

Plug-and-play posterior sampling. Motivated by the need to characterize the uncertainty, tackling image reconstruction as posterior sampling from a Bayesian perspective is another important approach. Our method is inspired by the plug-and-play framework but takes on a sampling perspective, exploiting the connection

between optimization and sampling [125]. Along similar lines, Bouman and Buzzard [11], Laumont et al. [65] proposed Bayesian counterparts of plug-and-play for posterior sampling, where they leveraged the connection to score matching for sampling from the image prior, but did not consider score-based diffusion models for the image prior, which is a key aspect of ours; see also Sun et al. [116]. Coeurdoux et al. [25] extended the split Gibbs sampler [123] in the plug-and-play framework, and advocated the use of score-based diffusion models such as DDPM [49] for image denoising based on heuristic observations. In contrast, we rigorously derive the denoising diffusion samplers from first principles, unraveling critical gaps from naïve applications of the generative samplers to denoising, and offer theoretical guarantees on the correctness of our approach.

Score-based diffusion models as image priors. Several representative methods for solving inverse problems using score-based diffusion priors alternates between taking steps along the diffusion process and projecting onto the measurement constraint, e.g., Chung et al. [23, 24], Graikos et al. [39], Kawar et al. [62], Song et al. [108, 110]. However, these approaches do not possess asymptotic consistency guarantees. Song et al. [111] proposed to use multiple Monte Carlo samples to reduce bias. On the other hand, Cardoso et al. [17] developed Monte Carlo guided diffusion methods for Bayesian linear inverse problems which tend to be computationally expensive, and Dou and Song [30] recently introduced a filtering perspective and applied particle filtering. Although asymptotically consistent, these approaches are limited to linear inverse problems. Trippe et al. [118], Wu et al. [126] introduced sequential Monte Carlo (SMC) algorithms for conditional sampling using unconditional diffusion models that are asymptotically exact. Mardani et al. [83] developed a variational perspective that connects to the regularization by denoising framework. Gupta et al. [44] showed that the worst-case complexity of diffusion posterior sampling can take super-polynomial time regardless of the algorithm in use.

3.7 Discussion

This work sets forth a rigorous and versatile algorithmic framework called DPnP for solving nonlinear inverse problems via posterior sampling, using image priors prescribed by score-based diffusion models with general forward models. DPnP alternates

between two sampling steps implemented by DDS and PCS, to promote consistency with the data prior and the measurement likelihood respectively. We provide both asymptotic and non-asymptotic convergence guarantees, establishing DPnP as the first provably consistent and robust score-based diffusion posterior sampling method for general nonlinear inverse problems. Our work opens up many interesting questions, which we single out a few below.

- *Accelerated posterior sampling.* Due to the modular design, it is straightforward to incorporated existing accelerated samplers for both DDS [80] and PCS [81] to speed up the inference, which is of broad practical interest.
- *Non-differentiable forward models.* While we assume the log-likelihood function $\mathcal{L}(\cdot; y)$ to be differentiable to apply MALA for PCS, it is straightforward to adopt other samplers that only assume zero-order oracle access to $\mathcal{L}(\cdot; y)$ for non-differentiable forward models.
- *Guided generation.* While we focus on solving inverse problems, our design might provide some insights into improving the quality of controlled or guided generation [111] as well.

Chapter 4

Explicit Rates for Diffusion-Based Solvers under Sparse Priors

The DPnP framework developed in the previous chapter establishes a general non-asymptotic theory for diffusion-based solvers. At the same time, it leaves several important questions unresolved.

- The convergence bound in Theorem 2 depends on the quantity $1 - \lambda$, known as the spectral gap of a Markov chain tied to the DPnP dynamics. Unfortunately, this spectral gap may be small in theory. In the worst case, it can decay exponentially with the problem dimension [44]. As a result, the abstract guarantee provided by the general theory may not be sufficient to explain the practical performance of the method.
- More broadly, the existing theory does not account for the empirical observation that many heuristic diffusion-based solvers, such as DPS, often produce meaningful reconstructions on structured inverse problems encountered in practice, even in regimes where their stationary distributions are provably different than the true posterior distribution.

The goal of this chapter is to take a first step toward resolving these issues in a concrete and mathematically tractable setting. Rather than pursuing complete generality, we focus on denoising problems equipped with sparse priors. This setting serves as a useful testing ground: it is rich enough to capture important phenomena arising in applications, while remaining amenable for analysis.

Our main message is that, under sparse priors, one can go substantially beyond the

abstract worst-case theory. In particular, we show that a simple heuristic diffusion-based solver can effectively leverage sparsity and achieve explicit quantitative rates for basic inverse problems. These results paves the way to further understanding of why diffusion-based methods can succeed in practice even when general-purpose spectral-gap bounds appear prohibitively pessimistic.

4.1 Background

We begin by introducing the sparse prior that will be used throughout this chapter. Our canonical model is the *Gaussian–Bernoulli prior*, which provides a simple and analytically convenient description of sparsity. Informally, each coordinate is zero with high probability and is otherwise drawn from a Gaussian distribution. This prior captures the basic principle that the unknown signal is supported on only a small fraction of coordinates, while the nonzero entries themselves have a continuous distribution.

Definition 1 (Gaussian–Bernoulli prior)**.** Let $s, n \in \mathbb{N}$, with $s < n$. A random vector

$$X = (X_1, \dots, X_n) \in \mathbb{R}^n$$

is said to follow the *Gaussian–Bernoulli model* with sparsity s and dimension n, denoted as $\mathsf{GB}(s, n)$, if its coordinates are independent and identically distributed according to

$$X_i \sim (1 - \frac{s}{n})\, \delta_0 + \frac{s}{n} \mathcal{N}(0, 1/s), \qquad i = 1, \dots, n.$$

Equivalently, one may write

$$X_i = B_i Z_i,$$

where

$$B_i \sim \text{Bernoulli}(\frac{s}{n}), \qquad Z_i \sim \mathcal{N}(0, 1/s),$$

and all random variables $\{B_i, Z_i\}_{i=1}^n$ are mutually independent.

In particular, each coordinate satisfies

$$\mathbb{P}(X_i = 0) = 1 - \frac{s}{n},$$

and conditional on $X_i \neq 0$, the value of X_i is distributed as $\mathcal{N}(0, 1/s)$. Thus X is

sparse when $s \ll n$, with expected support size

$$\mathbb{E}\big[|\operatorname{supp}(X)|\big] = s.$$

A heuristic diffusion-based solver. Existing diffusion-based posterior sampling algorithms typically involves adding a guidance term in the reverse diffusion equation. In the present chapter, we work with a particularly simple instance of this principle, which dates back at least to Aali et al. [1]. The algorithm generates iterates by

$$x_{k+1} = x_k + \eta_k\Big(\nabla \log p_{t_k}(x_k) + \lambda_k \nabla \mathcal{L}(x_k; y)\Big) + \sqrt{2\eta_k}\,\xi_k, \qquad \xi_k \sim \mathcal{N}(0, I), \quad (4.1)$$

where $\eta_k > 0$ is the step size, $t_k > 0$ is the diffusion level, $\lambda_k > 0$ is the guidance strength, and $\mathcal{L}(x; y)$ is the measurement log-likelihood introduced earlier.

To keep the discussion focused, we specialize throughout this chapter to the most basic inverse problem, namely denoising. Thus the observation model is

$$y = x_\star + \sigma\zeta, \qquad x_\star \sim \mathsf{GB}(s, n), \quad \zeta \sim \mathcal{N}(0, I). \quad (4.2)$$

Accordingly, the measurement log-likelihood takes the form

$$\mathcal{L}(x; y) = -\frac{\|x - y\|^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2), \qquad \nabla\mathcal{L}(x; y) = -\frac{x - y}{\sigma^2}.$$

Inserting this expression into (4.1), and using an approximate score $\hat{s}_t(x)$ in place of the exact score $\nabla \log p_t(x)$, we obtain

$$x_{k+1} = x_k + \eta_k\Big(\hat{s}_{t_k}(x_k) - \frac{\lambda_k}{\sigma^2}(x_k - y)\Big) + \sqrt{2\eta_k}\,\xi_k, \qquad \xi_k \sim \mathcal{N}(0, I), \quad (4.3)$$

We see that the guidance term simply pulls the iterate toward the observation y, with a strength modulated by λ_k.

We now specify the choice of algorithmic parameters that will be analyzed in the sequel. Fix an iteration budget $K \in \mathbb{N}$, a terminal diffusion level

$$T = C_T\sqrt{n} \quad (4.4)$$

for a sufficiently large universal constant $C_T > 0$, and an early stopping threshold

$\tau > 0$. The quantity T will be the largest value attained by the schedule (t_k), while τ will be the smallest. We also introduce a parameter $\Lambda > 1$, which we refer to as the *time-dilation parameter*; its role will be clarified below.

The parameters are chosen as follows.

- *Exponential diffusion schedule.* We use a geometrically decreasing diffusion schedule from T down to τ, namely

$$t_k = T\left(\frac{\tau}{T}\right)^{k/K}, \qquad k = 0, 1, \dots, K. \tag{4.5}$$

  In particular, $t_K = \tau$. The associated step sizes are defined by

$$\eta_k = \Lambda(t_k - t_{k+1}), \qquad k = 0, 1, \dots, K-1. \tag{4.6}$$

  Thus the physical time scale determined by the discretization is dilated by the factor Λ relative to the decay of the diffusion level.

- *Increasing guidance strength.* The guidance coefficient is chosen to be

$$\lambda_k = \frac{\sigma^2}{2t_k + 16\sigma^2 \log n}, \qquad k = 0, 1, \dots, K. \tag{4.7}$$

  In particular, λ_k is small when t_k is large and increases gradually as the diffusion level decreases.

Let us briefly comment on the motivation behind these choices. The exponential schedule (4.5) is standard in implementations related to the variance-exploding SDE, whose reverse-time discretization resembles (4.3) without the guidance term. The extra factor Λ in (4.6) is borrowed from the time-dilation idea of Guo et al. [43]: it distinguishes the rate at which the diffusion level t_k decreases from the amount of algorithmic time spent at that scale, thereby allowing the chain more opportunity to mix before the noise level is further reduced. At the level of intuition, one may think of t_k as controlling the landscape being explored, while η_k controls how long the algorithm is allowed to explore that landscape. We suspect that this parameter is mainly an artifact of the proof, and is removable with more refined analysis.

Finally, the schedule (4.7) is designed to reflect the empirical behavior of diffusion posterior sampling heuristics such as DPS. At early stages, when the diffusion level is high and the score term dominates, the guidance is deliberately weak; this

prevents the dynamics from being overly constrained by the measurement before sufficient exploration has taken place. As the diffusion level decreases, the guidance becomes stronger, gradually steering the iterates toward configurations that are more consistent with the observation. In this sense, the algorithm transitions from a predominantly prior-driven exploration phase to a more data-driven refinement phase.

Assumption on score estimation error. We assume that the score estimator satisfies an L^2 error bound along a family of auxiliary distributions designed for the denoising problem. More precisely, define

$$\Pi_t^y(x) \propto p_t(x) \exp\left(-\frac{\|x-y\|^2}{2(2t+16\sigma^2 \log n)}\right).$$

We assume that, for a sufficiently small universal constant $c > 0$,

$$\sum_{k=1}^{K-1} \eta_k \mathbb{E}_{X_{t_k} \sim \Pi_{t_k}^y} \|\nabla \log p_{t_k}(X_{t_k}) - s_{t_k}(X_{t_k})\|^2 \le c. \tag{4.8}$$

Assumptions of this type are standard in the diffusion sampling literature, where the expectation is typically taken with respect to the marginal p_{t_k}; see, for example, [7, 72]. In the present denoising inverse problem setting, however, the reference distribution must reflect the additional information provided by the observation y. The measure Π_t^y serves this purpose. Because our guidance strength decays according to (4.7), whose continuous-time analogue appears as the Gaussian tilt in the definition of Π_t^y, the distribution Π_t^y behaves similarly to p_t at large noise levels, while at small noise levels it concentrates near the observation y, and hence near the ground truth $x_\star$. This captures the intuition that, as the inverse problem solver approaches the true signal, score accuracy in a neighborhood of $x_\star$ becomes increasingly important.

4.2 Main results

We now state the main guarantee for the heuristic solver introduced above in the sparse denoising setting.

Theorem 3. *Fix an early stopping time $0 < \tau \le \sigma^2$. Assume the score estimation error bound (4.8), and that*

$$\sigma^2 \le \frac{c_\sigma}{s^3 \log^2 n} \tag{4.9}$$

for some sufficiently small universal constant $c_\sigma > 0$.

Consider the denoising problem (4.2). Run the heuristic solver (4.3) with the parameter schedules specified above by (4.4)–(4.7), with initialization

$$x_0 \sim \mathcal{N}\Big(0, \frac{2t_0}{3} I\Big).$$

If

$$\Lambda \ge C_\Lambda \Big(n \log \frac{T}{\tau} + \frac{n^2}{\sigma s^{3/2}} \log^2\big(\frac{n}{\sigma}\big) \Big) \tag{4.10}$$

and

$$K \ge C_K \left(\frac{(1 + n\sigma^2 \log n) \log^2(n/\tau)}{\tau^2} + \frac{\Lambda^2 n \log^5(n/\tau)}{\tau} \right) \tag{4.11}$$

with sufficiently large universal constants $C_\Lambda, C_K > 0$, then the output x_K satisfies, with probability at least 0.99, that for every coordinate $i \in [n]$,

$$|x_K(i) - x_\star(i)| \lesssim \begin{cases} \sigma \log n, & i \in \mathrm{supp}(x_\star), \\ \sqrt{\tau \log n}, & i \notin \mathrm{supp}(x_\star). \end{cases}$$

Moreover, with probability at least 0.99, we have perfect support recovery:

$$\Big\{ i : |x_K(i)| \ge C\sqrt{\tau \log n} \Big\} = \mathrm{supp}(x_\star)$$

for some universal constant $C > 0$.

When choosing $\Lambda = \widetilde{\Theta}(\frac{n^2}{\sigma s^{3/2}})$, the above theorem gives an iteration complexity of

$$K = \widetilde{\Theta}\left(\frac{n^5}{s^3 \sigma^2 \tau} + \frac{1 + n\sigma^2}{\tau^2} \right).$$

The theorem shows that, in the sparse denoising model, the diffusion-based heuristic recovers the signal with coordinatewise accuracy that cleanly distinguishes between the active and inactive coordinates. On the true support, the reconstruction error is bounded by the noise level up to a logarithmic factor, while away from the support

the algorithm drives the coordinates down to the much smaller scale $\sqrt{\tau}$ determined by the early-stopping time.

It is worth emphasizing that the theorem concerns a highly nonconvex, score-driven dynamics rather than a classical thresholding or convex optimization method. The result therefore gives theoretical evidence that the combination of score-based diffusion models and likelihood guidance can indeed exploit the sparse prior to solve inverse problems in a quantitatively meaningful way. This provides one explanation for the empirical success of heuristic diffusion-based solvers in inverse problems.

A few remarks are in order.

Remark 1 (Interpretation of the coordinatewise bounds). The conclusion separates the two tasks that any sparse recovery procedure must perform: identifying the support and estimating the nonzero amplitudes. For indices $i \notin \mathrm{supp}(x_\star)$, the theorem guarantees that the output remains extremely close to zero, with residual magnitude controlled by $\sqrt{\tau}$. Since τ may be chosen polynomially small in the problem parameters, this provides a strong form of approximate support recovery. For indices on the support, the error bound is of order $\sigma \log n$, which is consistent (up to logarithmic factors) with the scale of the fluctuation of the ambient Gaussian noise.

Remark 2 (Comparison with prior art). Our analysis requires only a Gaussian initialization, in line with the standard practice of diffusion models. This contrasts with existing analyses that rely on a carefully designed initialization [129] or require an unspecified burn-in period before entering the regime where explicit convergence can be established [68].

Remark 3 (Comparison with Langevin dynamics). One can also obtain a convergence guarantee for Langevin dynamics targeting Π_τ^y with sufficiently small $\tau > 0$, for instance through a Holley–Stroock perturbation argument. However, such a guarantee relies on access to the exact score function. This requirement is difficult to meet in practice, especially when τ is small and the target distribution becomes highly singular.

Remark 4 (Polynomial complexity). A key feature of Theorem 3 is that both the iteration count K and the time-dilation parameter Λ are bounded by quantities that are polynomial in n, τ^{-1}, and σ^{-1}. Thus, at least in this model problem, the algorithm avoids the exponentially slow behavior that might be suggested by the

worst-case spectral-gap perspective discussed earlier.

Remark 5 (Role of the stopping time τ). The parameter τ should be viewed as the target resolution at which the algorithm is terminated. The theorem does not claim exact recovery of the zero coordinates; instead, it shows that these coordinates are shrunk to scale $\sqrt{\tau}$. Decreasing τ therefore improves the off-support accuracy, but it also affects the required computational budget through the polynomial dependence on τ^{-1}. In this sense, τ governs the tradeoff between computational effort and the final level of sparsity one wishes to enforce.

4.3 Proof outline

Our proof hinges upon the concept of Wasserstein action [5], which we introduce below.

Definition 2 (Wasserstein action)**.** Let $(\rho_t)_{t\in[a,b]}$ be a family of probability distributions in $\mathbb{R}^d$ with finite second moments, which is further assumed to be an absolutely continuous curve in the W_2 metric, and let

$$|\dot{\rho}|_t \coloneqq \lim_{\delta\to 0} \frac{W_2(\rho_{t+\delta}, \rho_t)}{|\delta|}$$

denote its metric derivative whenever the limit exists. The Wasserstein action of the curve is

$$\mathcal{A}\big((\rho_t)_{t\in[a,b]}\big) \coloneqq \int_a^b |\dot{\rho}|_t^2 \, dt.$$

Equivalently, if v_t is a velocity field solving

$$\partial_t \rho_t + \nabla\cdot(\rho_t v_t) = 0,$$

then

$$\mathcal{A}\big((\rho_t)_{t\in[a,b]}\big) = \inf_{\partial_t\rho_t+\nabla\cdot(\rho_t v_t)=0} \int_a^b \|v_t\|^2_{L^2(\rho_t)} \, dt.$$

We divide the proof of Theorem 3 into a series of lemmas. We define an auxiliary family of measures by

$$\Pi_t^y(\mathrm{d}x) \propto p_t(x) \exp\Big(-\frac{\|x-y\|^2}{2(2t+16\sigma^2\log n)}\Big)\,\mathrm{d}x.$$

We will compare the distribution of x_k against the law of Π_t^y at appropriate time points.

We begin with a technical lemma that bounds the Wasserstein action of the curve of the auxiliary measures.

Lemma 5 (Action bound)**.** *Let*

$$\mathcal{A}_y \coloneqq \mathcal{A}\big((\Pi_t^y)_{t\in[\tau,T]}\big).$$

Then

$$\mathcal{A}_y \lesssim n\log\frac{T}{\tau} + s\|y\|^2 + \frac{n^2}{\sigma_0 s^{3/2}}\log^2\Big(\frac{n}{s\sigma_0}\Big)$$

where we define $\sigma_0 \coloneqq 4\sigma\sqrt{\log n}$.

The next lemma compares the law of the algorithm output x_K with the auxiliary measure.

Lemma 6 (Comparing with the auxiliary measure)**.** *There exists a universal constant* $C>0$ *such that*

$$\begin{aligned}\mathsf{KL}\big(\Pi_\tau^y\|\mathsf{Law}(x_K)\big) \le \mathsf{KL}(\Pi_T^y\|\mathsf{Law}(x_0)) + C\left(\frac{1}{\Lambda}+\frac{1}{K\tau}\log\frac{T}{\tau}\cdot\log^2\frac{n}{\tau}\right)\mathcal{A}_y \\ + C\left(\frac{\|y\|_2^2}{K\tau^2}+\frac{n\sigma^2}{K\tau^2}\log n+\frac{\Lambda^2 n}{K\tau}\log^3\frac{n}{\tau}\right)\log^2\frac{T}{\tau}.\end{aligned} \tag{4.12}$$

In the above error bound, we further need to control the initialization error $\mathsf{KL}(\Pi_T^y\|\mathsf{Law}(x_0))$, which is the goal of our next lemma.

Lemma 7 (Initialization error)**.** *Under the assumption of Theorem 3,*

$$\mathsf{KL}(\Pi_T^y\|\mathsf{Law}(x_0)) \lesssim \frac{n}{T^2}+\frac{\|y\|_2^2}{T}.$$

It remains to study the auxiliary measure, as done by the following lemma.

Lemma 8 (Localization of the auxiliary measure)**.** *There exist universal constants* $c, C>0$ *such that the following holds. There exists an event* $\mathcal{G}$ *with* $\mathbb{P}(\mathcal{G})\ge 0.999$*, and on* $\mathcal{G}$*, if* $X\sim\Pi_\tau^y$*, then with probability at least* 0.999*,*

$$|X(i)-x_\star(i)| \lesssim \begin{cases}\sigma\log n, & i\in\mathrm{supp}(x_\star),\\ \sqrt{\tau\log n}, & i\notin\mathrm{supp}(x_\star),\end{cases} \qquad \forall i\in[n].$$

Putting these together, we are now ready to prove Theorem 3.

Proof of Theorem 3. For simplicity, let

$$L \coloneqq \log \frac{n}{\tau}.$$

Let

$$E \coloneqq \left\{ x \in \mathbb{R}^n : \forall i \in [n],\ |x(i) - x_\star(i)| \lesssim \begin{cases} \sigma \log n, & i \in \operatorname{supp}(x_\star), \\ \sqrt{\tau \log n}, & i \notin \operatorname{supp}(x_\star) \end{cases} \right\}.$$

By Lemma 8, there exists an event $\mathcal{G}$ such that

$$\mathbb{P}(\mathcal{G}) \geq 0.999,$$

and on $\mathcal{G}$, if $X \sim \Pi_\tau^y$, then

$$\Pi_\tau^y(E) \geq 0.999.$$

It therefore suffices to show that $\mathsf{Law}(x_K)$ is close to Π_τ^y in total variation.

We shall also condition on the event

$$\|y\| \lesssim 1,$$

which holds with high probability, say, 0.9999, by Chebyshev's inequality.

Step I: action bound. By Lemma 5,

$$\mathcal{A}_y \lesssim n \log \frac{T}{\tau} + s\|y\|^2 + \frac{n^2}{\sigma_0 s^{3/2}} \log^2\Big(\frac{n}{s\sigma_0}\Big)$$

Using $\|y\|_2 \lesssim 1$ and $\sigma_0^2 = 16\sigma^2 \log n$, we get

$$\mathcal{A}_y \lesssim n \log \frac{T}{\tau} + s + \frac{n^2}{\sigma s^{3/2}} \log^2\Big(\frac{n}{s\sigma}\Big)$$

Since (4.9) implies

$$\sigma \leq \frac{c}{s^{3/2} \log n},$$

we have

$$\log \frac{n}{s\sigma} \lesssim \log \frac{n}{\sigma}.$$

Recall that $T = C_T\sqrt{n}$. Then

$$\log \frac{T}{\tau} \asymp \log \frac{n}{\tau} = L, \qquad \log T \asymp \log n.$$

Therefore

$$\mathcal{A}_y \lesssim nL + \frac{n^2}{\sigma s^{3/2}} \log^2\Big(\frac{n}{\sigma}\Big).$$

Set

$$A := nL + \frac{n^2}{\sigma s^{3/2}} \log^2\Big(\frac{n}{\sigma}\Big).$$

We have shown that

$$\mathcal{A}_y \le CA. \tag{4.13}$$

Step II: initialization error. By Lemma 7,

$$\mathsf{KL}(\Pi_T^y \| \mathsf{Law}(x_0)) \lesssim \frac{n}{T^2} + \frac{\|y\|_2^2}{T} \lesssim \left(\frac{n}{T^2} + \frac{1}{T}\right).$$

Since $T = C_T\sqrt{n}$, choosing C_T sufficiently large yields

$$\mathsf{KL}(\Pi_T^y \| \mathsf{Law}(x_0)) \le 10^{-5}. \tag{4.14}$$

Step III: KL control at time τ. Applying Lemma 6, and using (4.13), (4.14), $\|y\|_2 \lesssim 1$, and

$$\log(T/\tau) \asymp L,$$

we obtain

$$\mathsf{KL}(\Pi_\tau^y \| \mathsf{Law}(x_K)) \le 10^{-5} + C\left(\frac{1}{\Lambda} + \frac{L^3}{K\tau}\right) A + C\left(\frac{L^2}{K\tau^2} + \frac{n\sigma^2 \log n\, L^2}{K\tau^2} + \frac{\Lambda^2 n L^5}{K\tau}\right). \tag{4.15}$$

The condition (4.10) implies, after enlarging the constant there if necessary, that

$$C\frac{A}{\Lambda} \le 10^{-5}.$$

Next, the term $AL^3/(K\tau)$ is dominated by the last term on the right-hand side of (4.15). Indeed, since $\Lambda \geq C_\Lambda A$,

$$\frac{\Lambda^2 n L^5}{K\tau} \geq C_\Lambda^2 \frac{A^2 n L^5}{K\tau} = C_\Lambda^2 \, A n L^2 \cdot \frac{AL^3}{K\tau}.$$

It therefore suffices to note that

$$A\, nL^2 \gtrsim 1,$$

where in the inequality we used (4.9). Hence

$$\frac{AL^3}{K\tau} \leq C \frac{\Lambda^2 n L^5}{K\tau}.$$

Thus (4.15) simplifies to

$$\mathsf{KL}(\Pi_\tau^y \| \mathsf{Law}(x_K)) \leq 10^{-5} + C\frac{A}{\Lambda} + C\left(\frac{(1+n\sigma^2 \log n)L^2}{K\tau^2} + \frac{\Lambda^2 n L^5}{K\tau}\right).$$

Now choose the constant in (4.11) sufficiently large. Then each of the last two terms is at most 10^{-5}, and therefore

$$\mathsf{KL}(\Pi_\tau^y \| \mathsf{Law}(x_K)) \leq 10^{-4}.$$

Step IV: transfer of localization from Π_τ^y to $\mathsf{Law}(x_K)$. By Pinsker's inequality,

$$\mathsf{TV}(\mathsf{Law}(x_K), \Pi_\tau^y) \leq \sqrt{\frac{1}{2}\,\mathsf{KL}(\Pi_\tau^y \| \mathsf{Law}(x_K))} \leq \sqrt{5 \times 10^{-5}} < 0.008.$$

Hence, on the event $\mathcal{G}$,

$$\mathsf{Law}(x_K)(E) \geq \Pi_\tau^y(E) - \mathsf{TV}(\mathsf{Law}(x_K), \Pi_\tau^y) \geq 0.999 - 0.008 > 0.991.$$

Therefore

$$\mathbb{P}(x_K \in E) \geq \mathbb{P}(\mathcal{G}) \cdot 0.991 \geq 0.999 \times 0.991 > 0.99.$$

This proves the claimed coordinatewise error bound. □

4.4 Related works

In addition to the literature in diffusion discussed before, this work is further related to analysis of posterior sampling and sparse inference.

Theory of posterior sampling with diffusion priors. Most existing theory for posterior sampling with general diffusion priors only provides asymptotic consistency guarantees. Several provably consistent approaches are developed, based on techniques such as titled transport [14], plug-and-play [127], and sequential Monte Carlo methods [17, 30, 126]. However, none of these methods admit non-asymptotic performance guarantees. [68] developed a deterministic recovery theory leveraging the connection between diffusion and denoising operators [59], however it does not provide a convergence guarantee with polynomial complexities since it hinges on an unspecified burn-in phase for convergence. A notable recent development is Xun et al. [129], which developed a provably efficient posterior sampling algorithm using annealed Langevin dynamics with polynomial iteration complexities, assuming the prior distribution is log-concave. They subsequently went beyond the log-concavity assumption by assuming access to direct measurements of the signal of interest at sufficiently low noise levels, which generally are not available in practice.

High-dimensional sparse inference. Bayesian approaches has long been applied to high-dimensional sparse linear regression, e.g., George and McCulloch [37], Ishwaran and Rao [55], Yang et al. [130]. Due to the challenge of sampling from multimodal distributions, recent efforts have been focused on developing efficient sampling algorithms tailored to the sparsity-inducing distributions, such as the spike-and-slab prior [57, 64, 85]. Our work can be specialized to this case, providing a complementary algorithmic angle with efficient sampling guarantees, highlighting the benefit of annealing provided by diffusion priors.

4.5 Discussion

The result presented here applies to the heuristic solver (4.3) for the sparse denoising problem (4.2) under a Gaussian-Bernoulli prior. Beyond this specific setting, it provides a first step toward a mathematical understanding of diffusion-based inverse problem solvers and, more broadly, guided diffusion models. The analysis highlights

several directions for future work:

- *Compressed sensing.* An important next step is to extend the theory from denoising to linear inverse problems such as compressed sensing. Establishing analogous guarantees would likely require new ideas that combine structural properties of the forward operator, such as restricted isometry property, with a quantitative analysis of the diffusion dynamics.
- *General sparse prior.* The Gaussian-Bernoulli model serves as a convenient test case because the its diffused version admits explicit formulas. However, many applications involve richer sparse priors, such as low-rank or group-sparsity. Extending the result to broader classes of sparse priors would help clarify the extent to which diffusion-based solvers adapt to realistic low-dimensional structures.
- *Other types of solvers.* In practice, diffusion-based inverse problems are often solved using a wide range of tools, including predictor-corrector methods, proximal or projection-based updates, annealed Langevin variants, and our plug-and-play samplers. Developing a unified theory that covers such algorithms would be highly valuable.

Chapter 5

Polynomial Iteration Complexity for Riemannian Diffusion Models

In this chapter, we first introduce the RSGM algorithm in De Bortoli et al. [27] for completeness. Then, we offer a polynomial convergence guarantee in Theorem 4. This chapter is based on [128].

5.1 Diffusion on a manifold

We recall the setup for diffusion processes on Riemannian manifolds introduced in Cheng et al. [21], De Bortoli et al. [27]. Let $(B_t)_{t\geq 0}$ be a standard Brownian motion in $\mathbb{R}^d$ and $U_x : \mathbb{R}^d \to T_x\mathcal{M}$ any orthonormal frame at x. The *Geometric Brownian motion* solves

$$\mathrm{d}X_t = U_{X_t} \circ \mathrm{d}B_t,$$

where $\circ$ denotes Stratonovich integral, and its transition density $p_t(x, y)$ with respect to μ solves the heat equation

$$\partial_t p_t(\cdot, y) = \frac{1}{2}\Delta_{\mathcal{M}} p_t(\cdot, y).$$

Equivalently, Brownian motion can be defined abstractly as the solution to the martingale problem for the operator $\frac{1}{2}\Delta_{\mathcal{M}}$. Concretely, for any $f \in C^\infty([0, \infty) \times \mathcal{M})$,

the process

$$M_t^f := f(t, X_t) - f(0, X_0) - \int_0^t \left(\partial_s + \frac{1}{2}\Delta_{\mathcal{M}}\right) f(s, X_s)\mathrm{d}s$$

is a martingale with respect to the natural filtration of X. More generally, a forward diffusion process with drift is given by

$$\mathrm{d}X_t \;=\; b_t(X_t)\,\mathrm{d}t + U_{X_t} \circ \mathrm{d}B_t,$$

with Fokker–Planck equation $\partial_t p_t = -\nabla(b_t p_t) + \frac{1}{2}\Delta_{\mathcal{M}} p_t$. Note that in this setting, the following process is a martingale for smooth f:

$$M_t^f := f(t, X_t) - f(0, X_0) - \int_0^t \left(\partial_s f + \langle b_t, \nabla f\rangle + \frac{1}{2}\Delta_{\mathcal{M}} f\right)(s, X_s)\mathrm{d}s. \tag{5.1}$$

Let p_t denote the density of X_t w.r.t. μ, and define the *score* $s_t := \nabla \log p_t$. The time-reversal identity on manifolds yields a reverse SDE on a given time interval $[0, T]$:

$$\mathrm{d}\widetilde{X}_\tau = (-b_\tau(\widetilde{X}_\tau) + \nabla \log p_\tau(\widetilde{X}_\tau))\mathrm{d}t + U_{\widetilde{X}_\tau} \circ \mathrm{d}B_t, \quad p_{\widetilde{X}_T} = p_{X_T},\ t \in [0, \tau_\infty],\ \tau = T - t.$$

Again, in practice, we often only have access to an approximation $\hat{s}_t$ of $\nabla \log p_t$, which can be trained with the Riemannian score matching technique as in De Bortoli et al. [27].

On compact manifolds, $-\Delta_{\mathcal{M}}$ admits a spectral gap $\lambda_1 > 0$. Any initial distribution mixes to the uniform distribution μ along the heat flow with rate $\mathrm{e}^{-\lambda_1 t}$.

Forward and backward processes. For Riemannian diffusion, We use the standard heat flow as the forward process for simplicity:

$$\mathrm{d}X_t \;=\; U_{X_t} \circ \mathrm{d}B_t, \qquad X_0 \sim p_0.$$

The time-reversal identity yields the *reverse-time SDE*

$$\mathrm{d}Y_\tau \;=\; \nabla \log p_\tau(Y_\tau)\,\mathrm{d}t \;+\; U_{Y_\tau} \circ \mathrm{d}B_t, \quad Y_T \sim p_T,\ t \in [0, T],\ \tau = T - t. \tag{5.2}$$

Then time-reversal theory implies that $Y_0 \overset{(\mathrm{d})}{=} X_0$.

Assumptions on the manifold and the score. Throughout our analysis, we assume the manifold $\mathcal{M}$ has bounded curvature and has nondegenerate normal neighborhoods. We also need mild regularity assumptions on the score estimate $\hat{s}_t$. This is formalized by the following assumptions.

Assumption 2 (Regularity)**.** *Let $(\mathcal{M}, g)$ be a connected, compact d-dimensional Riemannian manifold. We assume:*

*(**A1**) **Positive injectivity radius:** there exists some $K \geq 1$ such that the injective radius $\geq 1/K$.*

*(**A2**) **Uniform curvature bounds:** for the same constant K (which obviously can be enlarged if necessary), we have*

$$\max\left\{ \mathrm{Diam}(\mathcal{M}), \|\mathrm{Rm}\|_{L^\infty}, \|\nabla \mathrm{Rm}\|_{L^\infty}, \|\nabla^2 \mathrm{Rm}\|_{L^\infty} \right\} \leq K.$$

*(**A3**) **Regularity of score estimates:** there exists a polynomial $\mathrm{poly}(d, K)$, such that*

$$\|\hat{s}_t(x)\| \leq \mathrm{poly}(d, K)\left(\|\nabla \log p_t(x)\| + t^{-1}\mathrm{Diam}^2(\mathcal{M})\right), \quad \forall x \in \mathcal{M}.$$

In Assumption 2, we made the standard "bounded geometry" assumption; similar assumptions also occur in Cheng et al. [20], De Bortoli et al. [27]. A positive injective radius ensures that we have sufficient room to operate on the tangent spaces as a proxy of operating on manifolds, since for every $x \in \mathcal{M}$, the exponential map $\exp_x$ is a diffeomorphism on the geodesic ball within injective radius. Bounds on Riemannian tensors rule out pathological cases, which helps to control the error propagation along the reverse diffusion. Lastly, compactness ensures a positive spectral gap of $\Delta_{\mathcal{M}}$ with $\lambda_1 > 0$, which is necessary to guarantee that the forward process mixes. The mild assumption (A3) on the score estimates avoids excessively large drifts in diffusion, and can be implemented easily in practice by clipping. In addition to the above, we also need a standard assumption on the score estimation error [18].

Assumption 3 (Score estimation error)**.** *There exists $\varepsilon_{\mathsf{score}} > 0$ such that*

$$\sum_{k=1}^{N}(t_k - t_{k-1})\mathbb{E}\|\hat{s}_{t_k}(Y_{t_k}) - \nabla \log p_{t_k}(Y_{t_k})\|^2 \leq \varepsilon_{\mathsf{score}}^2.$$

Goal. Our goal is to show that the reverse-time SDE (5.2) can be approximated with $\hat{s}_t$ in place of the exact score and with a discrete process (y_k) with polynomially many steps, such that the output of the discrete process obeys

$$\mathsf{TV}(p_0, \mathsf{Law}_{y_0}) \leq \varepsilon + \varepsilon_{\mathsf{score}},$$

for some small $\varepsilon > 0$ and for $\varepsilon_{\mathsf{score}}$ some characterization of the score estimation error.

5.2 Discretization of the reverse-time SDE

Recall the reverse-time SDE defined in (5.2). We discretize it at equidistributed time

$$\delta = t_0 < t_1 < \cdots < t_N = T, \quad t_k - t_{k-1} \equiv \frac{T-\delta}{N} =: h, \quad k = 1, \cdots, N.$$

In Algorithm 5, we provide an outline of discretized reverse-time SDE on Riemannian manifold, modified from De Bortoli et al. [27]. In each reverse step $k \in \{N, \ldots, 1, 0\}$, we select an orthonormal frame U_k at y_k, then sample Gaussian noise g_k and lift it to the tangent space $T_{y_k}\mathcal{M}$ using the orthonormal frame, obtaining $G_k \in T_{y_k}\mathcal{M}$. Afterwards, we propose a tangent update $\Delta_k = h\hat{s}_{t_k}(y_k) + \sqrt{h}G_k$ and the project to the manifold using the exponential map. To prevent the update from exitting the injective radius, we perform a rejection sampling step that rejects exceedingly large update. The algorithm terminates at $k = 0$ and returns the final iterate y_0. In this way, we ensure every update is well-defined in normal coordinates during the algorithm.

5.3 Theoretical guarantee

We are now ready to present our main quantitative guarantee of RSGM, as outlined in the following TV-accuracy bound.

Theorem 4. *Assume Assumptions 2 and 3 hold. There exists some universal constant $C, C' > 0$ such that the following holds. If $T \geq \frac{C}{\lambda_1}(d\log(Kd) + K + \log(\frac{N}{\varepsilon}))$,*

Algorithm 5 Riemannian Score-Based Generative Models (RSGM)

Manifold $(\mathcal{M}, g)$; score $\hat{s}_t$; early stopping time $\delta > 0$; reverse time grid $\delta = t_0 < t_1 < \cdots < t_N = T$; step size $h = t_k - t_{k-1}$; initial $x_N \sim \mu$ (uniform distribution);

for $k \in \{N, \ldots, 1, 0\}$ **do**
  Choose an orthonormal frame U_k at y_k, which is a linear map from $\mathbb{R}^d$ to $T_{y_k}\mathcal{M}$.
  $g_k \sim \mathcal{N}(0, I_d)$ in $\mathbb{R}^d$; $\quad G_k \leftarrow U_k g_k \in T_{y_k}\mathcal{M}$.
  $\Delta_k \leftarrow h\hat{s}_{t_k}(y_k) + \sqrt{h}\, G_k \in T_{y_k}\mathcal{M}$
  if $\|\Delta_k\| \le h^{1/4}$ **then**
    $y_{k-1} \leftarrow \exp_{y_k}(\Delta_k)$
  else
    $y_{k-1} \sim \mu$
  end if
end for
return y_0

then the output y_0 *of Algorithm 5 obeys*

$$\mathsf{TV}(p_\delta, \mathsf{Law}(y_0)) \le \varepsilon + C'\varepsilon_{\mathsf{score}} + \sqrt{hT}\,\mathrm{poly}(d, K, \delta^{-1}),$$

where h *is the discretization step size,* $\lambda_1 > 0$ *is the mixing rate of the geometric Brownian motion on* $\mathcal{M}$*, i.e., the smallest eigenvalue of* $-\Delta_{\mathcal{M}}$ *in* $L^2(\mu)$*.*

A few remarks are in order.

Iteration complexity. The error bound decomposes cleanly into three terms: ε results from mixing of the heat semigroup at the spectral gap λ_1, $\varepsilon_{\mathsf{score}}$ captures error from imperfect score estimation, and $\sqrt{hT}\,\mathrm{poly}(d, K, \delta^{-1})$ is the discretization error controlled by the step size and curvature. Consequently, choosing $T \asymp \lambda_1^{-1}(d\log d + \log(d/\varepsilon))$ and $h = \frac{\varepsilon^2}{\mathrm{poly}(d,K,\delta^{-1})T}$, then the TV error is bounded by $\varepsilon + \varepsilon_{\mathsf{score}}$ after polynomially many iterations

$$N = T/h \asymp \frac{\mathrm{poly}(d, K, \delta^{-1})}{(\lambda_1\varepsilon)^2}.$$

Compared to prior convergence rates in the Wasserstein metric [27], which require exponential complexity, we achieve polynomial iteration complexity for Riemannian diffusion models for the first time. Nonetheless, we emphasize that

Table 5.1: Comparison of the current theoretical guarantees on diffusion probabilistic models on Euclidean spaces and manifolds. Here, $\lambda_1 > 0$ is the spectral gap of the Laplace–Beltrami operator.

Work	Structure	Metric	Iteration complexity	Data distribution
[7]	Euclidean	TV	$\widetilde{O}(d/\varepsilon^2)$	bounded moment
[72]	Euclidean	TV	$\widetilde{O}(\mathrm{poly}(d)/\varepsilon)$	bounded support
[71]	Euclidean	TV	$\widetilde{O}(d/\varepsilon)$	bounded moment
[27]	Manifold	W_p	$\exp(O(d))\,\varepsilon^{-1/\lambda_1}$	smooth, strictly positive
This work	Manifold	TV	$\widetilde{O}\left(\frac{\mathrm{poly}(d)}{\lambda_1^2\varepsilon^2}\right)$	None (early stopping)

TV and Wasserstein distances are incomparable with each other in general, and our guarantee complements prior Wasserstein results [27] by ensuring distributional closeness in a different notion with a much smaller number of iterations. We provide a concrete comparison with prior art in Table 5.1.

Possible improvements. We note that the bound established in Theorem 4 holds under very mild geometric assumptions, requiring only constraints on the injective radius and Riemannian curvature. The purpose of this study is to demonstrate that, in the manifold setting, the exponential blow-up in T can be avoided and polynomial complexity can be achieved. To keep the exposition as simple as possible and to clearly highlight the key ideas, we have not attempted to optimize the current bound on the degree of the polynomial. Potential approaches for sharper bounds include: (i) a better design of discretization schedule, possibly adaptive to the manifold geoemetry, and a more careful computation of discretization error, such as those in Li and Jiao [69], Benton et al. [7] (notably, the dependence on δ might be improved to poly-logarithmic in this way); (ii) a tailored analysis for TV error that does not rely on Pinsker's inequality, like those in Li and Yan [71], may also be extended to manifolds; (iii) a tighter version of our Minakshisundaram-Pleijel parametrix bound. We leave these improvements as future work.

5.4 Proof outline

Throughout the proof, we assume that

$$h \le \frac{1}{\mathrm{poly}(d, K, \delta^{-1})}, \tag{5.3}$$

since otherwise the bound in Theorem 4 would be trivial (recall that TV distance is always bounded by 2). We start by recalling the sequence considered in RSGM. Let $(Y_k)_{k\in\{0,\dots,N\}}$ be given by $Y_N \sim \mu$ and for any $k \in \{0,\dots,N-1\}$:

$$Y_{k-1} = \begin{cases} \exp_{Y_k}\left[h\hat{s}_{t_k}(Y_k) + \sqrt{h}G_k\right], & \|h\hat{s}_{t_k}(Y_k) + \sqrt{h}\,G_k\| \le h^{1/4}, \\ \text{drawn from } \mu, & \text{otherwise.} \end{cases}$$

This defines a sequence of probability transition kernels $\widehat{\mathcal{K}}_{t_k,t_{k-1}}$. For simplicity, we denote this by $\widehat{\mathcal{K}}_k$. Let q_k be the law of Y_k. We have

$$q_0 = q_N\widehat{\mathcal{K}}_N\widehat{\mathcal{K}}_{N-1}\cdots\widehat{\mathcal{K}}_1.$$

Similarly, the probability transition kernel from time t_k to t_{k-1} in (5.2) is denoted by $\mathcal{K}_{t_k,t_{k-1}}$ or $\mathcal{K}_k$ in short. We have

$$p_0 = p_N\mathcal{K}_N\mathcal{K}_{N-1}\cdots\mathcal{K}_1.$$

Our goal would be to bound $\mathsf{TV}(p_0, q_0)$ as in Theorem 4, by decomposing the total error into four components:

$$(\text{initialization error}) + (\text{score error}) + (\text{drift discretization error}) + (\text{BM simulation error}).$$

More concretely:

- **Initialization error** arises from initializing Y_N with μ instead of the true marginal p_N;
- **Score error** arises from imperfect score estimation;
- **Drift discretization error** arises from approximating the continuous-time drift $\hat{s}_t(Y_t)$ by its "time-frozen" counterpart $\hat{s}_{t_k}(Y_{t_k})$;
- **Brownian motion (BM) simulation error** is a distinctive feature of the manifold setting. Unlike in Euclidean space — where the transition kernel of Brownian motion over $[t_k, t_{k-1}]$ is exactly Gaussian with variance $(t_k - t_{k-1})$ — the transition kernel of manifold-valued Brownian motion cannot be simulated exactly by any discrete-time process, even after time discretization. This inherent inexactness gives rise to this final error term.

The first two components are relatively easier to bound using well-established tools: mixing rate bounds of heat flow [119] and Girsanov transform [18]. For the drift discretization error, recent techniques developed in the Euclidean setting [7] can also be adapted with modifications that account for the manifold curvature. However, the last component — the Brownian motion simulation error — represents the core challenge in the manifold setting, which fundamentally denies a direct extension of Euclidean analysis.

Step I: Constructing auxiliary kernels via localization. In view of this, we first introduce an intermediate random process that separates the drift discretization error from the BM simulation error. Constructing such a process, however, involves additional technicality. In particular, the frozen drift $\hat{s}_{t_k}(Y_{t_k})$ is a vector in the tangent space $T_{Y_k}\mathcal{M}$, and is therefore only well-defined at the fixed point Y_k. This poses a compatibility issue: as Brownian motion evolves continuously on the manifold, it immediately departs from Y_k, rendering the frozen drift ill-defined. Careful geometric considerations are thus required to reconcile the piecewise-constant drift approximation with the intrinsic curvature of the manifold.

In our analysis, this is handled using localization by the construction of an auxiliary sequence of transition kernels $\mathcal{K}_k^{\mathsf{aux}}$. These kernels do not appear in the algorithm itself; they serve solely as an analytical tool to facilitate the proof. These kernels expose the behavior of the time-reverse SDE (5.2) when the estimated score $\hat{s}_t$ is frozen to be a constant vector field in between discretization steps, meanwhile keeping the continuous Brownian motion.

Let $\eta : [0, \infty) \to [0, 1]$ be a smooth cutoff function, i.e., η is decreasing, $\eta|_{[0,1]} \equiv 1$ and $\eta|_{[4,\infty)} \equiv 0$. Such a function can be chosen such that $|\eta'| + |\eta''| + |\eta'''| \le 100$. Recall that the injective radius of $\mathcal{M}$ is lower bounded by $1/K$, and the curvature is upper bounded by K. Define

$$\omega := \frac{c_\omega}{K d^4}, \qquad \eta_\omega(r) = \eta\left(\frac{4r^2}{\omega^2}\right), \quad r \ge 0, \tag{5.4}$$

where $c_\omega > 0$ is a small universal constant. We have $\eta_\omega|_{[0,\frac{\omega}{2}]} \equiv 1$ and $\eta_\omega|_{[\omega,\infty)} \equiv 0$. For $t > 0$, $x, y \in \mathcal{M}$, define the following vector field on $\mathcal{M}$:

$$\mathscr{S}_{t,x}(y) = (\mathrm{d}\exp_x)_{\log_x y}\left(\eta_\omega(\rho(x,y)) \cdot \hat{s}_t(x)\right) \in T_y\mathcal{M}.$$

Intuitively speaking, $\mathscr{S}_{t,x}(\cdot)$ is the "constant" velocity field $\hat{s}_t(x)$ in normal coordinates, which represents our idea of freezing the drift term for a time period. The $\mathrm{d}\exp_x$ in the formula is responsible for identifying $T_y\mathcal{M}$ with $T_x\mathcal{M}$. [1] On the other hand, the cut-off function η_ω is necessary to keep all our discussions restricted to the injective radius, so as to avoid pathologies of cut locus.

With this in mind, we are ready to define $\mathcal{K}_k^{\mathsf{aux}}$ as the transition kernel from time t_k to t_{k-1} of the reverse-time SDE

$$\mathrm{d}Y_\tau = \mathscr{S}_{t_k, Y_{t_k}}(Y_\tau)\mathrm{d}t + U_{Y_\tau} \circ \mathrm{d}W_t, \quad \tau = T - t, \quad \tau \in [t_{k-1}, t_k], \tag{5.5}$$

and in addition,

$$p_k^{\mathsf{aux}} = p_N \mathcal{K}_N^{\mathsf{aux}} \mathcal{K}_{N-1}^{\mathsf{aux}} \cdots \mathcal{K}_{k+1}^{\mathsf{aux}}, \quad k = N, N-1, \cdots, 0.$$

Step II: Decomposing different sources of error. We now decompose

$$\mathsf{TV}(p_0, q_0) \le \mathsf{TV}(p_0, p_0^{\mathsf{aux}}) + \mathsf{TV}(p_0^{\mathsf{aux}}, q_0) \le \sqrt{2\mathsf{KL}(p_0 \parallel p_0^{\mathsf{aux}})} + \mathsf{TV}(p_0^{\mathsf{aux}}, q_0),$$

where the last inequality used Pinsker's inequality. To control $\mathsf{KL}(p_0 \parallel p_0^{\mathsf{aux}})$, we further introduce the counterpart of $\mathscr{S}_{t,x}$ using the exact score function $\nabla \log p_t$:

$$\mathscr{S}^\star_{t,x}(y) = (\mathrm{d}\exp_x)_{\log_x y}\left(\eta_\omega(\rho(x,y)) \cdot \nabla \log p_t(x)\right) \in T_y\mathcal{M}.$$

We apply Girsanov's theorem [51] to compare (5.5) with (5.2), in a way that is standard in recent literature [18, 27]. Denote the path law of the solution of (5.2) by $\mathsf{Law}(Y)$, and the path law of the solution of (5.5) by $\mathsf{Law}(Y^{\mathsf{aux}})$. Girsanov's theorem asserts that the KL divergence $\mathsf{KL}(\mathsf{Law}(Y) \parallel \mathsf{Law}(Y^{\mathsf{aux}}))$ is upper bounded by the expectation of the squared norm of the difference between the drift terms in the two SDEs.[2] More concretely,

$$\mathsf{KL}\big(\mathsf{Law}(Y) \parallel \mathsf{Law}(Y^{\mathsf{aux}})\big) \le \sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \mathbb{E}\left\|\nabla \log p_t(Y_t) - \mathscr{S}_{t_k, Y_{t_k}}(Y_t)\right\|^2 \mathrm{d}t.$$

[1] Generally speaking, it is more natural to use parallel transport to identify different tangent spaces. However, this would later lead to a more complicated treatment of the perturbed heat equation with variable drifts. We choose to use parallelism in normal coordinates instead for simplicity.

[2] In its classical form, Girsanov's theorem requires integrability such as Novikov's condition to

Since p_0 and p_0^{aux} are marginals of $\mathsf{Law}(Y)$ and $\mathsf{Law}(Y^{\mathsf{aux}})$ respectively at time $t = t_0$, by post-processing inequality, we have

$$\begin{aligned}
\mathsf{KL}(p_0 \parallel p_0^{\mathsf{aux}}) &\le \sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \mathbb{E} \left\| \nabla \log p_t(Y_t) - \mathscr{S}_{t_k, Y_{t_k}}(Y_t) \right\|^2 \mathrm{d}t \\
&\le \underbrace{2 \sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \mathbb{E} \|\nabla \log p_t(Y_t) - \mathscr{S}^{\star}_{t_k, Y_{t_k}}(Y_t)\|^2 \mathrm{d}t}_{\text{drift discretization}} \\
&\quad + \underbrace{2 \sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \mathbb{E} \|\mathscr{S}_{t_k, Y_{t_k}}(Y_t) - \mathscr{S}^{\star}_{t_k, Y_{t_k}}(Y_t)\|^2 \mathrm{d}t}_{\text{score matching}} . \qquad (5.6)
\end{aligned}$$

It remains to decompose $\mathsf{TV}(p_0^{\mathsf{aux}}, q_0)$. To isolate the initialization error, we introduce

$$q_0^{\star} = p_N \widehat{\mathcal{K}}_N \widehat{\mathcal{K}}_{N-1} \cdots \widehat{\mathcal{K}}_1 .$$

By triangle inequality and post-processing inequality, we have

$$\mathsf{TV}(p_0^{\mathsf{aux}}, q_0) \le \mathsf{TV}(p_0^{\mathsf{aux}}, q_0^{\star}) + \mathsf{TV}(q_0^{\star}, q_0) \le \underbrace{\mathsf{TV}(p_0^{\mathsf{aux}}, q_0^{\star})}_{\text{BM simulation}} + \underbrace{\mathsf{TV}(p_N, q_N)}_{\text{initialization}} .$$

Step III: Controlling initialization and score errors. By our design, $q_N = \mu$, and $\mathsf{TV}(p_N, q_N) = \mathsf{TV}(p_N, \mu)$. This is known as the mixing rate of heat flow in total variation norm, and has well-established bounds, e.g., Urakawa [119]. The score-matching error, on the other hand, can be controlled with an analysis on the distortion on the Riemannian metric in normal coordinates. We compile the bounds into the following lemma.

Lemma 9. *There exists a universal constant $C > 0$, such that whenever $T \ge 1$, we have*

$$\mathsf{TV}(p_N, q_N) \le \mathrm{e}^{C(K + d \log d)} \mathrm{e}^{-\frac{\lambda_1}{2}(T - \frac{1}{2})},$$

hold. In our setting, this can be bypassed with a localization argument as in Chen et al. [18].

and

$$\sum_{k=1}^{N}\int_{t_{k-1}}^{t_k}\mathbb{E}\|\mathscr{S}_{t_k,Y_{t_k}}(Y_t)-\mathscr{S}^\star_{t_k,Y_{t_k}}(Y_t)\|^2\mathrm{d}t\le 2\varepsilon^2_{\mathsf{score}}.$$

Step IV: Controlling drift discretization error with Itô/Stratonovich calculus and Li-Yau estimates. The drift discretization error defined in (5.6) has a similar form to the discretization error for the Euclidean setting [7], though additional complications arise due to non-constant $\mathscr{S}^\star_{t_k,Y_{t_k}}$. The idea is to study the time derivative of $\mathbb{E}\|\nabla\log p_t(Y_t)-\mathscr{S}^\star_{t_k,Y_{t_k}}(Y_t)\|^2$, which in view of $\partial_\tau\log p_t=-\frac{1}{2}\Delta_{\mathcal{M}}p_t$ (negative sign due to reverse time) involves space derivatives of $\log p_t$ up to third order. Fortunately, after applying Itô/Stratonovich calculus to simplify the expression, a key property in the proof of the Euclidean setting carries over: third-order derivatives of $\log p_t$ cancel out. The remaining first and second-order derivatives can be controlled by Li-Yau estimates on the log-gradient of the heat kernel. We obtain

Lemma 10. *Under the assumptions in Theorem 4, there is a universal constant $C>0$ such that*

$$\sum_{k=1}^{N}\int_{t_{k-1}}^{t_k}\mathbb{E}\|\nabla\log p_t(Y_t)-\mathscr{S}^\star_{t_k,Y_{t_k}}(Y_t)\|^2\mathrm{d}t\le\frac{Cd^6K^8}{\delta^3}h^2N.$$

Step V: Controlling BM simulation error using parametrix estimates. Our approach is inspired by the following consequence of post-processing inequality and Pinsker's inequality:

$$\mathsf{TV}(p_0^{\mathsf{aux}},q_0^\star)\le\sqrt{2\mathsf{KL}(p_0^{\mathsf{aux}}\parallel q_0^\star)}\le\sqrt{2\sum_{k=1}^{N}\mathsf{KL}(p_k^{\mathsf{aux}}\mathcal{K}_k^{\mathsf{aux}}\parallel p_k^{\mathsf{aux}}\widehat{\mathcal{K}}_k)}.$$

This leads us to compare the kernel $\mathcal{K}_k^{\mathsf{aux}}$ and $\widehat{\mathcal{K}}_k$. In normal coordinates, the Fokker-Planck equation shows that these two are the solutions of the heat equations with the Euclidean Laplacian and with the manifold Laplace-Beltrami operator. We utilize the Minakshisundaram-Pleijel parametrix theory [8] in geometric analysis for this comparison, and establish a quantitative bound in polynomially small radius and polynomially short time (cf. Lemma 38).

5.5 Related works

There are two lines of work that are most closely related to ours: convergence theory for Euclidean diffusion models, and sampling algorithm on manifolds. We review these works below.

Non-asymptotic convergence theory for Euclidean diffusion models. Early convergence analyses of diffusion models require L_∞-accurate score estimates [26]. For stochastic samplers such as DDPM [49], early bounds under Lipschitz/smoothness assumptions of the data distribution admit an $O(T^{-\frac{1}{2}})$ iteration complexity in the total variation distance assuming L_2-accurate score estimates [18], with subsequent analyses relaxing the Lipschitz assumption yet retaining the same complexity [7, 66, 72]. More recently, Li and Yan [71] has improved the iteration complexity to $\widetilde{O}(T^{-1})$. For deterministic samplers, Chen et al. [18] established polynomial convergence with exact scores, and Li et al. [72] established a convergence rate of $O(T^{-1})$ under L_2-accurate scores. See Beyler and Bach [9], Li and Jiao [69], Liang et al. [77] for additional analyses that established convergence in the Wasserstein distance and improved discrete-time rates. Several works [53, 70, 76, 94] also developed non-asymptotic convergence rates of diffusion models under the manifold hypothesis, suggesting diffusion models are adaptive to low-dimensional structures. This line of work should not be confused with ours, where the diffusion process is designed specifically to be constrained on the manifold.

Sampling on Riemannian manifold. Cheng et al. [20, 21] analyzed the geometric Euler–Maruyama (EM) discretization for time-homogeneous SDEs, and proved a polynomial complexity guarantee under dissipative-distant geometric assumptions on the manifold. See also Bharath et al. [10] for follow-ups. Guan et al. [42] proposed a Riemannian proximal sampler with convergence guarantees under the log-Sobolev inequality. Various sampling algorithms are also studied for a related problem known as sampling from constrained spaces [3, 115]. Nonetheless, convergence analyses of Riemannian diffusion models under general data distributions remain highly limited, with De Bortoli et al. [27] being the only prior work with non-asymptotic convergence rates.

5.6 Discussion

We developed a discrete-time theory for Riemannian diffusion models showing that a polynomial stepsize suffices for TV-accurate sampling under mild geometric conditions. In particular, our results show that choosing a stepsize polynomially small in manifold parameters achieves any prescribed TV target without exponential blow-ups in dimension or curvature. This complements prior Wasserstein-type guarantees which requires exponentially manp steps. Several important future directions remain open.

- *Sharper bounds.* For simplicity, we did not attempt to establish sharp bounds for the error terms in our analysis, and it is likely that the degree of the polynomial in the bound could be improved significantly by refining our analysis.
- *Analysis of deterministic samplers.* We focused on DDPM-style stochastic samplers in our analysis. For practical purpose, it is also tempting to develop an analogous theory for DDIM-style deterministic samplers.

Chapter 6

Manifold DPnP: Constrained Solvers with Riemannian Diffusion Models

Having established the theoretical foundations of manifold diffusion models, we proceed to develop the framework to solve constrained inverse problems with prior knowledge encoded by diffusion models.

6.1 DPnP as a heat flow

We begin by recalling the structure of DPnP. At each iteration, the algorithm alternates between two sub-samplers.

1. A *proximal consistency step*, which, given the current iterate $\hat{x}_k$, samples from the distribution with density proportional to

$$\exp\left(\mathcal{L}(x;y) - \frac{1}{2\eta_k^2}\|x - \hat{x}_k\|^2\right).$$

   Here $\mathcal{L}(x;y)$ denotes the likelihood function, and $\eta_k > 0$ is the annealing parameter at iteration k.

2. A *diffusion denoising step*, which, given the intermediate point $\hat{x}_{k+\frac{1}{2}}$, samples from the distribution with density proportional to

$$p^\star(x)\exp\left(-\frac{1}{2\eta_k^2}\|x - \hat{x}_{k+\frac{1}{2}}\|^2\right),$$

where $p^\star$ is the distribution of clean data.

In Euclidean space, these two steps are natural: the quadratic penalty $\|x - z\|^2$ plays the role of a proximal regularizer, preventing drastic changes across iterates while allowing either the likelihood term or the prior term to reshape the distribution. On a manifold, however, this formulation immediately raises a conceptual difficulty. What should replace the Euclidean quadratic cost $\|x - z\|^2$ when x and z lie on a curved space?

At first sight, two natural candidates present themselves.

- *Extrinsic Euclidean distance.* One may embed the manifold isometrically into an ambient Euclidean space and use the squared ambient Euclidean distance. This approach is unsatisfactory for several reasons. Although Nash's embedding theorem guarantees the existence of an isometric embedding, such an embedding may be difficult to construct explicitly and may require a much higher-dimensional ambient space than the intrinsic dimension of the manifold, creating substantial computational overhead. More fundamentally, the resulting distance is not intrinsic: it depends on the chosen embedding. Different embeddings can distort global geometries in different ways. For example, points that are far apart along the manifold may become artificially close in the ambient space.
- *Geodesic distance.* A more intrinsic alternative is to replace $\|x-z\|^2$ by $\rho(x, z)^2$, where ρ is the Riemannian geodesic distance. This choice respects the geometry of the manifold, but it is generally expensive to compute and often difficult to differentiate globally. Even more importantly for our purposes, it is not clear how to sample efficiently from the corresponding Gibbs distributions in a way that is compatible with the Riemannian diffusion framework introduced earlier.

A crucial observation of this chapter is that the proximal form of DPnP admits an equivalent dynamical interpretation that avoids these difficulties altogether. Rather than viewing each substep as sampling from a Gibbs distribution with a quadratic penalty, we reinterpret the algorithm as a short-time evolution under a suitable heat flow. This reformulation is especially valuable on manifolds, since Brownian motion and the heat equation possess canonical intrinsic analogues in the Riemannian setting.

More precisely, each proximal sampling step can be realized as the terminal

distribution of a reverse-time diffusion over a time interval of length η^2. Composing the two substeps of DPnP therefore yields a piecewise diffusion process: the first segment enforces the prior through reverse heat flow from the data distribution, while the second enforces consistency through reverse heat flow from the likelihood-weighted distribution.

To state this precisely, let

$$q_0(x) \propto \exp(\mathcal{L}(x; y))$$

be the consistency density, normalized so that q_0 is a probability density whenever the normalizing constant is finite. For $t \geq 0$, let p_t and q_t denote the heat evolutions of $p_0 \coloneqq p^\star$ and q_0, respectively; that is, p_t and q_t are the densities of Brownian motion at time t initialized from $p^\star$ and q_0.

With this notation, one iteration of DPnP can be characterized as follows.

Theorem 5. *Fix $x \in \mathbb{R}^d$ and $\eta > 0$. Starting from $X_0 = x$, one step of the proximal consistency sampler followed by one step of the diffusion denoising sampler with annealing parameter η is equivalent in law to running the diffusion*

$$\mathrm{d}X_t = \begin{cases} \nabla \log q_{\eta^2 - t}(X_t)\,\mathrm{d}t + \mathrm{d}B_t, & 0 \leq t \leq \eta^2, \\ \nabla \log p_{2\eta^2 - t}(X_t)\,\mathrm{d}t + \mathrm{d}B_t, & \eta^2 < t \leq 2\eta^2, \end{cases}$$

where $(B_t)_{t \geq 0}$ is standard Brownian motion, p_t is the heat flow starting from $p^\star$, and q_t is the heat flow starting from

$$q_0(x) \propto \exp(\mathcal{L}(x; y)).$$

More precisely, the output of one full DPnP *iteration is given by the endpoint $X_{2\eta^2}$ of this piecewise diffusion process.*

This theorem provides the key bridge from the Euclidean formulation of DPnP to its manifold counterpart. The right-hand side is expressed entirely in terms of Brownian motion and score fields of heat-evolved densities, both of which admit natural intrinsic generalizations on a Riemannian manifold. In particular, the theorem suggests that the correct manifold analogue of DPnP is not obtained by searching for a direct replacement of the quadratic penalty $\|x - z\|^2$, but rather by transporting

the equivalent heat-flow formulation to the manifold setting.

This viewpoint has both conceptual and practical advantages. Conceptually, it reveals that the split Gibbs structure of DPnP is fundamentally a discretized alternation of heat flows. Practically, it leads to an implementation based on Riemannian Brownian motion and manifold score models, thereby avoiding the need for extrinsic embeddings or repeated geodesic optimization. This will serve as the starting point for the manifold constrained solvers developed in the remainder of this chapter.

6.2 Formulation of Manifold DPnP

We now formulate DPnP intrinsically on a general Riemannian manifold. The heat-flow interpretation from the previous section suggests that the correct manifold generalization should be built not from a direct replacement of the Euclidean quadratic penalty, but rather from reverse-time Riemannian diffusions associated with heat-evolved densities.

6.2.1 Notation

Recall that $(\mathcal{M}, g)$ is a connected, compact d-dimensional Riemannian manifold with normalized volume measure μ. Let $B_t^{\mathcal{M}}$ denote Brownian motion on $\mathcal{M}$, namely the diffusion process with generator $\frac{1}{2}\Delta_{\mathcal{M}}$. For any probability density ρ_0 on $\mathcal{M}$ with respect to μ, we denote by

$$\rho_t \coloneqq \mathrm{e}^{\frac{t}{2}\Delta_{\mathcal{M}}}\rho_0$$

its heat evolution, that is, the density at time t of $B_t^{\mathcal{M}}$ initialized from ρ_0.

Given an observation y, define the consistency density

$$q_y^\star(x) \coloneqq \frac{1}{Z(y)}\exp(\mathcal{L}(x;y)), \qquad Z(y) \coloneqq \int_{\mathcal{M}} \exp(\mathcal{L}(x;y))\,\mathrm{d}\mu(x),$$

Here $\mathcal{L}(\cdot;y)$ is the same measurement log-likelihood function as in the Euclidean formulation.

Let

$$p_t \coloneqq \mathrm{e}^{\frac{t}{2}\Delta_{\mathcal{M}}}p^\star, \qquad q_{y,t} \coloneqq \mathrm{e}^{\frac{t}{2}\Delta_{\mathcal{M}}}q_y^\star, \tag{6.1}$$

and define the corresponding time-dependent score fields

$$s_p(x,t) \coloneqq \nabla \log p_t(x), \qquad s_q(x,t) \coloneqq \nabla \log q_{y,t}(x).$$

6.2.2 Reverse Riemannian diffusion as the manifold proximal operator

The heat-flow characterization from the previous section immediately leads to an intrinsic sampling operator.

Definition 3 (Reverse Riemannian diffusion operator)**.** Let ρ_0 be a probability density on $\mathcal{M}$, let $\rho_t = \mathrm{e}^{\frac{t}{2}\Delta_{\mathcal{M}}}\rho_0$, and let

$$s_\rho(x,t) \coloneqq \nabla \log \rho_t(x).$$

For $\eta > 0$, define $\mathsf{RD}^{\eta}_{\rho}(x)$ to be the law of X_{η^2}, where X_t solves the reverse-time Riemannian diffusion

$$\mathrm{d}X_t = s_\rho(X_t, \eta^2 - t)\,\mathrm{d}t + \mathrm{d}B_t^{\mathcal{M}}, \qquad X_0 = x, \qquad 0 \le t \le \eta^2. \tag{6.2}$$

This operator is the manifold analogue of one Euclidean proximal sampling step. Indeed, when $\mathcal{M} = \mathbb{R}^d$, the Brownian motion $B_t^{\mathcal{M}}$ reduces to ordinary Brownian motion, and (6.2) becomes exactly the reverse heat-flow SDE from the previous section.

We may therefore define manifold DPnP as the alternation of two such reverse diffusions, one associated with the consistency density $q_y^\star$, and one associated with the prior density $p^\star$.

Definition 4 (Manifold DPnP)**.** Fix an annealing schedule $\{\eta_k\}_{k=0}^{N-1}$. Starting from $x_0 \in \mathcal{M}$, define the iterates

$$x_{k+\frac{1}{2}} \sim \mathsf{RD}^{\eta_k}_{q_y^\star}(x_k), \qquad x_{k+1} \sim \mathsf{RD}^{\eta_k}_{p^\star}(x_{k+\frac{1}{2}}), \qquad k = 0, 1, \ldots, N-1. \tag{6.3}$$

The output x_N is called one sample generated by manifold DPnP.

In expanded form, this means that the proximal consistency step is given by the

terminal point of

$$\mathrm{d}X_t = s_q(X_t, \eta_k^2 - t)\,\mathrm{d}t + \mathrm{d}B_t^{\mathcal{M}}, \qquad X_0 = x_k, \qquad 0 \le t \le \eta_k^2,$$

and the prior denoising step is given by the terminal point of

$$\mathrm{d}X_t = s_p(X_t, \eta_k^2 - t)\,\mathrm{d}t + \mathrm{d}B_t^{\mathcal{M}}, \qquad X_0 = x_{k+\frac{1}{2}}, \qquad 0 \le t \le \eta_k^2.$$

Implementation of the exponential map. In the above algorithm, we assume that the exponential map on the underlying manifold can be computed efficiently. This assumption is satisfied for many manifolds that arise in science and engineering. Examples include flat tori, or products of circles, which appear naturally in angular and phase synchronization problems [106], and Grassmannians, which provide the natural search spaces for subspace learning and related low-rank estimation problems [46]. For these matrix manifolds, explicit formulas and efficient numerical implementations of geodesics, exponential maps, and closely related retractions are standard in Riemannian optimization [2].

6.2.3 Discretization by geodesic random walk

To implement the above diffusions, we discretize them using an Euler–Maruyama scheme as in Chapter 5. Given an estimated score field $\hat{s}(\cdot, t)$, this yields a discrete approximation of (6.2). Specifically, over a time horizon $\eta^2 > 0$ with m inner steps and step size $h = \eta^2/m$, we set

$$z_0 = x, \qquad t_j = \eta^2 - jh,$$

and iterate

$$z_{j+1} = \exp_{z_j}\big(h\,\hat{s}(z_j, t_j) + \sqrt{h}\,\xi_j\big), \qquad \xi_j \sim \mathcal{N}(0, I_{T_{z_j}\mathcal{M}}), \qquad j = 0, \dots, m-1. \tag{6.4}$$

The output z_m is our numerical approximation of $\mathsf{RD}_\rho^\eta(x)$.

In practice, the score fields s_p and s_q are not available in closed form and must

Algorithm 6 Manifold DPnP

Input: complete Riemannian manifold $(\mathcal{M}, g)$, observation y, annealing schedule $\{\eta_k\}_{k=0}^{N-1}$, prior score oracle $\hat{s}_p$, consistency score oracle $\hat{s}_q(\cdot, \cdot\,; y)$.

Hyperparameter: numbers of inner discretization steps $\{m_k^{(q)}\}_{k=0}^{N-1}$ and $\{m_k^{(p)}\}_{k=0}^{N-1}$.

Initialization: sample $x_0 \sim \mu$ on $\mathcal{M}$.

Update: **for** $k = 0, 1, \cdots, N-1$ **do**

(1) *Proximal consistency sampler*, $\mathsf{PCS}_{\mathcal{M}}$: set

$$x_{k+\frac{1}{2}} = \mathsf{GRW}\Big(x_k, \hat{s}_q(\cdot, \cdot\,; y), \eta_k^2, m_k^{(q)}\Big)\,.$$

This approximates one sample from the reverse Riemannian diffusion associated with the consistency density $q_y^\star$.

(2) *Denoising diffusion sampler*, $\mathsf{DDS}_{\mathcal{M}}$: set

$$x_{k+1} = \mathsf{GRW}\Big(x_{k+\frac{1}{2}}, \hat{s}_p, \eta_k^2, m_k^{(p)}\Big)\,.$$

This approximates one sample from the reverse Riemannian diffusion associated with the prior density $p^\star$.

Output: x_N.

be replaced by learned or approximate score oracles

$$\hat{s}_p(x, t) \approx \nabla \log p_t(x), \qquad \hat{s}_q(x, t) \approx \nabla \log q_{y,t}(x).$$

For the prior step, $\hat{s}_p$ is provided by the manifold diffusion model. For the consistency step, $\hat{s}_q$ may be obtained either analytically when the forward model is simple, or by a separate score-approximation procedure.

We may now state the full algorithm in Algorithm 6. The subroutine for geodesic random walk in shown in Algorithm 7.

6.3 Theoretical analysis

In this section, we extend the theoretical analysis of DPnP to the Riemannian setting. As in the Euclidean setting, we establish both asymptotic consistency and non-asymptotic robustness of the resulting algorithm. These guarantees are completely similar to their Euclidean counterparts. To avoid repetition, we only

Algorithm 7 Geodesic Random Walk Sampler $\mathsf{GRW}(x, \hat{s}, T, m)$

Input: starting point $x \in \mathcal{M}$, time-dependent score field $\hat{s}(\cdot, \cdot)$, diffusion time $T > 0$, number of discretization steps m.
Initialization: set $h = T/m$, $z_0 = x$.
Update: **for** $n = 0, 1, \cdots, m-1$ **do**

(1) Set the reverse-time level

$$t_n = T - nh.$$

(2) Sample a Gaussian tangent vector

$$\xi_n \sim \mathcal{N}(0, I_{T_{z_n}\mathcal{M}}).$$

(3) Perform one intrinsic Euler step:

$$z_{n+1} = \exp_{z_n}\big(h\,\hat{s}(z_n, t_n) + \sqrt{h}\,\xi_n\big).$$

Output: z_m.

state the final results without elaborating intermediate lemmas, and only sketch the proof as needed.

6.3.1 Asymptotic consistency

W state the analogue of Theorem 1 for compact manifolds. Its formulation is intentionally parallel to the Euclidean case.

Theorem 6 (Asymptotic consistency of manifold DPnP)**.** *Under a setting similar to that of Lemma 4 and Lemma 3 but adapted to manifold* DPnP*, the following holds.*

Let $\varepsilon_1 > \varepsilon_2 > \cdots$ *be a decreasing sequence of positive numbers satisfying* $\lim_{l\to\infty} \varepsilon_l = 0$*, and let*

$$0 = k_0 < k_1 < k_2 < \cdots$$

be an increasing sequence of integers. Set the annealing schedule as follows:

$$\eta_k = \varepsilon_l, \qquad \text{for } k_{l-1} \le k < k_l, \quad l = 1, 2, \cdots.$$

If

$$\min_{l'=1,2,\cdots} |k_{l'} - k_{l'-1}| \to \infty,$$

then the output $\hat{x}_{k_l}$ of the manifold version of DPnP *converges in distribution to the posterior distribution $p^\star(\cdot \mid y)$ as $l \to \infty$.*

In words, Theorem 6 shows that once the Euclidean Gaussian corruption is replaced by heat-kernel corruption on $\mathcal{M}$, and the Euclidean proximal step is replaced by its Riemannian counterpart, the same asymptotic consistency mechanism continues to hold.

6.3.2 Non-asymptotic error analysis

As in the Euclidean analysis, practical implementations introduce approximation errors: the manifold proximal sampler is run for finitely many steps with nonzero discretization, and the Riemannian score-based denoising sampler uses approximate scores and finitely many time discretization steps. We model these errors abstractly through per-iteration total variation bounds, uniformly over all iterations.

Recall the definition of $q_{y,t}$ in the beginning of this chapter. We define the stationary distribution under annealing level $\eta > 0$ as

$$\pi_\eta(\mathrm{d}x) \propto p^\star(x) q_{y,\eta^2}(x)\, \mu(\mathrm{d}x). \tag{6.5}$$

Since $\mathcal{M}$ is compact and $\mathcal{L}$ is continuous, we again have, as in the Euclidean setting, that

$$q_{y,\eta^2}(x) \to \mathrm{e}^{\mathcal{L}(x;y)} \qquad \text{uniformly in } x \in \mathcal{M}$$

as $\eta \to 0$, and thus π_η converges weakly to the posterior $p^\star(\cdot \mid y)$.

Theorem 7 (Non-asymptotic robustness of manifold DPnP)**.** *With the notation in manifold* DPnP*, set $\eta_k \equiv \eta > 0$. There exists*

$$\lambda \coloneqq \lambda(p^\star, \mathcal{L}, \eta) \in (0, 1),$$

such that the following holds.

Define π_η by (6.5). If the manifold proximal sampler $\mathsf{PCS}_\mathcal{M}$ has error at most $\varepsilon_{\mathsf{PCS},\mathcal{M}}$ in total variation per iteration, and the Riemannian score-based denoising sampler $\mathsf{DDS}_\mathcal{M}$ has error at most $\varepsilon_{\mathsf{DDS},\mathcal{M}}$ in total variation per iteration, then for

any accuracy goal $\varepsilon_{\mathsf{acc}} > 0$*, with*

$$K \asymp \frac{\log(1/\varepsilon_{\mathsf{acc}})}{1-\lambda},$$

we have

$$\mathsf{TV}(p_{\hat{x}_K}, \pi_\eta) \lesssim \varepsilon_{\mathsf{acc}} \sqrt{\chi^2(p_{\hat{x}_1} \,\|\, \pi_\eta)} + \frac{1}{1-\lambda}\big(\varepsilon_{\mathsf{DDS},\mathcal{M}} + \varepsilon_{\mathsf{PCS},\mathcal{M}}\big) \log\left(\frac{1}{\varepsilon_{\mathsf{acc}}}\right). \tag{6.6}$$

Proof sketch. The proof is the same perturbative argument as in the Euclidean case. Let $\mathcal{K}_\eta$ denote the exact one-step kernel of manifold DPnP, and let $\widetilde{\mathcal{K}}_\eta$ denote the inexact kernel implemented in practice. By assumption, their one-step discrepancy is bounded uniformly in total variation by

$$\sup_{x \in \mathcal{M}} \mathsf{TV}\big(\widetilde{P}_\eta(x,\cdot), P_\eta(x,\cdot)\big) \le \varepsilon_{\mathsf{DDS},\mathcal{M}} + \varepsilon_{\mathsf{PCS},\mathcal{M}}.$$

Since $\mathcal{K}_\eta$ has invariant distribution π_η and $L^2(\pi_\eta)$ contraction factor λ, a standard stability bound for perturbed geometrically ergodic Markov chains yields

$$\mathsf{TV}(p_{\hat{x}_K}, \pi_\eta) \lesssim \lambda^K \sqrt{\chi^2(p_{\hat{x}_1} \,\|\, \pi_\eta)} + K\big(\varepsilon_{\mathsf{DDS},\mathcal{M}} + \varepsilon_{\mathsf{PCS},\mathcal{M}}\big).$$

Choosing

$$K \asymp \frac{\log(1/\varepsilon_{\mathsf{acc}})}{1-\lambda}$$

gives (6.6). □

6.4 Numerical experiments

6.4.1 Synthetic data

We demonstrate the correctness of our algorithm on a simple model, where the relevant score functions can be computed analytically. We take the state space to be the flat product torus $\mathbb{T}^d$, represented by angle coordinates $\theta \in [-\pi, \pi)^d$. We consider an inverse problem

$$\cos(\theta_j) = 0.1 + \lambda \xi_j, \quad \xi_j \sim \mathcal{N}(0,1), \quad j = 1, \cdots, d-1.$$

Intuitively, these observations reveal only part of the Euclidean embedding coordinates of a point on $\mathbb{T}^d$. Thus, although the problem is expressed in angular coordinates, it corresponds to a simple linear inverse problem in the ambient Euclidean representation. The target posterior distribution is then of the form

$$\pi(\theta) \propto p(\theta) \exp\Big\{-\frac{1}{2} Q(\theta)\Big\}, \qquad Q(\theta) = \lambda^{-1} \sum_{j=1}^{d-1} \big(\cos(\theta_j) - 0.1\big)^2.$$

In the noiseless limit $\lambda \to 0$, the target is supported on the constraint set

$$\theta_j \in \{\pm\alpha\}, \qquad \alpha = \arccos(0.1), \qquad j = 1, \dots, d-1,$$

while the last coordinate θ_d remains unconstrained. Thus the target is supported on 2^{d-1} one-dimensional branches of $\mathbb{T}^d$. The prior $p^\star$ is chosen to be a wrapped Gaussian mixture. More precisely, we take

$$p^\star(\theta) = \sum_{\ell=1}^{L} w_\ell \, \varphi_\sigma^{\mathbb{T}^d}(\theta - \mu_\ell), \qquad \sigma = 0.35,$$

where $w_\ell \geq 0$ and $\sum_{\ell=1}^{L} w_\ell = 1$. Here $\varphi_\sigma^{\mathbb{T}^d}$ denotes the wrapped isotropic Gaussian density on $\mathbb{T}^d$, defined by

$$\varphi_\sigma^{\mathbb{T}^d}(\theta - \mu) = \sum_{m \in \mathbb{Z}^d} \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left\{-\frac{\|\theta - \mu + 2\pi m\|^2}{2\sigma^2}\right\}, \qquad \theta, \mu \in [-\pi, \pi)^d.$$

The mixture is designed to be branch-rich but not already constrained. Namely, for each sign pattern $s \in \{\pm 1\}^{d-1}$, we include one or more components whose means have the form

$$\mu_\ell = \big(s\alpha + \delta_\ell,\, z_\ell\big), \qquad \alpha = \arccos(0.1),$$

where $z_\ell \in [-\pi, \pi)$ specifies the free coordinate and $\delta_\ell \in \mathbb{R}^{d-1}$ is a nonzero offset. Thus these components place mass near each constraint branch $\theta_{1:d-1} = s\alpha$, but not exactly on it. In addition, we include a few off-branch components with means

$$\mu_\ell = (r_\ell, z_\ell), \qquad r_\ell \not\approx s\alpha \quad \text{for all } s \in \{\pm 1\}^{d-1},$$

so that the prior has nontrivial mass away from the constraint set. Such prior makes the projected law of raw prior samples nontrivially different from the constrained posterior and therefore provides an informative test of the algorithm.

Starting from the uniform distribution on $\mathbb{T}^d$, we run $K = 60$ manifold DPnP iterations with a geometrically decreasing noise schedule, with

$$\eta_0 = 1.8, \quad \eta_K = 0.03.$$

In the $\lambda \to 0$ limit, both p_η and q_η can be calculated exactly (up to numerical error) using the finite wrapped-Gaussian mixture representation of p. We run the experiment for $d = 2, 3, 4$.

Evaluation metric. Because the limiting target is singular with respect to the ambient volume measure on $\mathbb{T}^d$, the literal total variation distance between the continuous iterates and the limiting constrained target is always equal to one. We therefore report two complementary diagnostics. The first is a projected total variation distance: each constrained coordinate is projected to its nearest branch in $\{\pm\alpha\}$, and we compare the empirical law of the resulting branch label together with the free coordinate θ_d against the exact projected target distribution. The second diagnostic is the constraint RMSE, defined as the root-mean-square circular distance of the constrained coordinates to the nearest branch. The former measures whether the algorithm has the correct relative mass across branches and along the free coordinate, while the latter measures whether the samples concentrate near the constraint set.

The results are shown in Figure 6.1. Across all three dimensions, both the projected TV distance and the constraint RMSE decrease rapidly and converge to nearly zero as the number of iterations increases. This confirms that manifold DPnP correctly targets the desired posterior distribution in this analytically tractable setting.

6.4.2 Real-world data

We evaluate the performance of our algorithm on the earthquake dataset [87]. The dataset contains the date, time and location of over 5700 significant earthquakes

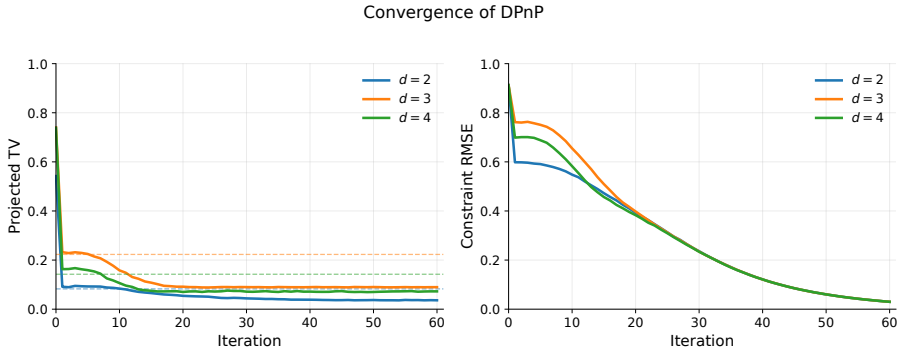

Figure 6.1: TV distance between the output distribution of manifold DPnP and the true posterior distribution.

from 2150 BC to the present. Per standard practice, we divide the dataset into training, validation, and test set, where the training set is used to train the score neural network, the validation set is used to regularize the training process, and only the test set is used to evaluate the performance of our algorithm. We visualize the dataset below and our data split in Figure 6.2.

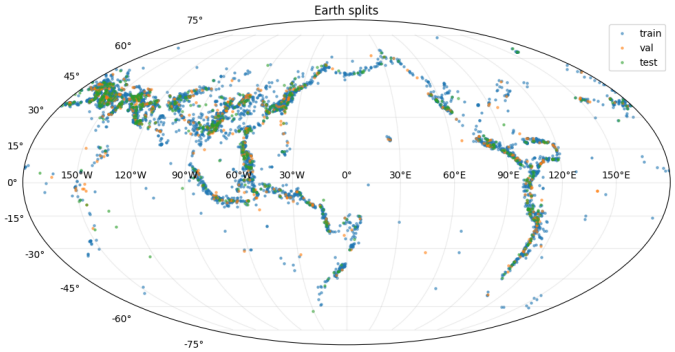

Figure 6.2: Illustration of the earthquake dataset.

Score function. The score function is trained through an implicit score matching precedure as in [27], on the training dataset specificed above.

Inverse problem. We consider the following inverse problem on the sphere. An earthquake occurs at an unknown location $x \in \mathbb{S}^2$, and a sensor located at $y \in \mathbb{S}^2$ records a scalar signal whose strength is maximal when $y = x$ and decays with the distance from the source.

A natural model is to let the noiseless signal be

$$I(x, y) = \exp\big(\beta\langle x, y\rangle - \beta\big),$$

where $\beta > 0$ is a fixed parameter controlling the decay rate. Then $I(x, y) = 1$ when $x = y$, and $m(x, y)$ decreases smoothly as the distance increases.

Given a measurement $u \in \mathbb{R}$, we model the observation noise by

$$U \mid x, y \sim \mathcal{N}\big(I(x, y), \sigma^2\big),$$

so that the likelihood is

$$p(u \mid x, y) \propto \exp\left(-\frac{(u - I(x, y))^2}{2\sigma^2}\right).$$

Suppose we have collected two measurements (y_1, u_1) and (y_2, u_2), and wish to recover the earthquake location x. Even in the noiseless case $\sigma = 0$, elementary geometry shows that this inverse problem has multiple solutions in general.

We therefore apply manifold DPnP to this problem, with the expectation that the prior learned by the diffusion model provides the regularization needed to select the correct solution among the geometrically feasible candidates. We compare manifold DPnP against a baseline that does not use any prior information on the data distribution. Formally, this baseline samples from the posterior

$$q(x) \propto p(u_1 \mid x, y_1)\, p(u_2 \mid x, y_2).$$

Figure 6.3 reports the root-mean-square error of the reconstruction produced by manifold DPnP, together with that of samples drawn from q. To reduce variance, each algorithm is run $N = 20$ times and the final output is taken to be the spherical average over all runs. The results show that manifold DPnP effectively exploits the learned prior and substantially improves recovery of the true source location.

6.5 Discussion

This chapter shows that DPnP admits an intrinsic extension to Riemannian manifolds. The key step is to transit from the direct proximal interpretation based on the

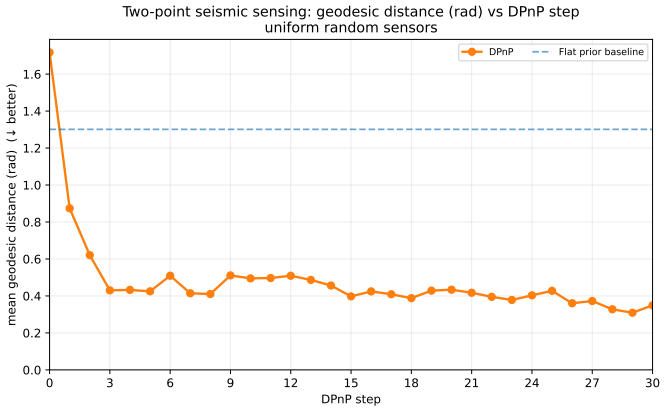

Figure 6.3: Performance of manifold DPnP on the earthquake dataset.

Euclidean quadratic penalty to a reinterpretation of each substep through heat flow. Once this is done, the manifold generalization becomes natural: Euclidean Brownian motion is replaced by Riemannian Brownian motion, Euclidean score fields are replaced by intrinsic score fields of heat-evolved densities, and the resulting algorithm can be implemented through geodesic random walk. In this sense, the heat-flow formulation is not merely a technical reformulation, but the conceptual mechanism that makes manifold DPnP possible.

A main message is that constrained inverse problems on manifolds can be handled without introducing extrinsic embeddings or ad hoc geometric surrogates. This is important both conceptually and computationally. Extrinsic embeddings generally depend on arbitrary modeling choices and may distort the geometry relevant to the inverse problem, while geodesic-distance-based proximal formulations are often difficult to compute and differentiate globally. By contrast, the present framework is fully intrinsic and aligns directly with the score-based Riemannian diffusion models developed earlier in the thesis.

The numerical experiments provide evidence that this intrinsic viewpoint is practically meaningful. On synthetic examples, manifold DPnP accurately recovers the target posterior distribution. On the earthquake localization task, it substantially improves reconstruction error relative to a baseline that uses only the measurement model and has no prior knowledge. Although the example is simple, it illustrates the intended use case clearly: when the forward model alone leaves multiple plausible solutions, the diffusion prior helps identify the one aligned with the data distribution

on the manifold.

Several directions remain open.

- Many heuristic solvers in the Euclidean setting have simpler forms while still exhibiting strong empirical performance. It would be interesting to extend such methods to the manifold setting and compare their behavior, efficiency, and robustness with those of manifold DPnP.
- The current framework is developed for compact manifolds. Extending manifold DPnP to noncompact manifolds would be highly desirable, but would require new control of heat kernels, reverse diffusions, and long-time stability in the absence of compactness.
- Deterministic solvers such as DDIM have enabled substantial acceleration for diffusion models in Euclidean space. Designing analogous deterministic solvers in the manifold setting is an interesting direction for future work.

Chapter 7

Conclusion

This thesis has developed a mathematical and algorithmic framework for using diffusion models as priors for inverse problems. The central perspective is that diffusion models are not only powerful generative models, but also effective tools for incorporating prior information into reconstruction algorithms in a principled plug-and-play manner. This viewpoint leads to a flexible class of solvers that combine measurement fidelity with learned prior from data, without requiring the prior to be expressible by an explicit handcrafted regularizer.

Beyond this algorithmic perspective, the thesis has established theoretical evidence for the effectiveness of such methods. The analysis shows that diffusion-based solvers are capable of efficiently integrating two complementary sources of information: the measurement, which enforces data consistency, and the diffusion prior, which regularizes the problem through learned statistical structure. In particular, the results demonstrate that this interaction is not merely heuristic, but can be understood rigorously in settings where provable convergence and recovery guarantees are available. These results help explain why diffusion-based inverse problem solvers can succeed in practice.

The thesis also extends this framework to constrained inverse problems on manifolds. In many applications, the unknown signal is constrained on a lower-dimensional manifold. By developing a manifold counterpart of diffusion plug-and-play methods, the thesis shows that diffusion-based priors can be adapted to such settings in an intrinsic way. Moreover, this extension retains efficient theoretical guarantees, indicating that the benefits of diffusion-based regularization continue to

hold on manifolds.

Taken together, these results support a broad conclusion: diffusion models provide a principled, versatile, and theoretically grounded framework for inverse problems. They offer a powerful mechanism for encoding prior information, admit rigorous analyses that clarify their effectiveness, and extend naturally to constrained and manifold-valued settings. These findings contribute to a growing understanding of diffusion models not only as generative tools, but also as a foundation for efficient and robust computational methods in statistical inference.

Several directions remain open for future work. On the theoretical side, it is important to extend the current guarantees to broader classes of inverse problems. On the methodological side, further understanding is needed for the design of faster, more stable, and more adaptive solvers. It is hoped that the ideas developed in this thesis provide a useful step toward a unified mathematical theory of diffusion-based inference.

Appendix A

Proofs for Chapter 3

A.1 Score functions of diffusion denoising samplers

A.1.1 Proof of Lemma 1

Proof. The marginal distribution (3.5) of the heat flow can be written as

$$Y_t \overset{(\mathrm{d})}{=} Y_0 + \sqrt{t}\varepsilon, \quad Y_0 \sim p^\star,\ \varepsilon \sim \mathcal{N}(0, I_d). \tag{A.1}$$

Comparing (2.6) and (A.1), it is not hard to check that

$$Y_t \overset{(\mathrm{d})}{=} \sqrt{1+t}\, X_{\frac{1}{2}\log(1+t)}.$$

Denote $\iota = \frac{1}{2}\log(1+t)$ as a short-hand. We have

$$p_{Y_t}(x) = p_{\sqrt{1+t}X_\iota}(x) \propto p_{X_\iota}\left(\frac{1}{\sqrt{1+t}}x\right).$$

Therefore it follows that

$$\nabla \log p_{Y_t}(x) = \nabla_x \log p_{X_\iota}\left(\frac{1}{\sqrt{1+t}}x\right) = \frac{1}{\sqrt{1+t}} s_\iota\left(\frac{1}{\sqrt{1+t}}x\right),$$

where we used the definition $s_\iota = \nabla \log p_{X_\iota}$. Plugging the definition $\iota = \frac{1}{2}\log(1+t)$ into the above equation yields the desired result. □

A.1.2 Proof of Lemma 2

Proof. We first compute the probability density function of z. Recall that $z = w - x_{\mathsf{noisy}}$, thus applying Bayes rule yields

$$\begin{aligned} p_z(x) = p_w(x + x_{\mathsf{noisy}}) &= p^\star(x^\star = x + x_{\mathsf{noisy}} | x^\star + \xi = x_{\mathsf{noisy}}) \\ &= \frac{p^\star(x + x_{\mathsf{noisy}}) p_\xi(-x)}{p_{x^\star + \xi}(x_{\mathsf{noisy}})} \\ &\propto p^\star(x + x_{\mathsf{noisy}}) p_\xi(-x), \end{aligned}$$

where $\xi \sim \mathcal{N}(0, \eta^2 I_d)$. It is straightforward to compute

$$p_\xi(-x) = \frac{1}{(2\pi)^{d/2}\eta^d} \mathrm{e}^{-\frac{1}{2\eta^2}\|x\|^2},$$

therefore

$$p_z(x) \propto p^\star(x + x_{\mathsf{noisy}}) \mathrm{e}^{-\frac{1}{2\eta^2}\|x\|^2}. \tag{A.2}$$

We proceed to compute the probability density function of Z_t. According to (3.10), it follows that

$$\begin{aligned} p_{Z_t}(x) &= p_{\mathrm{e}^{-t} z} * p_{\sqrt{1-\mathrm{e}^{-2t}}\varepsilon}(x) \\ &= \int p_{\mathrm{e}^{-t} z}(x') \, p_{\sqrt{1-\mathrm{e}^{-2t}}\varepsilon}(x - x') \, \mathrm{d}x' \\ &\propto \int p_z(\mathrm{e}^t x') \exp\left(-\frac{\|x - x'\|^2}{2(1 - \mathrm{e}^{-2t})} \right) \mathrm{d}x' \\ &\propto \int p^\star(x_{\mathsf{noisy}} + \mathrm{e}^t x') \exp\left(-\frac{\|\mathrm{e}^t x'\|^2}{2\eta^2} \right) \exp\left(-\frac{\|x - x'\|^2}{2(1 - \mathrm{e}^{-2t})} \right) \mathrm{d}x' \\ &\propto \int p^\star(x') \exp\left(-\frac{\|x' - x_{\mathsf{noisy}}\|^2}{2\eta^2} \right) \exp\left(-\frac{\|x - \mathrm{e}^{-t}(x' - x_{\mathsf{noisy}})\|^2}{2(1 - \mathrm{e}^{-2t})} \right) \mathrm{d}x'. \end{aligned} \tag{A.3}$$

where $*$ denotes convolution, the penultimate line follows from (A.2) and the last line follow from the change of variable $x' \mapsto \mathrm{e}^{-t}(x' - x_{\mathsf{noisy}})$. One may exercise some

brute force to verify that

$$\begin{aligned}
&\exp\left(-\frac{1}{2\eta^2}\|x'-x_{\mathsf{noisy}}\|^2\right)\exp\left(-\frac{1}{2(1-\mathrm{e}^{-2t})}\|x-\mathrm{e}^{-t}(x'-x_{\mathsf{noisy}})\|^2\right)\\
&=\exp\left(-\frac{\mathrm{e}^{2t}\|x\|^2}{2(\eta^2+\mathrm{e}^{2t}-1)}\right)\exp\left(-\frac{1}{2(1-\mathrm{e}^{-2\tilde{t}})}\left\|\mathrm{e}^{-\tilde{t}}x_{\mathsf{noisy}}+\frac{\mathrm{e}^{t-\tilde{t}}\eta^2 x}{\eta^2+\mathrm{e}^{2t}-1}-\mathrm{e}^{-\tilde{t}}x'\right\|^2\right)\\
&\propto\exp\left(-\frac{\mathrm{e}^{2t}\|x\|^2}{2(\eta^2+\mathrm{e}^{2t}-1)}\right)p_{\sqrt{1-\mathrm{e}^{-2\tilde{t}}}\varepsilon}\left(\mathrm{e}^{-\tilde{t}}x_{\mathsf{noisy}}+\frac{\mathrm{e}^{t-\tilde{t}}\eta^2 x}{\eta^2+\mathrm{e}^{2t}-1}-\mathrm{e}^{-\tilde{t}}x'\right),
\end{aligned}$$

where $\tilde{t}$ is as defined in (3.13). Define

$$D_t \coloneqq \eta^2+\mathrm{e}^{2t}-1, \qquad m_t(x)\coloneqq \mathrm{e}^{-\tilde{t}}x_{\mathsf{noisy}}+\frac{\mathrm{e}^{t-\tilde{t}}\eta^2}{D_t}x.$$

Plug the above displays back into (A.3), we see

$$\begin{aligned}
p_{Z_t}(x)&\propto\exp\left(-\frac{\mathrm{e}^{2t}\|x\|^2}{2D_t}\right)\int p^\star(x')\,p_{\sqrt{1-\mathrm{e}^{-2\tilde{t}}}\varepsilon}\left(m_t(x)-\mathrm{e}^{-\tilde{t}}x'\right)\mathrm{d}x'\\
&\propto\exp\left(-\frac{\mathrm{e}^{2t}\|x\|^2}{2D_t}\right)\int p^\star(\mathrm{e}^{\tilde{t}}x')\,p_{\sqrt{1-\mathrm{e}^{-2\tilde{t}}}\varepsilon}\left(m_t(x)-x'\right)\mathrm{d}x'\\
&\propto\exp\left(-\frac{\mathrm{e}^{2t}\|x\|^2}{2D_t}\right)\left(p_{\mathrm{e}^{-t}x_0}*p_{\sqrt{1-\mathrm{e}^{-2\tilde{t}}}\varepsilon}\right)\left(m_t(x)\right)\\
&\propto\exp\left(-\frac{\mathrm{e}^{2t}\|x\|^2}{2D_t}\right)p_{X_{\tilde{t}}}\left(m_t(x)\right).
\end{aligned}$$

where the second line applies the change of variable $x'\mapsto \mathrm{e}^{\tilde{t}}x'$ in the integral, the penultimate line follows from $p_{\mathrm{e}^{-\tilde{t}}x_0}(x')\propto p^\star(\mathrm{e}^{\tilde{t}}x')$ (since $x_0\sim p^\star$), and the last line follows from $X_{\tilde{t}}\overset{(\mathrm{d})}{=}\mathrm{e}^{-\tilde{t}}x_0+\sqrt{1-\mathrm{e}^{-2\tilde{t}}}\varepsilon$.

Finally, from the above formula, we obtain

$$\begin{aligned}
\nabla\log p_{Z_t}(x)&=\nabla_x\left(-\frac{\mathrm{e}^{2t}\|x\|^2}{2(\eta^2+\mathrm{e}^{2t}-1)}\right)+\nabla_x\log p_{X_{\tilde{t}}}\left(\mathrm{e}^{-\tilde{t}}x_{\mathsf{noisy}}+\frac{\mathrm{e}^{t-\tilde{t}}\eta^2 x}{\eta^2+\mathrm{e}^{2t}-1}\right)\\
&=-\frac{\mathrm{e}^{2t}x}{\eta^2+\mathrm{e}^{2t}-1}+\frac{\mathrm{e}^{t-\tilde{t}}\eta^2}{\eta^2+\mathrm{e}^{2t}-1}s^{\mathsf{cont}}\left(\tilde{t},\,\mathrm{e}^{-\tilde{t}}x_{\mathsf{noisy}}+\frac{\mathrm{e}^{t-\tilde{t}}\eta^2 x}{\eta^2+\mathrm{e}^{2t}-1}\right),
\end{aligned}$$

where we used the definition $s_{\tilde{t}}=\nabla\log p_{X_{\tilde{t}}}$. □

A.2 Discretization with the exponential integrator

A.2.1 General form of the exponential integrator

Consider a SDE of the form:

$$\mathrm{d}M_\tau = \big(v(\tau)M_\tau + f(\tau, M_\tau)\big)\mathrm{d}\tau + \sqrt{\beta}\mathrm{d}B_\tau, \quad \tau \in [0, \tau_\infty], \quad M_0 \sim p_{M_0},$$

where $v : [0, \tau_\infty] \to \mathbb{R}$, $f : [0, \tau_\infty] \times \mathbb{R}^d \to \mathbb{R}^d$ are deterministic functions, and $\beta > 0$ is a constant. Given discretization time points $0 = \tau_0 \le \tau_1 \le \cdots \le \tau_k \le \tau_\infty$, a naïve way to discretize the SDE is

$$M_{\tau_{i+1}} - M_{\tau_i} \approx \big(v(\tau_i)M_{\tau_i} + f(\tau_i, M_{\tau_i})\big)(\tau_{i+1} - \tau_i) + \sqrt{\beta}\sqrt{\tau_{i+1} - \tau_i}\varepsilon_i, \quad i = 0, 1, \cdots, k-1,$$

where $\varepsilon_i \sim \mathcal{N}(0, I_d)$ is a standard d-dimensional Gaussian random vector which is independent of M_{τ_i}. Although this approach is straightforward, it has the drawback that the linear term $v(\tau)M_\tau$ is discretized rather crude. For example, for the OU process where $v \equiv -1$, $f \equiv 0$, $\beta = 2$, the SDE can be solved analytically as in (2.6), while the above approach still has a discretization error.

A more accurate discretization, known to significantly improve the quality of score-based generative models, is given by the *exponential integrator* [131], which preserves the linear term and discretizes the SDE to

$$\mathrm{d}\hat{M}_\tau = \big(v(\tau)\hat{M}_\tau + f(\tau_i, \hat{M}_{\tau_i})\big)\mathrm{d}\tau + \sqrt{\beta}\mathrm{d}B_\tau, \quad \tau \in [\tau_i, \tau_{i+1}], \quad i = 0, 1, \cdots, k,$$

with initialization $\hat{M}_0 \sim p_{M_0}$. On each time interval $[\tau_i, \tau_{i+1}]$, this is simply a linear SDE, which can be explicitly solved by

$$\hat{M}_\tau \overset{(\mathrm{d})}{=} \mathrm{e}^{V(\tau) - V(\tau_i)}\hat{M}_{\tau_i} + \left(\int_{\tau_i}^{\tau} \mathrm{e}^{V(\tau) - V(\tilde{\tau})}\mathrm{d}\tilde{\tau}\right) f(\tau_i, \hat{M}_{\tau_i}) + \sqrt{\beta}\left(\int_{\tau_i}^{\tau} \mathrm{e}^{2(V(\tau) - V(\tilde{\tau}))}\mathrm{d}\tilde{\tau}\right)^{1/2} \varepsilon_i,$$

where V is the antiderivative of v:

$$V(\tau) = \int_0^\tau v(\tilde{\tau})\mathrm{d}\tilde{\tau}.$$

Taking $\tau = \tau_{i+1}$, we obtain

$$\hat{M}_{\tau_{i+1}} \overset{(\mathrm{d})}{=} \mathrm{e}^{V(\tau_{i+1})-V(\tau_i)}\hat{M}_{\tau_i} + \left(\int_{\tau_i}^{\tau_{i+1}} \mathrm{e}^{V(\tau_{i+1})-V(\tilde{\tau})}\mathrm{d}\tilde{\tau}\right) f(\tau_i, \hat{M}_{\tau_i}) + \sqrt{\beta}\mathrm{e}^{V(\tau_{i+1})}\left(\int_{\tau_i}^{\tau_{i+1}} \mathrm{e}^{2(V(\tau_{i+1})-V(\tilde{\tau}))}\mathrm{d}\tilde{\tau}\right)^{1/2} \varepsilon_i, \tag{A.4}$$

which provides an iterative formula to compute $\hat{M}_{\tau_{i+1}}$.

A.2.2 Discretization of DDS-DDPM

Recall the definition of ε_t in (2.8). Plug the expression of $\nabla \log p_{Y_t}$ in Lemma 1 into (3.7), and use the notation $\tau = \eta^2 - t$, we obtain, for $t \in [0, \eta^2]$, that

$$\begin{aligned}\mathrm{d}Y_\tau^{\mathsf{rev}} &= \frac{1}{\sqrt{1+\tau}} s_{\frac{1}{2}\log(1+\tau)}\left(\frac{Y_\tau^{\mathsf{rev}}}{\sqrt{1+\tau}}\right)\mathrm{d}t + \mathrm{d}\tilde{B}_t \\ &= -\frac{1}{\sqrt{\tau}} \varepsilon_{\frac{1}{2}\log(1+\tau)}\left(\frac{Y_\tau^{\mathsf{rev}}}{\sqrt{1+\tau}}\right)\mathrm{d}t + \mathrm{d}\tilde{B}_t.\end{aligned}$$

Choosing discretization time points. To discretize this SDE, we first choose the discretization time points. Recalling (2.4) and Lemma 1, it is most reasonable to discretize at those time points $0 \le \tau_0 \le \cdots \le \tau_{T'} \le \eta^2$ which satisfy

$$\frac{1}{2}\log(1+\tau_\ell) = \frac{1}{2}\log\frac{1}{\bar{\alpha}_\ell}, \quad 0 \le \ell \le T'.$$

This solves to

$$\tau_\ell = \bar{\alpha}_\ell^{-1} - 1. \tag{A.5}$$

The requirement that $\tau_\ell \le \eta^2$ translates to $\bar{\alpha}_\ell \ge \frac{1}{1+\eta^2}$, which yields the following choice of T':

$$T' := \max\left\{t : 0 \le \ell \le T,\, \bar{\alpha}_\ell > \frac{1}{\eta^2+1}\right\}. \tag{A.6}$$

Applying the exponential integrator. Now we apply the exponential integrator to discretize the SDE on each time interval $\tau \in [t_{\ell-1}, t_\ell]$, $\ell = 1, \cdots, T'$ as follows:

$$\begin{aligned}
\mathrm{d}\widehat{Y}_\tau^{\mathsf{rev}} &= -\frac{1}{\sqrt{\tau}} \varepsilon_{\frac{1}{2}\log(1+\tau_\ell)} \left(\frac{\widehat{Y}_{\tau_\ell}^{\mathsf{rev}}}{\sqrt{1+\tau_\ell}} \right) \mathrm{d}t + \mathrm{d}\tilde{B}_t, \\
&= -\frac{1}{\sqrt{\tau}} \varepsilon_{t_\ell} \left(\sqrt{\bar{\alpha}_\ell} \widehat{Y}_{\tau_\ell}^{\mathsf{rev}} \right) \mathrm{d}t + \mathrm{d}\tilde{B}_t.
\end{aligned}$$

The SDE can be integrated directly on $\tau \in [t_{\ell-1}, t_\ell]$ (see also (A.4), with $v \equiv 0$), yielding

$$\begin{aligned}
\widehat{Y}_{\tau_{\ell-1}}^{\mathsf{rev}} &= \widehat{Y}_{\tau_t}^{\mathsf{rev}} - 2(\sqrt{\tau_\ell} - \sqrt{\tau_{\ell-1}}) \cdot \varepsilon_{t_\ell} \left(\sqrt{\bar{\alpha}_\ell} \widehat{Y}_{\tau_\ell}^{\mathsf{rev}} \right) + \int_{\eta^2 - t_\ell}^{\eta^2 - t_{\ell-1}} d\tilde{B}_t \mathrm{d}t \\
&\overset{(\mathrm{d})}{=} \widehat{Y}_{\tau_t}^{\mathsf{rev}} - 2(\sqrt{\tau_\ell} - \sqrt{\tau_{\ell-1}}) \cdot \varepsilon_{t_\ell} \left(\sqrt{\bar{\alpha}_\ell} \widehat{Y}_{\tau_\ell}^{\mathsf{rev}} \right) + \sqrt{\tau_\ell - \tau_{\ell-1}}\, w_\ell,
\end{aligned}$$

where $w_\ell \sim \mathcal{N}(0, I_d)$ is independent of $\widehat{Y}_{\tau_t}^{\mathsf{rev}}$. Set $\hat{x}_\ell = \widehat{Y}_{\tau_\ell}^{\mathsf{rev}}$, we obtain

$$\hat{x}_{\ell-1} \overset{(\mathrm{d})}{=} \hat{x}_\ell - 2(\sqrt{t_\ell} - \sqrt{t_{\ell-1}}) \cdot \varepsilon_{t_\ell} \left(\sqrt{\bar{\alpha}_\ell} \hat{x}_\ell \right) + \sqrt{t_\ell - t_{\ell-1}}\, w_\ell, \quad w_\ell \sim \mathcal{N}(0, I_d), \quad \text{(A.7)}$$

which is exactly the update equation in Algorithm 1, except that $\varepsilon_{t_\ell}^\star$ is replaced by the noise estimate $\hat{\varepsilon}_{t_\ell}$.

A.2.3 Discretization of DDS-DDIM

Plug in the expression of the score function of Z_t in Lemma 2 into the probability flow ODE (3.11), we obtain

$$\begin{aligned}
\mathrm{d}Z_\tau^{\mathsf{rev}} &= \frac{\eta^2 - 1}{\eta^2 + \mathrm{e}^{2\tau} - 1} Z_\tau^{\mathsf{rev}} \mathrm{d}t + \frac{\mathrm{e}^{\tau - \tilde{t}(\tau)} \eta^2}{\eta^2 + \mathrm{e}^{2\tau} - 1} s_{\tilde{t}(\tau)} \left(\mathrm{e}^{-\tilde{t}(\tau)} x_{\mathsf{noisy}} + \frac{\mathrm{e}^{\tau - \tilde{t}(\tau)} \eta^2 x}{\eta^2 + \mathrm{e}^{2\tau} - 1} \right) \mathrm{d}t \\
&= \frac{\eta^2 - 1}{\eta^2 + \mathrm{e}^{2\tau} - 1} Z_\tau^{\mathsf{rev}} \mathrm{d}t - \frac{\mathrm{e}^{2\tau}}{\mathrm{e}^{2\tau} - 1} \varepsilon_{\tilde{t}(\tau)} \left(\mathrm{e}^{-\tilde{t}(\tau)} x_{\mathsf{noisy}} + \frac{\mathrm{e}^{\tau - \tilde{t}(\tau)} \eta^2 x}{\eta^2 + \mathrm{e}^{2\tau} - 1} \right) \mathrm{d}t,
\end{aligned}$$

where the second line used the definition (3.13).

Choosing discretization time points. Similar to the derivation in Appendix A.2.2, we discretize at time points $0 = \tau_0 \le \tau_1 \le \cdots \le \tau_{T'} \le \eta^2$, which obey

$$\tilde{t}(\tau_\ell) = t_\ell = \frac{1}{2}\log\frac{1}{\bar{\alpha}_\ell}, \quad \ell = 0, 1, \ldots, T', \tag{A.8}$$

which solves to

$$\tau_\ell = \frac{1}{2}\log\frac{\eta^2 + \bar{\alpha}_\ell - 1}{(\eta^2+1)\bar{\alpha}_\ell - 1}. \tag{A.9}$$

To make this well-defined, we require

$$\frac{\eta^2 + \bar{\alpha}_\ell - 1}{(\eta^2+1)\bar{\alpha}_\ell - 1} > 0,$$

which is equivalent to

$$\bar{\alpha}_\ell > \frac{1}{1+\eta^2}.$$

This leads to the same choice of T' as in (A.6). We also set

$$t_\infty = \tau_{T'}.$$

It is convenient to introduce a notation for the corresponding discrete schedule of τ_t, denoted by

$$\bar{u}_\ell = \mathrm{e}^{-2\tau_\ell} = \frac{(\eta^2+1)\bar{\alpha}_\ell - 1}{\eta^2 + \bar{\alpha}_\ell - 1}, \quad t = 0, 1, \cdots, T'.$$

Applying the exponential integrator. Now we apply the exponential integrator, which discretizes the ODE on each time interval $\tau \in [t_{\ell-1}, t_\ell]$, $t = 1, \cdots, T'$, as

$$\begin{aligned}
\mathrm{d}\hat{Z}_\tau^{\mathsf{rev}} &= \frac{\eta^2 - 1}{\eta^2 + \mathrm{e}^{2\tau} - 1}\hat{Z}_\tau^{\mathsf{rev}}\mathrm{d}t - \frac{\mathrm{e}^{2\tau}}{\mathrm{e}^{2\tau}-1}\varepsilon_{t_\ell}\left(\mathrm{e}^{-\tilde{t}(\tau_\ell)}x_{\mathsf{noisy}} + \frac{\mathrm{e}^{\tau_\ell - \tilde{t}(\tau_\ell)}\eta^2\hat{Z}_{\tau_\ell}^{\mathsf{rev}}}{\eta^2 + \mathrm{e}^{2\tau_\ell} - 1}\right)\mathrm{d}t \\
&= \frac{\eta^2 - 1}{\eta^2 + \mathrm{e}^{2\tau} - 1}\hat{Z}_\tau^{\mathsf{rev}}\mathrm{d}t - \frac{\mathrm{e}^{2\tau}}{\mathrm{e}^{2\tau}-1}\varepsilon_{t_\ell}\left(\sqrt{\bar{\alpha}_\ell}x_{\mathsf{noisy}} + \frac{\sqrt{\bar{u}_\ell}\sqrt{\bar{\alpha}_\ell}\eta^2\hat{Z}_{\tau_\ell}^{\mathsf{rev}}}{(\eta^2-1)\bar{u}_\ell + 1}\right)\mathrm{d}t,
\end{aligned}$$

where the last line follows from dividing both the denominator and the numerator in the fraction inside ε_{t_ℓ} by $\mathrm{e}^{2\tau_\ell}$. This is a first-order linear ODE on $\tau \in [t_{\ell-1}, t_\ell]$, which

can be solved explicitly (cf. (A.4)) by

$$\hat{Z}_\tau^{\mathsf{rev}} = \frac{\sqrt{(\eta^2-1)\mathrm{e}^{-2\tau}+1}}{\sqrt{(\eta^2-1)\bar{u}_\ell+1}}\hat{Z}_{\tau_\ell}^{\mathsf{rev}} + \sqrt{(\eta^2-1)\mathrm{e}^{-2\tau}+1}\big(h(\eta,\mathrm{e}^{-2\tau})-h(\eta,\bar{u}_\ell)\big)\varepsilon_{t_\ell}(\Xi_\ell)\,,$$

for $\tau \in [t_{\ell-1}, t_\ell]$, where

$$h(\eta,u) \coloneqq -\arctan\frac{\eta}{\sqrt{u^{-1}-1}}, \tag{A.10}$$

$$\Xi_\ell \coloneqq \sqrt{\bar{\alpha}_\ell}x_{\mathsf{noisy}} + \frac{\sqrt{\bar{u}_\ell}\sqrt{\bar{\alpha}_\ell}\eta^2\hat{Z}_\tau^{\mathsf{rev}}}{(\eta^2-1)\bar{u}_\ell+1}. \tag{A.11}$$

Plug in $\tau = t_{\ell-1}$ in the above solution, and set $z_\ell = \hat{Z}_{\tau_\ell}^{\mathsf{rev}}$, we obtain

$$\begin{aligned} z_{\ell-1} &= \frac{\sqrt{(\eta^2-1)\bar{u}_{\ell-1}+1}}{\sqrt{(\eta^2-1)\bar{u}_\ell+1}}z_\ell + \sqrt{(\eta^2-1)\bar{u}_{\ell-1}+1}\big(h(\eta,\bar{u}_{\ell-1})-h(\eta,\bar{u}_\ell)\big)\varepsilon_{t_\ell}(\Xi_\ell)\,, \\ \Xi_\ell &\coloneqq \sqrt{\bar{\alpha}_\ell}x_{\mathsf{noisy}} + \frac{\sqrt{\bar{u}_\ell}\sqrt{\bar{\alpha}_\ell}\eta^2 z_\ell}{(\eta^2-1)\bar{u}_\ell+1}. \end{aligned} \tag{A.12}$$

The initialization, which should ideally be $z_{T'} = \hat{Z}_{t_\infty}^{\mathsf{rev}} \sim p_{Z_{t_\infty}}$, is approximated by $z_{T'} \sim \mathcal{N}(0, I_d)$. Recalling (A.11), this is exactly the update equation and the initialization in Algorithm 2, except that ε_{t_ℓ} is replaced by the noise estimate $\hat{\varepsilon}_{t_\ell}$.

A.2.4 Discretization of PCS

We first note that the Metropolis-adjustment step in PCS (cf. Algorithm 4) is standard following the classical form of MALA [98]. Therefore, we focus on explaining the Langevin step. Recall the continuous-time Langevin dynamics for sampling from the distribution $\exp(\mathcal{L}(\cdot;y) - \frac{1}{2\eta^2}\|\cdot - x\|^2)$:

$$\mathrm{d}Z_\tau = -\nabla_{Z_\tau}\mathcal{L}(Z_\tau;y)\mathrm{d}\tau + \frac{1}{\eta^2}(Z_\tau - x)\mathrm{d}\tau + \sqrt{2}\mathrm{d}B_\tau, \quad \tau \geq 0, \quad Z_0 \sim \mathcal{N}(0, I_d). \tag{A.13}$$

The classical form of MALA, as in [98], performs one step of a straightforward discretization of (A.13) as the Langevin step, as follows:

$$z_{n+\frac{1}{2}} \approx z_n - \gamma\nabla_{z_n}\mathcal{L}(z_n;y) + \frac{\gamma}{\eta^2}(z_n - x) + \sqrt{2\gamma}w_n, \quad w_n \sim \mathcal{N}(0, I_d).$$

In our setting, due to the presence of the linear drift term $\frac{1}{\eta^2}(Z_\tau - x)$, which can be quite large when η is small, we apply the exponential integrator instead. Set the discretization time points $\tau_n = n\gamma$, the exponential integrator reads as

$$\mathrm{d}Z_\tau = -\nabla_{Z_{n\gamma}} \mathcal{L}(Z_{n\gamma}; y)\mathrm{d}\tau + \frac{1}{\eta^2}(Z_\tau - x)\mathrm{d}\tau + \sqrt{2}\mathrm{d}B_\tau, \quad n\gamma \le \tau \le (n+1)\gamma.$$

Solve this linear SDE on $n\gamma \le \tau \le (n+1)\gamma$ directly (see also (A.4)) to obtain

$$Z_{(n+1)\gamma} \overset{(\mathrm{d})}{=} rZ_{n\gamma} + (1-r)x + \eta^2(1-r)\nabla_{Z_{n\gamma}} \mathcal{L}(Z_{n\gamma}; y) + \eta\sqrt{1-r^2}w_n, \quad w_n \sim \mathcal{N}(0, I_d),$$

where $r \coloneqq \mathrm{e}^{-\gamma/\eta^2}$. This is the same as the update equation for the Langevin step in PCS (cf. Algorithm 4).

A.3 Proof of main theorems

A.3.1 Proof of Theorem 1

Proof. The proof is based on two lemmas on the one-step transition kernel of DPnP and the asymptotic behavior of the transition kernel, which we will present soon. First, we set up some notation. Denote

$$p_\eta(x) \coloneqq p_{x^\star \sim p^\star, \varepsilon \sim \mathcal{N}(0, I_d)}(x^\star + \eta\varepsilon = x) = \frac{1}{(2\pi)^{d/2}\eta^d} \int p^\star(z) \mathrm{e}^{-\frac{1}{2\eta^2}\|x-z\|^2} \mathrm{d}z.$$

From the first equality, it is clear that $p_\eta \to p^\star$ when $\eta \to 0^+$. We will also use the notation q_η defined in (3.17), which we recall here:

$$q_\eta(x) \coloneqq \frac{1}{(2\pi)^{d/2}\eta^d} \int \mathrm{e}^{\mathcal{L}(z;y) - \frac{1}{2\eta^2}\|x-z\|^2} \mathrm{d}z.$$

In virtue of the Assumption 1, we know that q_η is finite for all $x \in \mathbb{R}^d$.

For convenience, we introduce a notation for application of transition kernels. For a probability distribution $p(x)$ and a probability transition kernel $K(x, x')$, denote by $p \circ K$ the probability distribution given by

$$p \circ K(x') = \int p(x) K(x, x') \mathrm{d}x.$$

The first lemma characterizes the one-step behavior of DPnP in terms of Markov transition kernels.

Lemma 11. *Under the settings of Lemma 4 and Lemma 3, the one-step transition kernel of* DPnP *with* $\eta_k = \eta$ *is given by:*

$$K_{\mathsf{DPnP},\eta}(x, x') = \left(\int \frac{q_0(z)}{p_\eta(z)} \mathrm{e}^{-\frac{1}{2\eta^2}\|z-x\|^2 - \frac{1}{2\eta^2}\|z-x'\|^2} \mathrm{d}z \right) \frac{p^\star(x')}{q_\eta(x)}.$$

In other words, if $\hat{x}_k$ *has distribution* $p_{\hat{x}_k}$*, then the distribution of* $\hat{x}_{k+1}$ *is*

$$p_{\hat{x}_{k+1}}(x') = p_{\hat{x}_k} \circ K_{\mathsf{DPnP},\eta}(x) = \int p_{\hat{x}_k}(x) K_{\mathsf{DPnP},\eta}(x, x') \mathrm{d}x.$$

The proof is postponed to Appendix A.3.3. The next lemma analyzes the ergodic properties of the Markov chain with transition kernel $K_{\mathsf{DPnP},\eta}$. These properties are known [11] but scattered in different literatures, so we will provide a brief proof to be self-contained.

Lemma 12. *The Markov transition kernel* $K_{\mathsf{DPnP},\eta}$ *has the following properties:*

(i) (Stationary distribution.) Let π_η *be the probability distribution defined by*

$$\pi_\eta(x) = c_\eta p^\star(x) q_\eta(x),$$

where $c_\eta > 0$ *is the normalization constant such that* $\int \pi_\eta(x) \mathrm{d}x = 1$*. Then* $K_{\mathsf{DPnP},\eta}$ *is reversible with stationary distribution* π_η*.*

(ii) (Convergence.) For any initial distribution p*, the distribution of the Markov chain with kernel* $K_{\mathsf{DPnP},\eta}$ *converges to* π_η*:*

$$\mathsf{TV}(p \circ K^{(n)}_{\mathsf{DPnP},\eta}, \pi_\eta) \to 0, \quad n \to \infty, \tag{A.14}$$

where $K^{(n)}_{\mathsf{DPnP},\eta}$ *is the* n*-step transition kernel of* $K_{\mathsf{DPnP},\eta}$*.*

The proof is postponed to Appendix A.3.4. We now show how to prove Theorem 1 with the above two lemmas. With the annealing schedule in Theorem 1, between steps $k_{l-1} \le k < k_l$, which consist of consecutive $(k_l - k_{l-1})$ steps, the transition

kernel of one-step of DPnP is $K_{\mathsf{DPnP},\varepsilon_l}$. As $(k_l - k_{l-1}) \to \infty$, Lemma 12 implies that

$$\mathsf{TV}(p_{\hat{x}_{k_l}}, \pi_{\varepsilon_l}) = \mathsf{TV}(p_{\hat{x}_{k_{l-1}}} \circ K_{\mathsf{DPnP},\varepsilon_l}^{(k_l - k_{l-1})}, \pi_{\varepsilon_l}) \to 0.$$

Under the assumption in Theorem 1 that $\varepsilon_l \to 0$, we let $l \to \infty$ to see $\lim_{l\to\infty} \pi_{\varepsilon_l} = c_0 p^\star(\cdot)\mathrm{e}^{\mathcal{L}(\cdot;y)} = p^\star(\cdot|y)$, thus $p_{\hat{x}_{k_l}} \to p^\star(\cdot|y)$, as claimed. □

A.3.2 Proof of Lemma 4

Proof. For DDS-DDPM, we note that under the continuous-time limit in Lemma 4, the discretization time points given by (A.5) verify

$$\tau_0^{\mathsf{rev}} = 0, \quad \sup_{0\le t\le T'-1} |\tau_t^{\mathsf{rev}} - \tau_{t+1}^{\mathsf{rev}}| \to 0, \quad \tau_{T'}^{\mathsf{rev}} \to \left(\frac{1}{1+\eta^2}\right)^{-1} - 1 = \eta^2, \quad T' \to \infty.$$

Therefore, these discretization time points $0 = \tau_0^{\mathsf{rev}} \le \cdots \le \tau_{T'}^{\mathsf{rev}} \le \eta^2$ form a partition of $[0, \eta^2]$, which becomes infinitely fine in the continuous-time limit. Thus the discretized integrator (A.7) converges to the solution of the SDE (3.7), which, as we have already argued in Appendix A.1, produces samples obeying the denoising posterior distribution $p^\star(\cdot|x_{\mathsf{noisy}})$, as claimed.

The proof for DDS-DDPM follows similarly, by observing that the discretization time points in (A.5) form an infinitely fine partition of $[0, \infty)$ in the continuous-time limit. □

A.3.3 Proof of Lemma 11

Proof. The proof is based on computing the transition kernel of the two subroutines. We claim that

(i) Sampling with probability density proportional to $\exp(\mathcal{L}(\cdot; y) - \frac{1}{2\eta^2}\|\cdot - x\|^2)$ is equivalent to applying the following Markov transition kernel

$$K_{\mathsf{PCS},\eta}(x, x') = \frac{1}{q_\eta(x)} \mathrm{e}^{\mathcal{L}(x';y) - \frac{1}{2\eta^2}\|x'-x\|^2}.$$

(ii) Sampling with probability $p^\star(x^\star \,|\, x^\star + \eta\varepsilon = x)$, where $\varepsilon \sim \mathcal{N}(0, I_d)$, is equivalent

to applying the following Markov transition kernel:

$$K_{\mathsf{DDS},\eta}(x,x') = \frac{1}{p_\eta(x)} p^\star(x') \mathrm{e}^{-\frac{1}{2\eta^2}\|x'-x\|^2}.$$

It is then clear that

$$\begin{aligned} K_{\mathsf{DPnP},\eta}(x,x') &= \int K_{\mathsf{PCS},\eta}(x,z) K_{\mathsf{DDS},\eta}(z,x') \mathrm{d}z \\ &= \left(\int \frac{q_0(z)}{p_\eta(z)} \mathrm{e}^{-\frac{1}{2\eta^2}\|z-x\|^2 - \frac{1}{2\eta^2}\|z-x'\|^2} \mathrm{d}z \right) \frac{p^\star(x')}{q_\eta(x)}, \end{aligned}$$

as desired. We now prove the above two claims. For (i), note that by (3.16), we know $K_{\mathsf{DDS},\eta}(x,\cdot) \propto p^\star(\cdot)\mathrm{e}^{-\frac{1}{2\eta^2}\|\cdot-x\|^2}$. Thus it suffices to compute the normalization constant, which is

$$\int p^\star(x')\mathrm{e}^{-\frac{1}{2\eta^2}\|x'-x\|^2}\mathrm{d}x' = p_\eta(x),$$

by the definition of p_η. Therefore

$$K_{\mathsf{DDS},\eta}(x,x') = \frac{1}{p_\eta(x)} p^\star(x') \mathrm{e}^{-\frac{1}{2\eta^2}\|x'-x\|^2},$$

as claimed. The proof of (ii) follows similarly. □

A.3.4 Proof of Lemma 12

Proof. We first introduce a fundamental lemma [93], which provides a simple method to bound the total variation between two distributions.

Lemma 13 (Data-processing inequality)**.** *Let p, q be two probability distributions, and K be a probability transition kernel. Then*

$$\mathsf{TV}(p \circ K, q \circ K) \le \mathsf{TV}(p, q).$$

We now prove the two items in Lemma 12 separately.

Proof of (i). We first show that π_η is well-defined, i.e., $\int p^\star(x) q_n \eta(x) \mathrm{d}x < \infty$. This

can be seen from Assumption 1, which implies $q_\eta(x) \lesssim \int \mathrm{e}^{-\frac{1}{2\eta^2}\|x-z\|^2}\mathrm{d}z \lesssim 1$, hence

$$\int p^\star(x) q_\eta(x)\mathrm{d}x \lesssim \int p^\star(x)\mathrm{d}x = 1.$$

To show that $K_{\mathsf{DPnP},\eta}$ is reversible with stationary distribution π_η, it suffices to verify

$$\pi_\eta(x) K_{\mathsf{DPnP},\eta}(x, x') = \pi_\eta(x') K_{\mathsf{DPnP},\eta}(x', x), \quad \forall x, x' \in \mathbb{R}^d.$$

However, it is easily checked that both sides are equal to

$$c_\eta \left(\int \frac{q_0(z)}{p_\eta(z)} \mathrm{e}^{-\frac{1}{2\eta^2}\|z-x\|^2 - \frac{1}{2\eta^2}\|z-x'\|^2} \mathrm{d}z \right) p^\star(x') p^\star(x).$$

Proof of (ii). We define an auxiliary Markov transition kernel $K_{\mathsf{aux},\eta} = K_{\mathsf{DDS},\eta} \circ K_{\mathsf{PCS},\eta}$. More explicitly,

$$\begin{aligned} K_{\mathsf{aux},\eta}(x, x') &= \int K_{\mathsf{DDS},\eta}(x, z) K_{\mathsf{PCS},\eta}(z, x')\mathrm{d}z \\ &= \left(\int \frac{p^\star(z)}{q_\eta(z)} \mathrm{e}^{-\frac{1}{2\eta^2}\|z-x\|^2 - \frac{1}{2\eta^2}\|z-x'\|^2} \mathrm{d}z \right) \frac{\mathrm{e}^{\mathcal{L}(x';y)}}{p_\eta(x)}. \end{aligned} \tag{A.15}$$

It is easy to see that

$$p \circ K^{(n)}_{\mathsf{DPnP},\eta} = p \circ K_{\mathsf{PCS},\eta} \circ K^{(n-1)}_{\mathsf{aux},\eta} \circ K_{\mathsf{DDS},\eta}. \tag{A.16}$$

Thus we are led to investigate the ergodic properties of $K_{\mathsf{aux},\eta}$. Similar to the proof of item (i) above, it is not hard to show that $K_{\mathsf{aux},\eta}$ is reversible with respect to the stationary distribution

$$\mu_\eta(x) \coloneqq c_\eta p_\eta(x) q_0(x) = c_\eta p_\eta(x) \mathrm{e}^{\mathcal{L}(x;y)}.$$

Moreover, one may check that

$$\pi_\eta = \mu_\eta \circ K_{\mathsf{DDS},\eta}. \tag{A.17}$$

It is apparent that $\mu(x') > 0$ and $K_{\mathsf{aux},\eta}(x, x')/\mu_\eta(x') > 0$ for all $x, x' \in \mathbb{R}^d$. By Tierney [117, Corollary 1], such a Markov transition kernel obeys, for any probability

distribution q, that

$$\mathsf{TV}(q \circ K_{\mathsf{aux},\eta}^{(n)},\ \mu_\eta) \to 0, \quad n \to \infty.$$

In view of (A.16) and (A.17), we set $q = p \circ K_{\mathsf{PCS},\eta}$ and invoke the data-processing inequality to obtain

$$\begin{aligned}
\mathsf{TV}(p \circ K_{\mathsf{DPnP},\eta}^{(n)},\ \pi_\eta) &= \mathsf{TV}(q \circ K_{\mathsf{aux},\eta}^{(n-1)} \circ K_{\mathsf{DDS},\eta},\ \mu_\eta \circ K_{\mathsf{DDS},\eta}) \\
&\leq \mathsf{TV}(q \circ K_{\mathsf{aux},\eta}^{(n-1)},\ \mu_\eta) \\
&\to 0,
\end{aligned}$$

as $n \to \infty$. This completes the proof. □

A.3.5 Proof of Theorem 2

Proof. Denote by $\tilde{K}_{\mathsf{PCS},\eta}$ and $\tilde{K}_{\mathsf{DDS},\eta}$ and the transition kernels for PCS and for DDS, respectively. Note that these may deviate from the transition kernels $K_{\mathsf{PCS},\eta}$ and $K_{\mathsf{DDS},\eta}$ defined for the idealized asymptotic setting in Appendix A.3. We have

$$\begin{aligned}
\mathsf{TV}(p_{\hat{x}_N}, \pi_\eta) &= \mathsf{TV}(p_{\hat{x}_{N-\frac{1}{2}}} \circ \tilde{K}_{\mathsf{DDS},\eta},\ \pi_\eta) \\
&\leq \mathsf{TV}(p_{\hat{x}_{N-\frac{1}{2}}} \circ K_{\mathsf{DDS},\eta},\ \pi_\eta) + \mathsf{TV}(p_{\hat{x}_{N-\frac{1}{2}}} \circ K_{\mathsf{DDS},\eta},\ p_{\hat{x}_{N-\frac{1}{2}}} \circ \tilde{K}_{\mathsf{DDS},\eta}) \\
&\leq \mathsf{TV}(p_{\hat{x}_{N-\frac{1}{2}}} \circ K_{\mathsf{DDS},\eta},\ \pi_\eta) + \varepsilon_{\mathsf{DDS}},
\end{aligned}$$

where the second line is triangle inequality, and the third line follows from the assumption in Theorem 2 that DDS has error at most $\varepsilon_{\mathsf{DDS}}$ in total variation, by taking the input of DDS to be $\hat{x}_{N-\frac{1}{2}}$.

Similarly, from $p_{\hat{x}_{N-\frac{1}{2}}} = p_{\hat{x}_{N-1}} \circ \tilde{K}_{\mathsf{PCS},\eta}$ and the assumption that PCS has error at most $\varepsilon_{\mathsf{PCS}}$ in total variation, we can show

$$\begin{aligned}
\mathsf{TV}(p_{\hat{x}_{N-\frac{1}{2}}} \circ K_{\mathsf{DDS},\eta},\ \pi_\eta) &\leq \mathsf{TV}(p_{\hat{x}_{N-1}} \circ K_{\mathsf{PCS},\eta} \circ K_{\mathsf{DDS},\eta},\ \pi_\eta) + \varepsilon_{\mathsf{PCS}} \\
&= \mathsf{TV}(p_{\hat{x}_{N-1}} \circ K_{\mathsf{DPnP},\eta},\ \pi_\eta) + \varepsilon_{\mathsf{PCS}}.
\end{aligned}$$

The above two inequalities together imply

$$\mathsf{TV}(p_{\hat{x}_N}, \pi_\eta) \leq \mathsf{TV}(p_{\hat{x}_{N-1}} \circ K_{\mathsf{DPnP},\eta},\ \pi_\eta) + \varepsilon_{\mathsf{DDS}} + \varepsilon_{\mathsf{PCS}}.$$

Iterating this process, we obtain

$$\mathsf{TV}(p_{\hat{x}_N}, \pi_\eta) \le \mathsf{TV}(p_{\hat{x}_1} \circ K_{\mathsf{DPnP},\eta}^{(N-1)},\ \pi_\eta) + (N-1)(\varepsilon_{\mathsf{DDS}} + \varepsilon_{\mathsf{PCS}}). \tag{A.18}$$

It remains to bound $\mathsf{TV}(p_{\hat{x}_1} \circ K_{\mathsf{DPnP},\eta}^{(N-1)},\ \pi_\eta)$. For this, we need the following two lemmas.

Lemma 14 (Comparing TV and χ^2-divergence, Polyanskiy and Wu [93])**.** *For any two distributions p, q, we have*

$$\mathsf{TV}(p, q) \le \sqrt{\chi^2(p \,\|\, q)}.$$

Lemma 15 (χ^2-contractivity of $K_{\mathsf{DPnP},\eta}$)**.** *There exists some $\lambda := \lambda(p^\star, \mathcal{L}, \eta) \in (0, 1)$, such that for any probability distribution $p(x)$, we have*

$$\chi^2(p \circ K_{\mathsf{DPnP},\eta}^{(N)} \parallel \pi_\eta) \le \lambda^{2N} \chi^2(p \,\|\, \pi_\eta).$$

A form of Lemma 15 is well-known for Markov chains with countable state spaces, but relatively few sources provide a complete proof for the abstract setting we consider here with continuous state space. For sake of completeness, we prove Lemma 15 in Appendix A.3.6.

Combining the above two lemmas, we obtain

$$\mathsf{TV}(p_{\hat{x}_1} \circ K_{\mathsf{DPnP},\eta}^{(N-1)},\ \pi_\eta) \le \sqrt{\chi^2(p_{\hat{x}_1} \circ K_{\mathsf{DPnP},\eta}^{(N-1)} \parallel \pi_\eta)} \le \lambda^{N-1} \sqrt{\chi^2(p_{\hat{x}_1} \,\|\, \pi_\eta)}.$$

Plug this into (A.18), we obtain

$$\mathsf{TV}(p_{\hat{x}_N}, \pi_\eta) \le \lambda^{N-1} \sqrt{\chi^2(p_{\hat{x}_1} \,\|\, \pi_\eta)} + (N-1)(\varepsilon_{\mathsf{DDS}} + \varepsilon_{\mathsf{PCS}}).$$

With $N \asymp \frac{\log(1/\varepsilon_{\mathsf{acc}})}{1-\lambda}$ such that $\lambda^{N-1} \le \exp\big(-(N-1)(1-\lambda)\big) \le \varepsilon_{\mathsf{acc}}$, the desired result readily follows. □

A.3.6 Proof of Lemma 15

Proof. We need a few fundamental properties of reversible Markov chains, which are collected below.

First we set up some notation. Define the Hilbert space $L^2(\pi)$ to be the space of square-integrable functions with respect to measure π, i.e., those functions $f : \mathbb{R}^d \to \mathbb{C}$ such that

$$\|f\|_{L^2(\pi)} \coloneqq \left(\int |f(x)|^2 \pi(x) \mathrm{d}x \right)^{1/2} < \infty.$$

The first well-known property [103] offers a way to represent a reversible transition kernel as a self-adjoint operator (infinite-dimensional symmetric matrix).

Lemma 16 (Self-adjoint representation of reversible Markov operator)**.** *Assume $K(x, x')$ is a Markov transition kernel that is reversible with respect to the stationary distribution $\pi(x)$. Then the integral operator $\mathcal{K} : L^2(\pi) \to L^2(\pi)$ defined by*

$$\mathcal{K}f(x) = \int K(x, x') f(x') \mathrm{d}x'$$

is self-adjoint and compact. For any probability distribution $p(x)$ such that $\int \frac{p^2(x)}{\pi(x)} \mathrm{d}x < \infty$, we have

$$\int p(x) \cdot \mathcal{K}f(x) \mathrm{d}x = \int p \circ K(x') f(x') \mathrm{d}x'.$$

Moreover, the eigenvalues of $\mathcal{K}$ are the same as those of K.

The following theorem is a generalization of the classical Perron-Frobenius theory for finite-dimensional transition matrix to strictly positive operators. The form we present here can be found in Schaefer [104, Theorem V.6.6]; see also Bourbaki [13, Theorem III.6.7] for a more elementary treatment which can also be adapted to the form we need.

Theorem 8 (Jentzsch)**.** *Let $K(x, x')$ be a Markov transition kernel. If $K(x, x') > 0$ for any $x, x' \in \mathbb{R}^d$, then K has a unique stationary distribution π. Moreover, 1 is a simple eigenvalue of K, with π being the only left eigenfunction, and the constant function 1 being the only right eigenfunction. In addition, there exists $\lambda \in (0, 1)$ such that any other eigenvalue of K has modulus no larger than λ.*

We are now ready to prove Lemma 15. We divide the proof into the following steps.

Step I: controlling the eigenvalues of $\mathcal{K}_{\mathsf{DPnP},\eta}$. Recall the auxiliary kernel $K_{\mathsf{aux},\eta}$ defined in (A.15). It is a standard result in linear algebra or function analysis

[12] that $K_{\mathsf{aux},\eta} = K_{\mathsf{PCS},\eta} \circ K_{\mathsf{DDS},\eta}$ has same eigenvalues as $K_{\mathsf{DPnP},\eta} = K_{\mathsf{DDS},\eta} \circ K_{\mathsf{PCS},\eta}$. From (A.15), it is easy to check $K_{\mathsf{aux},\eta}(x, x') > 0$, thus Theorem 8 implies 1 is a simple eigenvalue of $\mathcal{K}_{\mathsf{DPnP},\eta}$. Moreover, there exists $\lambda := \lambda(p^\star, \mathcal{L}, \eta) \in (0, 1)$, such that any other eigenvalue of $K_{\mathsf{aux},\eta}$ has modulus no larger than λ.

Since $K_{\mathsf{DPnP},\eta}$ has the same eigenvalues as $K_{\mathsf{aux},\eta}$, and, by Lemma 16, the operator $\mathcal{K}_{\mathsf{DPnP},\eta}$ also has the same eigenvalues as these two, we conclude that $\mathcal{K}_{\mathsf{DPnP},\eta}$ is a self-adjoint compact operator on $L^2(\pi_\eta)$, of whom 1 is a simple eigenvalue. Moreover, any other eigenvalue of $\mathcal{K}_{\mathsf{DPnP},\eta}$ has modulus no larger than λ.

Step II: establishing the contractivity of $\mathcal{K}_{\mathsf{DPnP},\eta}$ in $L^2(\pi_\eta)$. It is easy to verify that the constant function $\mathbf{1}$, which takes value 1 for any $x \in \mathbb{R}^d$, is a eigenfunction of $\mathcal{K}_{\mathsf{DPnP},\eta}$ associated to the simple eigenvalue 1, thus is the only (up to scaling) eigenfunction associated to that eigenvalue. It is also a unit-length eigenfunction, since $\|\mathbf{1}\|_{L^2(\pi_\eta)} = (\int 1 \cdot \pi_\eta(x) dx)^{1/2} = 1$. Therefore, the operator $\mathcal{K}_{\mathsf{DPnP},\eta} - \mathbf{1}\mathbf{1}^\top$ is a self-adjoint operator whose eigenvalues have modulus no larger than λ, where $\mathbf{1}\mathbf{1}^\top$ is the orthogonal projection onto $\mathbf{1}$ in $L^2(\pi_\eta)$, defined by

$$\mathbf{1}\mathbf{1}^\top f(x) \equiv \int f(x')\pi_\eta(x')\mathrm{d}x', \quad \forall x \in \mathbb{R}^d.$$

Using the fact that $\mathcal{K}_{\mathsf{DPnP},\eta}\mathbf{1}\mathbf{1}^\top = \mathbf{1}\mathbf{1}^\top\mathcal{K}_{\mathsf{DPnP},\eta} = \mathbf{1}\mathbf{1}^\top$, one may show $(\mathcal{K}_{\mathsf{DPnP},\eta} - \mathbf{1}\mathbf{1}^\top)^N = \mathcal{K}^{(N)}_{\mathsf{DPnP},\eta} - \mathbf{1}\mathbf{1}^\top$ by expanding the product, see e.g. Saloff-Coste [103]. Consequently, $\mathcal{K}^N_{\mathsf{DPnP},\eta} - \mathbf{1}\mathbf{1}^\top$ is a self-adjoint operator whose eigenvalues have modulus no larger than λ^N, i.e.,

$$\left\|\mathcal{K}^N_{\mathsf{DPnP},\eta} - \mathbf{1}\mathbf{1}^\top\right\|_{L^2(\pi_\eta)\to L^2(\pi_\eta)} \le \lambda^N, \tag{A.19}$$

where $\|\cdot\|_{L^2(\pi_\eta)\to L^2(\pi_\eta)}$ denotes the operator norm on $L^2(\pi_\eta)$.

Step III: bounding the inner product of $p \circ K^{(N)}_{\mathsf{DPnP},\eta} - \pi_\eta$ with any square-integrable function. Note that when $\chi^2(p \,\|\, \pi_\eta) = \infty$, the conclusion is trivially true. For the rest part of the proof, we assume $\chi^2(p \,\|\, \pi_\eta) < \infty$. Now, for any

$f \in L^2(\pi_\eta)$, by applying Lemma 16 iteratively, we obtain

$$\begin{aligned}\int p \circ K^{(N)}_{\mathsf{DPnP},\eta}(x) f(x) \mathrm{d}x &= \int p(x') \mathcal{K}^N_{\mathsf{DPnP},\eta} f(x') \mathrm{d}x' \\ &= \int p(x') \cdot (\mathcal{K}^N_{\mathsf{DPnP},\eta} - \mathbf{1}\mathbf{1}^\top) f(x') \mathrm{d}x' + \int p(x') \mathbf{1}\mathbf{1}^\top f(x') \mathrm{d}x' \\ &= \int p(x') \cdot (\mathcal{K}^N_{\mathsf{DPnP},\eta} - \mathbf{1}\mathbf{1}^\top) f(x') \mathrm{d}x' + \int f(x') \pi_\eta(x') \mathrm{d}x',\end{aligned}$$

where the last line follows from the definition of $\mathbf{1}\mathbf{1}^\top$ and $\int p(x')\mathrm{d}x' = 1$. Rearrange the terms to see

$$\int \big(p \circ K^{(N)}_{\mathsf{DPnP},\eta}(x) - \pi_\eta(x)\big) f(x) \mathrm{d}x = \int p(x') \cdot (\mathcal{K}^N_{\mathsf{DPnP},\eta} - \mathbf{1}\mathbf{1}^\top) f(x') \mathrm{d}x'. \tag{A.20}$$

In particular, taking $p = \pi_\eta$ yields

$$0 = \int \pi_\eta(x') \cdot (\mathcal{K}^N_{\mathsf{DPnP},\eta} - \mathbf{1}\mathbf{1}^\top) f(x') \mathrm{d}x'. \tag{A.21}$$

Substract (A.21) from (A.20), and then take absolute value, we obtain

$$\begin{aligned}&\left| \int \big(p \circ K^{(N)}_{\mathsf{DPnP},\eta}(x) - \pi_\eta(x)\big) f(x) \mathrm{d}x \right| \\ &= \left| \int \big(p(x') - \pi_\eta(x')\big) \cdot (\mathcal{K}^N_{\mathsf{DPnP},\eta} - \mathbf{1}\mathbf{1}^\top) f(x') \mathrm{d}x' \right| \\ &\le \left(\int \frac{\big(p(x') - \pi_\eta(x')\big)^2}{\pi_\eta(x)} \mathrm{d}x \right)^{1/2} \cdot \big\| (\mathcal{K}^N_{\mathsf{DPnP},\eta} - \mathbf{1}\mathbf{1}^\top) f(x') \big\|_{L^2(\pi_\eta)} \\ &\le \sqrt{\chi^2(p \,\|\, \pi_\eta)} \cdot \lambda^N \|f\|_{L^2(\pi_\eta)}.\end{aligned} \tag{A.22}$$

Step IV: choosing an appropriate square-integrable function. Now, set

$$f(x) = \frac{p \circ K^{(N)}_{\mathsf{DPnP},\eta}(x) - \pi_\eta(x)}{\pi_\eta(x)}.$$

It is easily checked that

$$\int \big(p \circ K^{(N)}_{\mathsf{DPnP},\eta}(x) - \pi_\eta(x)\big) f(x) \mathrm{d}x = \chi^2(p \circ K^{(N)}_{\mathsf{DPnP},\eta} \parallel \pi_\eta),$$
$$\|f\|_{L^2(\pi_\eta)} = \sqrt{\chi^2(p \circ K^{(N)}_{\mathsf{DPnP},\eta} \parallel \pi_\eta)}.$$

Plug these equations into (A.22), we obtain

$$\chi^2(p \circ K^{(N)}_{\mathsf{DPnP},\eta} \parallel \pi_\eta) \leq \lambda^{2N} \chi^2(p \,\|\, \pi_\eta),$$

as claimed. □

Appendix B

Proofs for Chapter 4

We collect here the auxiliary notation and the supporting lemmas used in the proof. Throughout, we write

$$\lambda \coloneqq \frac{s}{n}, \qquad q_t \coloneqq (1-\lambda)\phi_t + \lambda\phi_{t+1/s},$$

where ϕ_u denotes the density of $\mathcal{N}(0,u)$. By definition, we can verify

$$p_t = q_t^{\otimes n}$$

We also set

$$a_t \coloneqq 2t + 16\sigma^2 \log n, \qquad b_0(t) \coloneqq t, \qquad b_1(t) \coloneqq t + \frac{1}{s}.$$

For $t \in [\tau, T]$ and $z \in \mathbb{R}$, let

$$r_{t,z}(x) \coloneqq \frac{1}{Z_{t,z}} q_t(x) \exp\Big(-\frac{(x-z)^2}{2a_t}\Big), \qquad \rho_{t,z}(dx) \coloneqq r_{t,z}(x)\,\mathrm{d}x,$$

and write

$$h_{t,z}(x) \coloneqq \log r_{t,z}(x).$$

Recall that we define an auxiliary family of measures by

$$\Pi_t^y(\mathrm{d}x) \coloneqq \frac{1}{\mathcal{Z}_t(y)} p_t(x) \exp\Big(-\frac{\|x-y\|^2}{2a_t}\Big)\,\mathrm{d}x = \bigotimes_{i=1}^n \rho_{t,y_i}(x_i)\,\mathrm{d}x_i.$$

Lemma 17 (Explicit formula for Wasserstein action in dimension one [121])**.** *Let $(\rho_t)_{t\in[a,b]} \subset \mathcal{P}_2(\mathbb{R})$ be absolutely continuous in W_2. Assume that each ρ_t has density f_t, and let F_t be the corresponding c.d.f. Then for a.e. t,*

$$|\dot\rho|_t^2 = \int_{\mathbb{R}} \frac{(\partial_t F_t(x))^2}{f_t(x)}\,\mathrm{d}x,$$

so that

$$\mathcal{A}\big((\rho_t)_{t\in[a,b]}\big) = \int_a^b \int_{\mathbb{R}} \frac{(\partial_t F_t(x))^2}{f_t(x)}\,\mathrm{d}x\,\mathrm{d}t.$$

B.1 Proof of the action bound

Proof of Lemma 5. The density Π_t^y factorizes over coordinates:

$$\Pi_t^y = \bigotimes_{i=1}^n \rho_{t,y_i}.$$

For each coordinate i, choose a one-dimensional velocity field $v_{t,i}$ generating the curve $t \mapsto \rho_{t,y_i}$, and define the product velocity field

$$V_t(x_1,\dots,x_n) \coloneqq (v_{t,1}(x_1),\dots,v_{t,n}(x_n)).$$

Because Π_t^y is a product measure, V_t transports the curve $t \mapsto \Pi_t^y$, and

$$\int_{\mathbb{R}^n} |V_t(x)|^2\,\Pi_t^y(dx) = \sum_{i=1}^n \int_{\mathbb{R}} |v_{t,i}(x)|^2\,\rho_{t,y_i}(dx).$$

Taking the infimum over admissible velocity fields yields

$$\mathcal{A}_y \le \sum_{i=1}^n \mathcal{I}(y_i),$$

where $\mathcal{I}(z) \coloneqq \mathcal{A}\big((\rho_{t,z})_{t\in[\tau,T]}\big)$. Thus it suffices to prove the following one-dimensional estimate.

Lemma 18 (One-dimensional action bound)**.** *Assume the standing assumptions of the sparse denoising model, with $0 < \tau \le \sigma_0^2$ and $\sigma_0^2 \le c/s^3$ for a sufficiently small*

universal constant $c > 0$. Fix $z \in \mathbb{R}$, and consider

$$\rho_{t,z}(\mathrm{d}x) \propto q_t(x) \exp\left(-\frac{(x-z)^2}{2(\sigma_0^2 + 2t)}\right) \mathrm{d}x, \qquad t \in [\tau, T].$$

Let $f_{t,z}$ and $F_{t,z}$ be the density and c.d.f. of $\rho_{t,z}$, and set

$$\mathcal{I}(z) \coloneqq \int_\tau^T \int_{\mathbb{R}} \frac{|\partial_t F_{t,z}(x)|^2}{f_{t,z}(x)} \,\mathrm{d}x \,\mathrm{d}t.$$

Define

$$L_\sigma \coloneqq \log \frac{n}{s\sigma_0}.$$

Then

$$\mathcal{I}(z) \lesssim \log \frac{T}{\tau} + sz^2 + \frac{n}{\sigma_0 s^{3/2}} L_\sigma^2.$$

The proof is postponed to Appendix B.1.1.

Applying Lemma 18 coordinatewise with $\sigma_0 = 4\sigma\sqrt{\log n}$ and $z = y_i$, and summing over i, yields

$$\mathcal{A}_y \lesssim n \log \frac{T}{\tau} + s\|y\|^2 + \frac{n^2}{\sigma_0 s^{3/2}} L_\sigma^2,$$

which is exactly the claimed bound, given $\sigma_0^2 \le c/s^3$ for sufficiently small constant $c > 0$. □

B.1.1 Proof of Lemma 18

Proof. Write

$$\lambda = \frac{s}{n}, \qquad a_t = \sigma_0^2 + 2t, \qquad b_0(t) = t, \qquad b_1(t) = t + \frac{1}{s},$$

and

$$\pi_0 \coloneqq 1 - \lambda, \qquad \pi_1 \coloneqq \lambda, \qquad \Delta_j(t) \coloneqq a_t + b_j(t), \qquad j \in \{0, 1\}.$$

Thus

$$\Delta_0(t) = \sigma_0^2 + 3t, \qquad \Delta_1(t) = \sigma_0^2 + 3t + \frac{1}{s}.$$

A completion of squares shows that $\rho_{t,z}$ is a two-Gaussian mixture:

$$f_{t,z} = \omega_0(t,z) g_{0,t,z} + \omega_1(t,z) g_{1,t,z}, \qquad F_{t,z} = \omega_0(t,z) G_{0,t,z} + \omega_1(t,z) G_{1,t,z},$$

where $g_{j,t,z}$ and $G_{j,t,z}$ are the density and c.d.f. of the Gaussian law $\mathcal{N}(\mu_j(t), v_j(t))$, with

$$\mu_j(t) \coloneqq \frac{b_j(t)}{\Delta_j(t)} z, \qquad v_j(t) \coloneqq \frac{a_t b_j(t)}{\Delta_j(t)}.$$

The mixture weights are

$$\omega_j(t,z) = \frac{\widetilde{W}_j(t,z)}{\widetilde{W}_0(t,z) + \widetilde{W}_1(t,z)},$$

where

$$\widetilde{W}_j(t,z) = \pi_j \Delta_j(t)^{-1/2} \exp\left(-\frac{z^2}{2\Delta_j(t)}\right).$$

For $j \in \{0,1\}$, define

$$\psi_j(t,x) \coloneqq \dot{\mu}_j(t) + \frac{\dot{v}_j(t)}{2v_j(t)}(x - \mu_j(t)).$$

Since

$$\partial_t G_{j,t,z}(x) = -g_{j,t,z}(x)\psi_j(t,x),$$

we obtain

$$\partial_t F_{t,z} = \dot{\omega}_1(t,z)(G_{1,t,z} - G_{0,t,z}) - \omega_0(t,z) g_{0,t,z}\psi_0 - \omega_1(t,z) g_{1,t,z}\psi_1.$$

Therefore,

$$\frac{|\partial_t F_{t,z}|^2}{f_{t,z}} \le 3\dot{\omega}_1(t,z)^2 \frac{(G_{1,t,z} - G_{0,t,z})^2}{f_{t,z}} + 3\omega_0(t,z)^2 \frac{g_{0,t,z}^2 \psi_0^2}{f_{t,z}} + 3\omega_1(t,z)^2 \frac{g_{1,t,z}^2 \psi_1^2}{f_{t,z}}.$$

Set

$$S_t(z) \coloneqq \dot{\omega}_1(t,z)^2 \int_{\mathbb{R}} \frac{(G_{1,t,z}(x) - G_{0,t,z}(x))^2}{f_{t,z}(x)} \,\mathrm{d}x,$$

and

$$H_{j,t}(z) \coloneqq \omega_j(t,z)^2 \int_{\mathbb{R}} \frac{g_{j,t,z}(x)^2 \psi_j(t,x)^2}{f_{t,z}(x)} \,\mathrm{d}x.$$

Then

$$\mathcal{I}(z) \le 3 \int_\tau^T \big(S_t(z) + H_{0,t}(z) + H_{1,t}(z)\big) \,\mathrm{d}t.$$

Step I: Gaussian-shape terms $H_{0,t}$ and $H_{1,t}$. Since $f_{t,z} \geq \omega_j(t,z) g_{j,t,z}$,

$$H_{j,t}(z) \leq \omega_j(t,z) \int_{\mathbb{R}} g_{j,t,z}(x) \psi_j(t,x)^2 \, \mathrm{d}x = \omega_j(t,z) \left(\dot{\mu}_j(t)^2 + \frac{\dot{v}_j(t)^2}{4 v_j(t)} \right).$$

A direct calculation gives

$$\dot{\mu}_j(t) = \frac{(a_t - 2b_j(t))z}{\Delta_j(t)^2}, \qquad \dot{v}_j(t) = \frac{a_t^2 + 2b_j(t)^2}{\Delta_j(t)^2}.$$

Hence

$$\frac{\dot{v}_j(t)^2}{4 v_j(t)} \lesssim \frac{\Delta_j(t)}{a_t b_j(t)} = \frac{1}{a_t} + \frac{1}{b_j(t)}.$$

Using $b_0(t) = t$, $b_1(t) = t + 1/s$, $a_t = \sigma_0^2 + 2t$, and $\tau \leq \sigma_0^2$, we get

$$\int_\tau^T \sum_{j=0}^{1} \omega_j(t,z) \frac{\dot{v}_j(t)^2}{4 v_j(t)} \, \mathrm{d}t \lesssim \log \frac{T}{\tau}.$$

It remains to bound the mean-motion terms. For the slab component,

$$\dot{\mu}_1(t) = \frac{(\sigma_0^2 - 2/s)z}{\Delta_1(t)^2}.$$

Since $\sigma_0^2 \leq c/s$,

$$\int_\tau^T \omega_1(t,z) \dot{\mu}_1(t)^2 \, \mathrm{d}t \leq \frac{C z^2}{s^2} \int_0^\infty \frac{\mathrm{d}t}{(\sigma_0^2 + 3t + 1/s)^4} \lesssim s z^2.$$

For the spike component,

$$\dot{\mu}_0(t) = \frac{\sigma_0^2 z}{\Delta_0(t)^2}.$$

We claim that

$$\int_\tau^T \omega_0(t,z) \frac{\sigma_0^4 z^2}{\Delta_0(t)^4} \, \mathrm{d}t \lesssim L_\sigma. \tag{B.1}$$

To prove this, put

$$u = \Delta_0(t), \qquad r = \frac{1}{s}, \qquad Z = z^2.$$

Then $\Delta_1(t) = u + r$, $\mathrm{d}t = \mathrm{d}u/3$, and

$$\omega_0(t,z) = \frac{1}{1 + \widetilde{R}_t \mathrm{e}^{\Gamma_t(z)}},$$

where

$$\widetilde{R}_t = \frac{\lambda}{1-\lambda}\sqrt{\frac{u}{u+r}}, \qquad \Gamma_t(z) = \frac{Z}{2su(u+r)}.$$

Since $\lambda = s/n$,

$$\log \frac{1}{\widetilde{R}_t} = \log \frac{1-\lambda}{\lambda} + \frac{1}{2}\log \frac{u+r}{u} \lesssim L_\sigma \qquad \text{for all } u \geq \sigma_0^2.$$

Thus

$$\omega_0(t,z) \leq \min\left\{1, \exp\left(CL_\sigma - \frac{Z}{2su(u+r)}\right)\right\}.$$

Increasing L_σ by a universal constant factor, it suffices to bound

$$\sigma_0^4 Z \int_{\sigma_0^2}^{\infty} \frac{1}{u^4} \min\left\{1, \exp\left(L_\sigma - \frac{Z}{2su(u+r)}\right)\right\} \mathrm{d}u.$$

On $u \leq r$, we have $su(u+r) \leq 2u$, hence

$$\frac{Z}{2su(u+r)} \geq \frac{Z}{4u}.$$

Therefore the contribution of $u \leq r$ is at most

$$C\sigma_0^4 Z \int_{\sigma_0^2}^{r} u^{-4} \min\left\{1, \exp\left(L_\sigma - \frac{Z}{4u}\right)\right\} \mathrm{d}u.$$

If $Z \leq \sigma_0^2 L_\sigma$, this is bounded trivially by

$$C\sigma_0^4 Z \int_{\sigma_0^2}^{\infty} u^{-4}\,\mathrm{d}u \lesssim \frac{Z}{\sigma_0^2} \lesssim L_\sigma.$$

If $Z > \sigma_0^2 L_\sigma$, the change of variables $y = Z/u$ gives the bound

$$C\frac{\sigma_0^4}{Z^2} \int_0^{\infty} y^2 \min\{1, \mathrm{e}^{L_\sigma - cy}\}\,\mathrm{d}y \lesssim \frac{\sigma_0^4}{Z^2} L_\sigma^3 \lesssim L_\sigma.$$

On $u \geq r$, we have $su(u+r) \leq 2su^2$, hence

$$\frac{Z}{2su(u+r)} \geq \frac{Z}{4su^2}.$$

The contribution of $u \geq r$ is therefore at most

$$C\sigma_0^4 Z \int_r^\infty u^{-4} \min\left\{1, \exp\left(L_\sigma - \frac{Z}{4su^2}\right)\right\} \mathrm{d}u.$$

If $Zs \leq L_\sigma$, this is bounded by

$$C\sigma_0^4 Z \int_r^\infty u^{-4} \,\mathrm{d}u \lesssim \sigma_0^4 s^3 Z \leq \sigma_0^4 s^2 L_\sigma \lesssim L_\sigma.$$

If $Zs > L_\sigma$, the change of variables $y = Z/(su^2)$ gives

$$C\sigma_0^4 s^{3/2} Z^{-1/2} \int_0^\infty y^{1/2} \min\{1, \mathrm{e}^{L_\sigma - cy}\} \,\mathrm{d}y \lesssim \sigma_0^4 s^{3/2} Z^{-1/2} L_\sigma^{3/2} \lesssim \sigma_0^4 s^2 L_\sigma \lesssim L_\sigma.$$

This proves (B.1). Consequently,

$$\int_\tau^T \big(H_{0,t}(z) + H_{1,t}(z)\big) \,\mathrm{d}t \lesssim \log \frac{T}{\tau} + sz^2 + L_\sigma. \tag{B.2}$$

Step II: the switching term S_t. Let

$$\mathcal{H}_t(x; z) \coloneqq G_{1,t,z}(x) - G_{0,t,z}(x).$$

Since $f_{t,z} \geq \omega_1(t,z) g_{1,t,z}$,

$$S_t(z) \leq \frac{\dot{\omega}_1(t,z)^2}{\omega_1(t,z)} \int_{\mathbb{R}} \frac{\mathcal{H}_t(x;z)^2}{g_{1,t,z}(x)} \,\mathrm{d}x.$$

By the Gaussian Hardy inequality, applied with $\partial_x \mathcal{H}_t = g_{1,t,z} - g_{0,t,z}$,

$$\int_{\mathbb{R}} \frac{\mathcal{H}_t(x;z)^2}{g_{1,t,z}(x)} \,\mathrm{d}x \leq v_1(t) \int_{\mathbb{R}} \frac{(g_{1,t,z}(x) - g_{0,t,z}(x))^2}{g_{1,t,z}(x)} \,\mathrm{d}x.$$

For Gaussian densities q_0, q_1 with variances v_0, v_1 and means μ_0, μ_1,

$$\int_{\mathbb{R}} \frac{q_0(x)^2}{q_1(x)}\,\mathrm{d}x = \frac{v_1}{\sqrt{v_0(2v_1 - v_0)}} \exp\left(\frac{(\mu_1 - \mu_0)^2}{2v_1 - v_0}\right),$$

provided $2v_1 > v_0$. In the present setting,

$$2v_1(t) - v_0(t) = \frac{a_t\,(3t^2 + t\sigma_0^2 + 5t/s + 2\sigma_0^2/s)}{\Delta_0(t)\Delta_1(t)} > 0.$$

Thus

$$S_t(z) \lesssim \frac{\dot{\omega}_1(t,z)^2}{\omega_1(t,z)} \frac{v_1(t)^2}{\sqrt{v_0(t)(2v_1(t) - v_0(t))}} \mathrm{e}^{E_t(z)}, \tag{B.3}$$

where

$$E_t(z) \coloneqq \frac{(\mu_1(t) - \mu_0(t))^2}{2v_1(t) - v_0(t)}.$$

We proceed to bound the factor $\frac{v_1(t)^2}{\sqrt{v_0(t)(2v_1(t) - v_0(t))}}$. Set

$$N_t \coloneqq 3t^2 + t\sigma_0^2 + \frac{5t}{s} + \frac{2\sigma_0^2}{s}.$$

Since

$$(t + 1/s)\Delta_0(t) = 3t^2 + t\sigma_0^2 + \frac{3t}{s} + \frac{\sigma_0^2}{s} \le N_t \le \frac{8}{3}(t + 1/s)\Delta_0(t),$$

we have

$$N_t \asymp (t + 1/s)\Delta_0(t).$$

Using

$$2v_1(t) - v_0(t) = \frac{a_t N_t}{\Delta_0(t)\Delta_1(t)},$$

it follows that

$$2v_1(t) - v_0(t) \asymp \frac{a_t(t + 1/s)}{\Delta_1(t)} = v_1(t).$$

Therefore

$$\frac{v_1(t)^2}{\sqrt{v_0(t)(2v_1(t) - v_0(t))}} \lesssim \frac{v_1(t)^{3/2}}{\sqrt{v_0(t)}}. \tag{B.4}$$

Since

$$v_0(t) = \frac{a_t t}{\Delta_0(t)} \qquad \text{and} \qquad v_1(t) = \frac{a_t(t+1/s)}{\Delta_1(t)} \le a_t \le \Delta_0(t),$$

we obtain

$$\frac{v_1(t)^{3/2}}{\sqrt{v_0(t)}} = v_1(t)^{3/2}\sqrt{\frac{\Delta_0(t)}{a_t t}} \le a_t\sqrt{\frac{\Delta_0(t)}{t}} \le \frac{\Delta_0(t)^{3/2}}{\sqrt{t}}.$$

Thus

$$\frac{v_1(t)^2}{\sqrt{v_0(t)(2v_1(t)-v_0(t))}} \lesssim \frac{\Delta_0(t)^{3/2}}{\sqrt{t}}.$$

Next write

$$R_t(z) := \frac{\omega_1(t,z)}{\omega_0(t,z)} = \widetilde{R}_t \mathrm{e}^{\Gamma_t(z)}, \tag{B.5}$$

where

$$\widetilde{R}_t = \frac{\lambda}{1-\lambda}\sqrt{\frac{\Delta_0(t)}{\Delta_1(t)}}, \qquad \Gamma_t(z) = \frac{z^2}{2s\Delta_0(t)\Delta_1(t)}. \tag{B.6}$$

Since

$$\omega_1 = \frac{R_t}{1+R_t}, \qquad \omega_0 = \frac{1}{1+R_t},$$

we have

$$\dot\omega_1 = \omega_0\omega_1\partial_t \log R_t, \qquad \frac{\dot\omega_1^2}{\omega_1} = \omega_0^2\omega_1|\partial_t \log R_t|^2. \tag{B.7}$$

Moreover,

$$\mu_1(t) - \mu_0(t) = \frac{z a_t}{s\Delta_0(t)\Delta_1(t)}$$

and

$$E_t(z) = \frac{z^2 a_t}{s^2\Delta_0(t)\Delta_1(t)\left(3t^2 + t\sigma_0^2 + 5t/s + 2\sigma_0^2/s\right)} \le \Gamma_t(z).$$

Therefore

$$\omega_0(t,z)^2\omega_1(t,z)\mathrm{e}^{E_t(z)} \le \frac{R_t(z)}{(1+R_t(z))^3}\mathrm{e}^{\Gamma_t(z)} = \frac{\widetilde{R}_t \mathrm{e}^{2\Gamma_t(z)}}{\left(1+\widetilde{R}_t\mathrm{e}^{\Gamma_t(z)}\right)^3}. \tag{B.8}$$

Also, we will need an explicit formula for $\partial_t \log R_t(z)$. Recall (B.5)–(B.6). Since $\dot\Delta_0(t) = \dot\Delta_1(t) = 3$, we have

$$\partial_t \log \widetilde{R}_t = \frac{1}{2}\left(\frac{\dot\Delta_0(t)}{\Delta_0(t)} - \frac{\dot\Delta_1(t)}{\Delta_1(t)}\right) = \frac{3}{2}\left(\frac{1}{\Delta_0(t)} - \frac{1}{\Delta_1(t)}\right).$$

Using $\Delta_1(t) - \Delta_0(t) = 1/s$, this becomes

$$\partial_t \log \widetilde{R}_t = \frac{3}{2s\Delta_0(t)\Delta_1(t)}.$$

Moreover,

$$\partial_t \Gamma_t(z) = -\frac{z^2}{2s} \frac{\dot{\Delta}_0(t)\Delta_1(t) + \Delta_0(t)\dot{\Delta}_1(t)}{\Delta_0(t)^2\Delta_1(t)^2} = -\frac{3z^2(\Delta_0(t) + \Delta_1(t))}{2s\Delta_0(t)^2\Delta_1(t)^2}.$$

Therefore

$$\partial_t \log R_t(z) = \partial_t \log \widetilde{R}_t + \partial_t \Gamma_t(z) = \frac{3}{2s\Delta_0(t)\Delta_1(t)} - \frac{3z^2(\Delta_0(t) + \Delta_1(t))}{2s\Delta_0(t)^2\Delta_1(t)^2}.$$

Hence

$$|\partial_t \log R_t(z)| \lesssim \frac{1}{s\Delta_0(t)\Delta_1(t)} + \frac{\Gamma_t(z)}{\Delta_0(t)}. \tag{B.9}$$

Combine the preceding estimates (B.3), (B.4), (B.7), (B.8), and (B.9), we obtain the pointwise bound

$$S_t(z) \lesssim \frac{\Delta_0(t)^{3/2}}{\sqrt{t}} \left(\frac{1}{s\Delta_0(t)\Delta_1(t)} + \frac{\Gamma_t(z)}{\Delta_0(t)} \right)^2 \frac{\widetilde{R}_t \mathrm{e}^{2\Gamma_t(z)}}{\left(1 + \widetilde{R}_t \mathrm{e}^{\Gamma_t(z)}\right)^3}. \tag{B.10}$$

The next claim provides a deterministic integral estimate needed to integrate this bound.

Claim 1. *For every $z \in \mathbb{R}$,*

$$\int_\tau^T \frac{\Delta_0(t)^{3/2}}{\sqrt{t}} \left(\frac{1}{s\Delta_0(t)\Delta_1(t)} + \frac{\Gamma_t(z)}{\Delta_0(t)} \right)^2 \frac{\widetilde{R}_t \mathrm{e}^{2\Gamma_t(z)}}{\left(1 + \widetilde{R}_t \mathrm{e}^{\Gamma_t(z)}\right)^3} \, \mathrm{d}t \lesssim \frac{n}{s} L_\sigma^2 \sqrt{1 + \frac{1}{s\sigma_0^2}}.$$

The proof of the claim is postponed to the end of this proof.

Combining (B.10) with Claim 1, we obtain

$$\int_\tau^T S_t(z) \, \mathrm{d}t \lesssim \frac{n}{s} L_\sigma^2 \sqrt{1 + \frac{1}{s\sigma_0^2}}. \tag{B.11}$$

Step III: putting things together. Using

$$\mathcal{I}(z) \le 3 \int_\tau^T \big(S_t(z) + H_{0,t}(z) + H_{1,t}(z)\big)\,\mathrm{d}t,$$

together with (B.2) and (B.11), we get

$$\mathcal{I}(z) \lesssim \log \frac{T}{\tau} + sz^2 + \frac{n}{s} L_\sigma^2 \sqrt{1 + \frac{1}{s\sigma_0^2}}.$$

Since $s\sigma_0^2 \lesssim 1$ by the assumption (4.9), the proof is complete. □

We now prove Claim 1.

Proof of Claim 1. Set

$$u = \Delta_0(t), \qquad a = \sigma_0^2, \qquad r = \frac{1}{s}, \qquad Z = z^2, \qquad \alpha = \frac{\lambda}{1-\lambda}.$$

Then

$$\Delta_1(t) = u + r, \qquad \Gamma_t(z) = \Gamma(u) := \frac{Z}{2su(u+r)} = \frac{Zr}{2u(u+r)},$$

and

$$\widetilde{R}_t = \alpha \sqrt{\frac{u}{u+r}}.$$

Also $u = a + 3t$, so $t = t(u) = (u-a)/3$ and $\mathrm{d}t = \mathrm{d}u/3$. We shall freely extend the resulting u-integral to $[a, \infty)$, since the integrand is nonnegative.

Define

$$A(u) := \log \frac{1}{\widetilde{R}_t} = \log \frac{1}{\alpha} + \frac{1}{2} \log \frac{u+r}{u}.$$

Since $\lambda = s/n$, for all $u \ge a$, it can be verified that

$$\left|\log \frac{1}{\alpha}\right| = \left|\log \frac{n-s}{s}\right| \lesssim L_\sigma, \qquad \left|\log \frac{u+r}{u}\right| \lesssim \log\Big(1 + \frac{1}{s\sigma_0^2}\Big) \lesssim L_\sigma,$$

where the first inequality can be shown by considering the case $s \le n/2$ and the case $s > n/2$ where $L_\sigma \gg \log n$ by (4.9) separately. Therefore

$$|A(u)| \lesssim L_\sigma, \qquad \mathrm{e}^{A(u)} = \frac{1}{\widetilde{R}_t} \lesssim \frac{n}{s} \sqrt{\frac{u+r}{u}}.$$

Let

$$\Phi(w) := \frac{\mathrm{e}^{2w}}{(1+\mathrm{e}^{w})^3}.$$

Then

$$\frac{\widetilde{R}_t \mathrm{e}^{2\Gamma_t(z)}}{\left(1+\widetilde{R}_t \mathrm{e}^{\Gamma_t(z)}\right)^3} = \mathrm{e}^{A(u)}\Phi(\Gamma(u)-A(u)),$$

and

$$\Phi(w) \lesssim \mathrm{e}^{-|w|}.$$

Write

$$B(u) \coloneqq \frac{1}{su(u+r)} = \frac{r}{u(u+r)}.$$

Using the previous display and changing variables from t to u, the left-hand side of the claim is bounded by

$$\underbrace{\frac{n}{s}\int_a^\infty \frac{u\sqrt{u+r}}{\sqrt{t(u)}}\left(B(u)+\frac{\Gamma(u)}{u}\right)^2 \Phi(\Gamma(u)-A(u))\,\mathrm{d}u}_{=:\mathcal{I}}.$$

We contend that

$$\mathcal{I} \lesssim L_\sigma^2\sqrt{1+\frac{r}{a}},$$

which is the desired estimate.

We split the u-integral into the boundary region

$$u \in [a, 2a]$$

and the bulk region

$$u \in [2a, \infty).$$

On the boundary region, $u \asymp a$, and $t(u)^{-1/2}$ is integrable. Moreover,

$$B(u) = \frac{r}{u(u+r)} \le \frac{1}{u},$$

so

$$B(u) + \frac{\Gamma(u)}{u} \le \frac{1+\Gamma(u)}{u}.$$

Therefore

$$\mathcal{I}_{\mathrm{bd}} \lesssim \int_a^{2a} \frac{\sqrt{u+r}}{u\sqrt{t(u)}}(1+\Gamma(u))^2\Phi(\Gamma(u)-A(u))\,\mathrm{d}u.$$

Since $u \asymp a$,

$$\frac{\sqrt{u+r}}{u} \lesssim \frac{1}{\sqrt{a}}\sqrt{1+\frac{r}{a}},$$

and

$$\int_a^{2a} \frac{\mathrm{d}u}{\sqrt{t(u)}} \lesssim \sqrt{a}.$$

Finally, it is elementary to verify

$$\sup_{y\geq 0,\ |A|\lesssim L_\sigma} (1+y)^2\Phi(y-A) \lesssim L_\sigma^2.$$

Hence

$$\mathcal{I}_{\mathrm{bd}} \lesssim L_\sigma^2\sqrt{1+\frac{r}{a}}.$$

We now consider the bulk region $u \geq 2a$. Since $t(u) \asymp u$,

$$\frac{u\sqrt{u+r}}{\sqrt{t(u)}} \lesssim u\sqrt{1+\frac{r}{u}} \leq u\sqrt{1+\frac{r}{a}}.$$

Thus

$$\mathcal{I}_{\mathrm{bulk}} \lesssim \sqrt{1+\frac{r}{a}}\int_{2a}^{\infty} u\left(B(u)+\frac{\Gamma(u)}{u}\right)^2\Phi(\Gamma(u)-A(u))\,\mathrm{d}u.$$

It remains to show that the last integral is $O(L_\sigma^2)$.

We split the bulk region into

$$D_< := \{\Gamma < 1\}, \qquad D_{\mathrm{loc}} := \{\Gamma \geq 1,\ \Gamma \leq 4A\}, \qquad D_{\mathrm{tail}} := \{\Gamma \geq 1,\ \Gamma > 4A\}.$$

First consider $D_<$. Expanding the square,

$$u\left(B+\frac{\Gamma}{u}\right)^2 = uB^2 + 2B\Gamma + \frac{\Gamma^2}{u}.$$

We estimate the three terms separately.

For the uB^2 term, since $\Phi \lesssim 1$, we have

$$\begin{aligned}\int_{D_<} uB(u)^2\Phi(\Gamma - A)\,\mathrm{d}u &\lesssim \int_{2a}^{\infty} u\left(\frac{r}{u(u+r)}\right)^2 \mathrm{d}u \\ &= \int_{2a}^{\infty}\left(\frac{1}{u} - \frac{1}{u+r} - \frac{r}{(u+r)^2}\right)\mathrm{d}u \\ &= \log\left(1+\frac{r}{2a}\right) - \frac{r}{2a+r} \\ &\lesssim \log\left(1+\frac{1}{s\sigma_0^2}\right) \\ &\lesssim L_\sigma.\end{aligned}$$

For the mixed term, we again use $\Phi \lesssim 1$. If $Z = 0$, then $\Gamma \equiv 0$ and there is nothing to prove. Otherwise, Γ is strictly decreasing in u and

$$-\Gamma'(u) = \Gamma(u)\left(\frac{1}{u} + \frac{1}{u+r}\right) = \Gamma(u)\frac{2u+r}{u(u+r)}.$$

Hence

$$B(u)\Gamma(u)\,\mathrm{d}u = \frac{r}{u(u+r)}\Gamma(u)\,\mathrm{d}u = \frac{r}{2u+r}\,(-\mathrm{d}\Gamma(u)).$$

Since $0 \le r/(2u+r) \le 1$ and $\Gamma < 1$ on $D_<$,

$$\int_{D_<} B(u)\Gamma(u)\Phi(\Gamma - A)\,\mathrm{d}u \lesssim \int_{D_<} B(u)\Gamma(u)\,\mathrm{d}u \le \int_{D_<}(-\mathrm{d}\Gamma) \lesssim 1.$$

Similarly,

$$\frac{\Gamma(u)^2}{u}\,\mathrm{d}u = \Gamma(u)\frac{u+r}{2u+r}(-\mathrm{d}\Gamma(u)) \le \Gamma(u)(-\mathrm{d}\Gamma(u)).$$

Therefore

$$\int_{D_<}\frac{\Gamma(u)^2}{u}\Phi(\Gamma - A)\,\mathrm{d}u \lesssim \int_0^1 y\,\mathrm{d}y \lesssim 1.$$

Combining the three estimates,

$$\int_{D_<} u\left(B(u) + \frac{\Gamma(u)}{u}\right)^2 \Phi(\Gamma - A)\,\mathrm{d}u \lesssim 1.$$

Next consider D_{loc}. Since $\Gamma \geq 1$,

$$B(u) \leq \frac{1}{u} \leq \frac{\Gamma(u)}{u},$$

and hence

$$B(u) + \frac{\Gamma(u)}{u} \lesssim \frac{\Gamma(u)}{u}.$$

Thus

$$u\left(B(u) + \frac{\Gamma(u)}{u}\right)^2 \lesssim \frac{\Gamma(u)^2}{u}.$$

Set

$$W(u) := \Gamma(u) - A(u).$$

We have

$$A'(u) = \frac{1}{2}\left(\frac{1}{u+r} - \frac{1}{u}\right) = -\frac{r}{2u(u+r)}.$$

Therefore, on $\{\Gamma \geq 1\}$,

$$-W'(u) = -\Gamma'(u) + A'(u) = \Gamma(u)\left(\frac{1}{u} + \frac{1}{u+r}\right) - \frac{r}{2u(u+r)} \asymp \frac{\Gamma(u)}{u}.$$

Consequently,

$$\frac{\mathrm{d}u}{u} \asymp \frac{-\mathrm{d}W}{\Gamma(u)} \qquad \text{on } D_{\mathrm{loc}}.$$

Since on D_{loc},

$$\Gamma(u) \leq 4A(u) \lesssim L_\sigma,$$

we get

$$\begin{aligned}\int_{D_{\mathrm{loc}}} \frac{\Gamma(u)^2}{u}\Phi(W(u))\,\mathrm{d}u &\lesssim \int_{D_{\mathrm{loc}}} \Gamma(u)\Phi(W(u))\,(-\mathrm{d}W) \\ &\lesssim L_\sigma \int_{\mathbb{R}} \Phi(W)\,\mathrm{d}W \\ &\lesssim L_\sigma.\end{aligned}$$

Finally consider D_{tail}. Again,

$$B(u) + \frac{\Gamma(u)}{u} \lesssim \frac{\Gamma(u)}{u}.$$

On D_{tail},

$$\Gamma(u) > 4A(u),$$

so

$$\Gamma(u) - A(u) \geq \frac{3}{4}\Gamma(u).$$

Therefore

$$\Phi(\Gamma(u) - A(u)) \lesssim \mathrm{e}^{-c\Gamma(u)}.$$

Using the change of variables $y = \Gamma(u)$, and the relation

$$-\Gamma'(u) \asymp \frac{\Gamma(u)}{u},$$

we obtain

$$\begin{aligned}\int_{D_{\text{tail}}} \frac{\Gamma(u)^2}{u}\Phi(\Gamma(u) - A(u))\,\mathrm{d}u &\lesssim \int_{D_{\text{tail}}} \frac{\Gamma(u)^2}{u}\mathrm{e}^{-c\Gamma(u)}\,\mathrm{d}u \\ &\lesssim \int_0^\infty y\mathrm{e}^{-cy}\,\mathrm{d}y \\ &\lesssim 1.\end{aligned}$$

Putting the three bulk estimates together gives

$$\mathcal{I}_{\text{bulk}} \lesssim \sqrt{1 + \frac{r}{a}}\,(1 + L_\sigma + 1) \lesssim L_\sigma^2\sqrt{1 + \frac{r}{a}}.$$

Together with the boundary estimate,

$$\mathcal{I} \lesssim L_\sigma^2\sqrt{1 + \frac{r}{a}}.$$

Recalling that $r = 1/s$ and $a = \sigma_0^2$, the original left-hand side is bounded by

$$\frac{n}{s}L_\sigma^2\sqrt{1 + \frac{1}{s\sigma_0^2}},$$

as claimed. □

B.2 Proof of comparison error

We need a few estimates on the score function before we proceed, whose proof is postponed to Appendix B.2.1.

Lemma 19 (Regularity of score)**.** *Let*

$$\mathsf{M}_1 \coloneqq \sup_{t\in[\tau,T]} t \int_{\mathbb{R}^n} \|\nabla \log \Pi_t^y(x)\|^2 \, \Pi_t^y(dx),$$

$$\mathsf{L}_2 \coloneqq \sup_{t\in[\tau,T]} \sup_{x\in\mathbb{R}^n} t \, \|\nabla^2 \log \Pi_t^y(x)\|, \qquad \mathsf{L}_3 \coloneqq \sup_{t\in[\tau,T]} \sup_{x\in\mathbb{R}^n} t^{3/2} \, \|\nabla^3 \log \Pi_t^y(x)\|,$$

and

$$\mathsf{H}_1 \coloneqq \sup_{t\in[\tau,T]} t^3 \int_{\mathbb{R}^n} \|\partial_t \nabla \log \Pi_t^y(x)\|^2 \, \Pi_t^y(dx),$$

where the tensor norms are Euclidean operator norms. Then

$$\mathsf{M}_1 \le 2n,$$

$$\mathsf{L}_2 \lesssim \log(n\tau^{-1}), \qquad \mathsf{L}_3 \lesssim \log^{3/2}(n\tau^{-1}),$$

and

$$\mathsf{H}_1 \lesssim \frac{\|y\|_2^2}{\tau} + \frac{n\sigma^2}{\tau} \log n + n \log^3(n\tau^{-1}).$$

We are now ready to prove Lemma 6.

Proof of Lemma 6. Recall the exponential schedule

$$t_k = T\Big(\frac{\tau}{T}\Big)^{k/K}, \qquad \eta_k = \Lambda(t_k - t_{k+1}), \qquad u_k \coloneqq \sum_{j=0}^{k-1} \eta_j = \Lambda(T - t_k),$$

and define

$$\alpha \coloneqq \Lambda\Big(1 - \Big(\frac{\tau}{T}\Big)^{1/K}\Big) = \frac{\eta_k}{t_k}, \qquad \bar{t}(u) \coloneqq T - \frac{u}{\Lambda}, \qquad S \coloneqq \Lambda(T - \tau).$$

Then $\bar{t}(u_k) = t_k$.

Let $\bar{X}$ be the piecewise-frozen interpolation of (4.3), namely

$$\bar{X}_0 = x_0, \qquad \mathrm{d}\bar{X}_u = \big(s_{t_k}(\bar{X}_{u_k}) + \lambda_k \nabla \mathcal{L}(X_{u_k}; y)\big)\,\mathrm{d}u + \sqrt{2}\,\mathrm{d}B_u, \qquad u \in [u_k, u_{k+1}].$$

By construction, $\bar{X}_{u_k}$ has the same law as x_k for every k. Let Q denote the path law of $\bar{X}$.

Let P denote the path law of the reference process

$$\mathrm{d}\widetilde{X}_u = \Big(\nabla \log \Pi^y_{\bar{t}(u)}(\widetilde{X}_u) + \widetilde{v}_u(\widetilde{X}_u)\Big)\,\mathrm{d}u + \sqrt{2}\,\mathrm{d}B_u, \qquad \widetilde{X}_0 \sim \Pi^y_T,$$

where $\widetilde{v}_u$ is a velocity field generating the reparametrized curve $u \mapsto \Pi^y_{\bar{t}(u)}$, chosen so that

$$\int_0^S \|\widetilde{v}_u\|^2_{L^2(\Pi^y_{\bar{t}(u)})}\,\mathrm{d}u = \frac{\mathcal{A}_y}{\Lambda}.$$

To see why this is possible, let v_t be a velocity field generating the original curve $t \mapsto \Pi^y_t$ on $[\tau, T]$. Then $\widetilde{X}_u \sim \Pi^y_{\bar{t}(u)}$, so in particular $\widetilde{X}_S \sim \Pi^y_\tau$. Furthermore, v_t can be chosen so that

$$\int_\tau^T \|v_t\|^2_{L^2(\Pi^y_t)}\,\mathrm{d}t = \mathcal{A}_y.$$

Since $\bar{t}(u) = T - u/\Lambda$, we have

$$\frac{\mathrm{d}\bar{t}}{\mathrm{d}u} = -\frac{1}{\Lambda}.$$

Therefore the reparametrized continuity equation is generated by

$$\widetilde{v}_u = -\frac{1}{\Lambda}\, v_{\bar{t}(u)}.$$

Consequently,

$$\int_0^S \|\widetilde{v}_u\|^2_{L^2(\Pi^y_{\bar{t}(u)})}\,\mathrm{d}u = \frac{1}{\Lambda^2}\int_0^S \|v_{\bar{t}(u)}\|^2_{L^2(\Pi^y_{\bar{t}(u)})}\,\mathrm{d}u.$$

Changing variables $t = \bar{t}(u)$, so that $\mathrm{d}u = -\Lambda\,\mathrm{d}t$, gives

$$\int_0^S \|\widetilde{v}_u\|^2_{L^2(\Pi^y_{\bar{t}(u)})}\,\mathrm{d}u = \frac{1}{\Lambda}\int_\tau^T \|v_t\|^2_{L^2(\Pi^y_t)}\,\mathrm{d}t = \frac{\mathcal{A}_y}{\Lambda}.$$

We proceed to compare the law of $\bar{X}$ and $\widetilde{X}$. By Girsanov's theorem,

$$\begin{aligned}\mathsf{KL}(P\|Q) \le \mathsf{KL}(\Pi_T^y\|\mathsf{Law}(x_0)) \\ + \frac{1}{4}\sum_{k=0}^{K-1}\int_{u_k}^{u_{k+1}} \mathbb{E}_P\Big\|\widetilde{v}_u(\widetilde{X}_u) + \nabla \log \Pi^y_{\bar{t}(u)}(\widetilde{X}_u) \\ - s_{t_k}(\widetilde{X}_{u_k}) - \lambda_k \nabla \mathcal{L}(\widetilde{X}_{u_k}; y)\Big\|^2 \mathrm{d}u.\end{aligned}$$

Recall that $\nabla \log \Pi^y_{t_k}(x) = \nabla \log p_{t_k}(x) - \lambda_k \nabla \mathcal{L}(x; y)$. Using triangle inequality gives

$$\begin{aligned}\mathsf{KL}(P\|Q) \le &\mathsf{KL}(\Pi_T^y\|\mathsf{Law}(x_0)) + \frac{\mathcal{A}_y}{\Lambda} + \sum_{k=0}^{K-1}\int_{u_k}^{u_{k+1}} \mathbb{E}_P\|\Delta_{k,u}\|^2 \,\mathrm{d}u \\ &+ \sum_{k=1}^{K-1}(u_{k+1}-u_k)\mathbb{E}_P\|\nabla \log p_{t_k}(\widetilde{X}_{u_k}) - s_{t_k}(\widetilde{X}_{u_k})\|^2, \end{aligned} \tag{B.12}$$

where

$$\Delta_{k,u} \coloneqq \nabla \log \Pi^y_{\bar{t}(u)}(\widetilde{X}_u) - \nabla \log \Pi^y_{t_k}(\widetilde{X}_{u_k}).$$

Note that $\Delta_{k,u} = \int_{u_k}^{u_{k+1}} \mathrm{d}\nabla \log \Pi^y_{\bar{t}(u)}(\widetilde{X}_u)$. Apply Itô's formula to obtain

$$\mathrm{d}\nabla \log \Pi^y_{\bar{t}(u)}(\widetilde{X}_u) = A_u \,\mathrm{d}u + \sqrt{2}\,\nabla^2 \log \Pi^y_{\bar{t}(u)}(\widetilde{X}_u)\,\mathrm{d}B_u,$$

where

$$\begin{aligned}A_u = &-\frac{1}{\Lambda}\partial_t \nabla \log \Pi^y_{\bar{t}(u)}(\widetilde{X}_u) + \nabla^2 \log \Pi^y_{\bar{t}(u)}(\widetilde{X}_u)\Big(\nabla \log \Pi^y_{\bar{t}(u)}(\widetilde{X}_u) + \widetilde{v}_u(\widetilde{X}_u)\Big) \\ &+ \Delta \nabla \log \Pi^y_{\bar{t}(u)}(\widetilde{X}_u).\end{aligned}$$

For $u \in [u_k, u_{k+1}]$, one has $t_k \asymp \bar{t}(u)$ provided $\alpha \le 1/2$. Hence Itô's isometry and Cauchy–Schwarz imply

$$\mathbb{E}_P\|\Delta_{k,u}\|^2 \le 2(u-u_k)\int_{u_k}^{u} \mathbb{E}_P\|A_r\|^2 \,\mathrm{d}r + 4\int_{u_k}^{u} \mathbb{E}_P\|\nabla^2 \log \Pi^y_{\bar{t}(r)}(\widetilde{X}_r)\|_{\mathsf{F}}^2 \,\mathrm{d}r.$$

Using Lemma 19, we obtain

$$\mathbb{E}_P\|A_r\|^2 \le C\left(\frac{\mathsf{H}_1}{\Lambda^2 t_k^3} + \frac{\mathsf{L}_2^2\mathsf{M}_1}{t_k^3} + \frac{\mathsf{L}_2^2}{t_k^2}\|\widetilde{v}_r\|^2_{L^2(\Pi^y_{\bar{t}(r)})} + \frac{n\mathsf{L}_3^2}{t_k^3}\right),$$

while

$$\mathbb{E}_P\|\nabla^2 \log \Pi^y_{\widetilde{t}(r)}(\widetilde{X}_r)\|_{\mathrm{F}}^2 \le \frac{Cn\mathsf{L}_2^2}{t_k^2}.$$

Integrating over $[u_k, u_{k+1}]$ yields

$$\int_{u_k}^{u_{k+1}} \mathbb{E}_P\|\Delta_{k,u}\|^2\, du \lesssim \frac{\mathsf{L}_2^2\eta_k}{t_k^2}\int_{u_k}^{u_{k+1}} \|\widetilde{v}_r\|^2_{L^2(\Pi^y_{\widetilde{t}(r)})}\, dr + \Big(\frac{\mathsf{H}_1}{\Lambda^2 t_k^3} + \frac{\mathsf{L}_2^2\mathsf{M}_1}{t_k^3} + \frac{n\mathsf{L}_2^2}{t_k^2} + \frac{n\mathsf{L}_3^2}{t_k^3}\Big)\eta_k^2. \tag{B.13}$$

Summing over k and using $\eta_k/t_k = \alpha$ gives

$$\sum_{k=0}^{K-1}\int_{u_k}^{u_{k+1}} \mathbb{E}_P\|\Delta_{k,u}\|^2\, du \lesssim \frac{\mathsf{L}_2^2\alpha}{\tau}\int_0^S \|\widetilde{v}_u\|^2_{L^2(\Pi^y_{\widetilde{t}(u)})}\, du + \Big(\tau^{-1}\Lambda^{-2}\mathsf{H}_1 + \tau^{-1}\mathsf{L}_2^2\mathsf{M}_1 + n\mathsf{L}_2^2 + \tau^{-1}n\mathsf{L}_3^2\Big)K\alpha^2.$$

Substituting this into (B.12), and using

$$\int_0^S \|\widetilde{v}_u\|^2_{L^2(\Pi^y_{\widetilde{t}(u)})}\, du = \frac{\mathcal{A}_y}{\Lambda},$$

we obtain

$$\mathsf{KL}(P\|Q) \le \mathsf{KL}(\Pi_T^y\|\mathsf{Law}(x_0)) + C\Big(1 + \frac{\alpha}{\tau}\log^2(n\tau^{-1})\Big)\frac{\mathcal{A}_y}{\Lambda} + C\left(\frac{\|y\|_2^2}{\Lambda^2\tau^2} + \frac{n\sigma^2}{\Lambda^2\tau^2}\log n + \frac{n}{\tau}\log^3(n\tau^{-1})\right)K\alpha^2.$$

From the definition of α, it is readily verified that

$$\alpha \lesssim \frac{\Lambda\log(T/\tau)}{K}, \qquad K\alpha^2 \lesssim \frac{\Lambda^2\log^2(T/\tau)}{K}.$$

Hence

$$\mathsf{KL}(P\|Q) \le \mathsf{KL}(\Pi_T^y\|\mathsf{Law}(x_0)) + C\left(\frac{1}{\Lambda} + \frac{1}{K\tau}\log\frac{T}{\tau}\cdot\log^2\frac{n}{\tau}\right)\mathcal{A}_y + C\left(\frac{\|y\|_2^2}{K\tau^2} + \frac{n\sigma^2}{K\tau^2}\log n + \frac{\Lambda^2 n}{K\tau}\log^3\frac{n}{\tau}\right)\log^2\frac{T}{\tau}.$$

Finally, by data processing under the endpoint map $\omega \mapsto \omega_S$,

$$\mathsf{KL}(\Pi_\tau^y \| \mathsf{Law}(x_K)) = \mathsf{KL}(\mathsf{Law}_P(\widetilde{X}_S) \| \mathsf{Law}_Q(\bar{X}_S)) \le \mathsf{KL}(P\|Q),$$

which proves the lemma. □

B.2.1 Proof of Lemma 19

We first prove the following one-dimensional version of Lemma 19.

Proposition 1 (One-dimensional regularity)**.** *For every $t \in [\tau, T]$ and every $z \in \mathbb{R}$,*

$$\sup_{x\in\mathbb{R}} t\,|h''_{t,z}(x)| \lesssim \log(n\tau^{-1}), \tag{B.14}$$

$$\sup_{x\in\mathbb{R}} t^{3/2}|h'''_{t,z}(x)| \lesssim \log^{3/2}(n\tau^{-1}), \tag{B.15}$$

$$t\int_{\mathbb{R}} |h'_{t,z}(x)|^2 r_{t,z}(x)\,dx \le 2, \tag{B.16}$$

$$t^3\int_{\mathbb{R}} |\partial_t h'_{t,z}(x)|^2 r_{t,z}(x)\,dx \lesssim \Big(\frac{|z|^2}{\tau} + \frac{\sigma^2}{\tau}\log n + \log^3(n\tau^{-1})\Big). \tag{B.17}$$

Proof. Write as before

$$a_t \coloneqq 2t + 16\sigma^2 \log n, \qquad b_0(t) \coloneqq t, \qquad b_1(t) \coloneqq t + \frac{1}{s}.$$

Recall that

$$r_{t,z}(x) \propto \Big((1-\lambda)\phi_{b_0(t)}(x) + \lambda\phi_{b_1(t)}(x)\Big)\exp\Big(-\frac{(x-z)^2}{2a_t}\Big).$$

It is convenient to work first with the two unnormalized summands. For $j \in \{0,1\}$, set

$$\tilde{f}_{j,t,z}(x) \coloneqq A_j(t,z)\exp\Big(-\frac{\beta_j(t)}{2}x^2 + \gamma_{t,z}x\Big),$$

where

$$\beta_j(t) \coloneqq \frac{1}{b_j(t)} + \frac{1}{a_t}, \qquad \gamma_{t,z} \coloneqq \frac{z}{a_t},$$

and

$$A_0(t,z) \coloneqq (1-\lambda)(2\pi b_0(t))^{-1/2}\mathrm{e}^{-z^2/(2a_t)}, \qquad A_1(t,z) \coloneqq \lambda(2\pi b_1(t))^{-1/2}\mathrm{e}^{-z^2/(2a_t)}.$$

Then

$$r_{t,z}(x) = \frac{\tilde{f}_{0,t,z}(x) + \tilde{f}_{1,t,z}(x)}{Z_{t,z}}, \qquad Z_{t,z} \coloneqq \int_{\mathbb{R}} \big(\tilde{f}_{0,t,z}(u) + \tilde{f}_{1,t,z}(u)\big)\, du.$$

Accordingly,

$$h_{t,z}(x) = \log\big(\tilde{f}_{0,t,z}(x) + \tilde{f}_{1,t,z}(x)\big) - \log Z_{t,z},$$

so every x-derivative of $h_{t,z}$ can be computed directly from $\tilde{f}_{0,t,z} + \tilde{f}_{1,t,z}$.

Introduce

$$\Delta_t \coloneqq \beta_0(t) - \beta_1(t) = \frac{1}{t} - \frac{1}{t + 1/s} = \frac{1}{t(1+st)} > 0,$$

and define the local posterior slab weight

$$\omega_{t,z}(x) \coloneqq \frac{\tilde{f}_{1,t,z}(x)}{\tilde{f}_{0,t,z}(x) + \tilde{f}_{1,t,z}(x)}.$$

Since the two summands share the same linear term $\gamma_{t,z}x$, their log-ratio is purely quadratic:

$$\log \frac{\tilde{f}_{1,t,z}(x)}{\tilde{f}_{0,t,z}(x)} = B_t + \frac{\Delta_t}{2} x^2,$$

where

$$B_t \coloneqq \log \frac{A_1(t,z)}{A_0(t,z)} = \log \frac{\lambda}{1-\lambda} + \frac{1}{2} \log \frac{t}{t + 1/s}.$$

In particular, B_t is independent of z. Since $\lambda = s/n$ and $t \geq \tau$,

$$|B_t| \lesssim \log(n\tau^{-1}). \tag{B.18}$$

Moreover,

$$\partial_t B_t = \frac{1}{2}\Big(\frac{1}{t} - \frac{1}{t + 1/s}\Big) = \frac{\Delta_t}{2}.$$

Thus

$$\omega_{t,z}(x) = \vartheta\Big(B_t + \frac{\Delta_t}{2} x^2\Big), \qquad \vartheta(r) \coloneqq \frac{1}{1 + \mathrm{e}^{-r}}.$$

We shall also use the coefficient bounds

$$\beta_j(t) \lesssim t^{-1}, \qquad \Delta_t \lesssim t^{-1}, \qquad |\partial_t \beta_j(t)| \lesssim t^{-2}, \qquad |\partial_t \gamma_{t,z}| \lesssim |z|\, t^{-2}, \tag{B.19}$$

together with

$$|\partial_t \Delta_t| = \frac{1+2st}{t^2(1+st)^2} \lesssim \frac{\Delta_t}{t} \lesssim t^{-2}, \qquad |\partial_t B_t| = \frac{\Delta_t}{2}. \tag{B.20}$$

We next record the elementary softmax estimate that will be used repeatedly.

Claim. For every $m \geq 0$, there exists $C_m < \infty$ such that

$$\sup_{u \geq 0} u^m \, \vartheta(A+u)\big(1-\vartheta(A+u)\big) \leq C_m(1+|A|^m), \qquad A \in \mathbb{R}. \tag{B.21}$$

Proof of claim. Since $\vartheta(r)(1-\vartheta(r)) \leq \mathrm{e}^{-|r|}$ for all $r \in \mathbb{R}$, it is enough to bound $u^m \mathrm{e}^{-|A+u|}$.

If $A \geq 0$, then $A+u \geq u$, and hence

$$u^m \mathrm{e}^{-|A+u|} \leq u^m \mathrm{e}^{-u} \leq C_m.$$

If $A < 0$, we split into two regimes. When $0 \leq u \leq 2|A|$, we simply use $\vartheta(A+u)(1-\vartheta(A+u)) \leq 1/4$, obtaining

$$u^m \vartheta(A+u)(1-\vartheta(A+u)) \leq 2^m |A|^m.$$

When $u \geq 2|A|$, we have $A+u \geq u/2$, so

$$u^m \vartheta(A+u)(1-\vartheta(A+u)) \leq u^m \mathrm{e}^{-u/2} \leq C_m.$$

Combining the two cases proves (B.21). □

We now compute the x-derivatives of $h_{t,z}$. Since

$$\partial_x \log \tilde{f}_{j,t,z}(x) = -\beta_j(t)x + \gamma_{t,z},$$

we obtain

$$h'_{t,z}(x) = \gamma_{t,z} - \beta_0(t)x + \omega_{t,z}(x)\Delta_t x. \tag{B.22}$$

Differentiating $\omega_{t,z}(x) = \vartheta(B_t + \frac{\Delta_t}{2}x^2)$ in x yields

$$\omega'_{t,z}(x) = \omega_{t,z}(x)\big(1-\omega_{t,z}(x)\big)\Delta_t x,$$

and therefore

$$h''_{t,z}(x) = -\beta_0(t) + \omega_{t,z}(x)\Delta_t + \omega_{t,z}(x)\big(1-\omega_{t,z}(x)\big)\Delta_t^2 x^2. \tag{B.23}$$

Differentiating once more gives

$$h'''_{t,z}(x) = 3\omega_{t,z}(x)\big(1-\omega_{t,z}(x)\big)\Delta_t^2 x + \omega_{t,z}(x)\big(1-\omega_{t,z}(x)\big)\big(1-2\omega_{t,z}(x)\big)\Delta_t^3 x^3. \tag{B.24}$$

We now prove the four stated estimates.

Bound (B.14)**.** From (B.23),

$$|h''_{t,z}(x)| \le \beta_0(t) + \Delta_t + \omega_{t,z}(x)\big(1-\omega_{t,z}(x)\big)\Delta_t^2 x^2.$$

Set

$$u := \frac{\Delta_t}{2}x^2 \ge 0.$$

Then

$$\omega_{t,z}(x) = \vartheta(B_t + u), \qquad \Delta_t^2 x^2 = 2\Delta_t u,$$

so by (B.21) with $m = 1$,

$$\omega_{t,z}(x)\big(1-\omega_{t,z}(x)\big)\Delta_t^2 x^2 = 2\Delta_t\, u\, \vartheta(B_t+u)\big(1-\vartheta(B_t+u)\big) \lesssim \Delta_t(1+|B_t|).$$

Using $\beta_0(t) \lesssim t^{-1}$ and $\Delta_t \lesssim t^{-1}$, we conclude that

$$\sup_{x\in\mathbb{R}} |h''_{t,z}(x)| \lesssim \frac{1+|B_t|}{t}.$$

Multiplying by t and invoking (B.18) proves (B.14).

Bound (B.15)**.** From (B.24) and $|1-2\omega_{t,z}(x)| \le 1$,

$$|h'''_{t,z}(x)| \le 3\omega_{t,z}(x)\big(1-\omega_{t,z}(x)\big)\Delta_t^2|x| + \omega_{t,z}(x)\big(1-\omega_{t,z}(x)\big)\Delta_t^3|x|^3.$$

With the same substitution $u = \frac{\Delta_t}{2}x^2$, we have

$$\Delta_t^2|x| = \sqrt{2}\,\Delta_t^{3/2}u^{1/2}, \qquad \Delta_t^3|x|^3 = 2^{3/2}\,\Delta_t^{3/2}u^{3/2}.$$

Applying (B.21) with $m = \frac{1}{2}$ and $m = \frac{3}{2}$, we obtain

$$\sup_{x\in\mathbb{R}} |h'''_{t,z}(x)| \lesssim \Delta_t^{3/2}\big(1 + |B_t|^{3/2}\big).$$

Since $\Delta_t \lesssim t^{-1}$, it follows that

$$t^{3/2} \sup_{x\in\mathbb{R}} |h'''_{t,z}(x)| \lesssim 1 + |B_t|^{3/2} \lesssim \log^{3/2}(n\tau^{-1}),$$

which is (B.15).

Bound (B.16)**.** Let

$$r_{t,z} = w_0(t,z)\, g_{0,t,z} + w_1(t,z)\, g_{1,t,z}$$

be the normalized Gaussian-mixture representation, where $g_{j,t,z}$ is the density of $\mathcal{N}(\mu_j(t,z), v_j(t))$, with

$$\mu_j(t,z) = \frac{b_j(t)}{a_t + b_j(t)}\, z, \qquad v_j(t) = \frac{a_t b_j(t)}{a_t + b_j(t)}.$$

Then

$$\partial_x \log g_{j,t,z}(x) = \partial_x \log \tilde{f}_{j,t,z}(x) = -\beta_j(t)x + \gamma_{t,z}.$$

Let $J \in \{0,1\}$ be the latent mixture index. Conditional on $X = x$, the posterior law of J is

$$\mathbb{P}(J = 1 \mid X = x) = \omega_{t,z}(x), \qquad \mathbb{P}(J = 0 \mid X = x) = 1 - \omega_{t,z}(x),$$

and therefore

$$h'_{t,z}(x) = \mathbb{E}[\partial_x \log g_{J,t,z}(x) \,|\, X = x]\,.$$

By Jensen's inequality,

$$|h'_{t,z}(x)|^2 \le \mathbb{E}\big[|\partial_x \log g_{J,t,z}(x)|^2 \,\big|\, X = x\big]\,.$$

Integrating against $r_{t,z}(x)\,dx$ and then taking expectation in J gives

$$\int_{\mathbb{R}} |h'_{t,z}(x)|^2 r_{t,z}(x)\,dx \le \sum_{j=0}^{1} w_j(t,z) \int_{\mathbb{R}} |\partial_x \log g_{j,t,z}(x)|^2 g_{j,t,z}(x)\,dx.$$

For a one-dimensional Gaussian with precision β, the Fisher information equals β. Hence

$$\int_{\mathbb{R}} |\partial_x \log g_{j,t,z}(x)|^2 g_{j,t,z}(x)\,dx = \beta_j(t).$$

Since $\beta_0(t) \geq \beta_1(t)$,

$$\int_{\mathbb{R}} |h'_{t,z}(x)|^2 r_{t,z}(x)\,dx \leq \beta_0(t) = \frac{1}{t} + \frac{1}{a_t} \leq \frac{3}{2t} \leq \frac{2}{t}.$$

This proves (B.16).

Bound (B.17)**.** Differentiating (B.22) in t, we obtain

$$\partial_t h'_{t,z}(x) = \partial_t \gamma_{t,z} - (\partial_t \beta_0(t))x + \omega_{t,z}(x)(\partial_t \Delta_t)x + \Delta_t x\,\partial_t \omega_{t,z}(x).$$

Since

$$\omega_{t,z}(x) = \vartheta\Big(B_t + \frac{\Delta_t}{2}x^2\Big),$$

its t-derivative is

$$\partial_t \omega_{t,z}(x) = \omega_{t,z}(x)\big(1 - \omega_{t,z}(x)\big)\Big(\partial_t B_t + \frac{\partial_t \Delta_t}{2}x^2\Big).$$

Therefore

$$\begin{aligned}\partial_t h'_{t,z}(x) = \partial_t \gamma_{t,z} - (\partial_t \beta_0(t))x + \omega_{t,z}(x)(\partial_t \Delta_t)x \\ + \omega_{t,z}(x)\big(1 - \omega_{t,z}(x)\big)\Delta_t(\partial_t B_t)x \\ + \frac{1}{2}\omega_{t,z}(x)\big(1 - \omega_{t,z}(x)\big)\Delta_t(\partial_t \Delta_t)x^3. \qquad \text{(B.25)}\end{aligned}$$

We now estimate the nonlinear terms. Set again

$$u := \frac{\Delta_t}{2}x^2 \geq 0, \qquad \omega_{t,z}(x) = \vartheta(B_t + u).$$

Using (B.21) with $m = \frac{1}{2}$, we obtain

$$\begin{aligned}\omega_{t,z}(x)\big(1-\omega_{t,z}(x)\big)\Delta_t|\partial_t B_t|\,|x| &= \sqrt{2}\,|\partial_t B_t|\,\Delta_t^{1/2}\,u^{1/2}\vartheta(B_t+u)\big(1-\vartheta(B_t+u)\big)\\ &\lesssim |\partial_t B_t|\,\Delta_t^{1/2}\big(1+|B_t|^{1/2}\big)\\ &\lesssim \Delta_t^{3/2}\big(1+|B_t|^{1/2}\big)\\ &\lesssim t^{-3/2}\big(1+|B_t|^{1/2}\big),\end{aligned} \tag{B.26}$$

since $|\partial_t B_t| = \Delta_t/2$.

Similarly, using (B.21) with $m = \frac{3}{2}$,

$$\begin{aligned}\omega_{t,z}(x)\big(1-\omega_{t,z}(x)\big)\Delta_t|\partial_t \Delta_t|\,|x|^3 &= 2^{3/2}|\partial_t\Delta_t|\,\Delta_t^{-1/2}\,u^{3/2}\vartheta(B_t+u)\big(1-\vartheta(B_t+u)\big)\\ &\lesssim |\partial_t\Delta_t|\,\Delta_t^{-1/2}\big(1+|B_t|^{3/2}\big)\\ &\lesssim \frac{\Delta_t^{1/2}}{t}\big(1+|B_t|^{3/2}\big)\\ &\lesssim t^{-3/2}\big(1+|B_t|^{3/2}\big),\end{aligned} \tag{B.27}$$

where we used $|\partial_t\Delta_t| \lesssim \Delta_t/t$.

Combining (B.25), (B.19), (B.20), (B.26), and (B.27), we arrive at the pointwise estimate

$$|\partial_t h'_{t,z}(x)| \lesssim \frac{|z|}{t^2} + \frac{|x|}{t^2} + \frac{1+|B_t|^{3/2}}{t^{3/2}}. \tag{B.28}$$

Consequently,

$$|\partial_t h'_{t,z}(x)|^2 \lesssim \frac{|z|^2}{t^4} + \frac{|x|^2}{t^4} + \frac{1+|B_t|^3}{t^3}. \tag{B.29}$$

It remains to bound the second moment of $r_{t,z}$. From the Gaussian-mixture representation above,

$$\int_{\mathbb{R}} x^2 r_{t,z}(x)\,dx = \sum_{j=0}^{1} w_j(t,z)\big(\mu_j(t,z)^2 + v_j(t)\big).$$

Since

$$|\mu_j(t,z)| = \frac{b_j(t)}{a_t+b_j(t)}|z| \le |z|, \qquad v_j(t) = \frac{a_t b_j(t)}{a_t+b_j(t)} \le a_t,$$

we obtain

$$\int_{\mathbb{R}} x^2 r_{t,z}(x)\,dx \le |z|^2 + a_t \le |z|^2 + 2t + 16\sigma^2\log n.$$

Integrating (B.29) against $r_{t,z}(x)\,dx$ therefore yields

$$\int_{\mathbb{R}} |\partial_t h'_{t,z}(x)|^2 r_{t,z}(x)\,dx \lesssim \frac{|z|^2}{t^4} + \frac{|z|^2 + t + \sigma^2 \log n}{t^4} + \frac{1 + |B_t|^3}{t^3}.$$

Multiplying by t^3, we find

$$t^3 \int_{\mathbb{R}} |\partial_t h'_{t,z}(x)|^2 r_{t,z}(x)\,dx \lesssim 1 + \frac{|z|^2}{t} + \frac{\sigma^2 \log n}{t} + |B_t|^3.$$

Finally, since $t \geq \tau$ and $|B_t| \lesssim \log(n\tau^{-1})$, the constant 1 is absorbed into the logarithmic term, and we conclude that

$$t^3 \int_{\mathbb{R}} |\partial_t h'_{t,z}(x)|^2 r_{t,z}(x)\,dx \lesssim \Big(\frac{|z|^2}{\tau} + \frac{\sigma^2}{\tau} \log n + \log^3(n\tau^{-1}) \Big).$$

This is exactly (B.17). □

Now we return to the proof of Lemma 19.

Proof of Lemma 19. Using the product representation

$$\Pi_t^y = \bigotimes_{i=1}^n \rho_{t,y_i},$$

we may write

$$\log \Pi_t^y(x) = \sum_{i=1}^n h_{t,y_i}(x_i) + C_t(y),$$

where $C_t(y)$ is independent of x. Hence

$$\nabla \log \Pi_t^y(x) = \big(h'_{t,y_1}(x_1), \dots, h'_{t,y_n}(x_n) \big),$$

$$\nabla^2 \log \Pi_t^y(x) = \operatorname{diag}\big(h''_{t,y_1}(x_1), \dots, h''_{t,y_n}(x_n) \big),$$

and $\nabla^3 \log \Pi_t^y(x)$ is the diagonal order-three tensor with entries $h'''_{t,y_i}(x_i)$ on the coordinate axes. Likewise,

$$\partial_t \nabla \log \Pi_t^y(x) = \big(\partial_t h'_{t,y_1}(x_1), \dots, \partial_t h'_{t,y_n}(x_n) \big).$$

It follows that

$$\|\nabla^2 \log \Pi_t^y(x)\| = \max_{1\le i\le n} |h''_{t,y_i}(x_i)|, \qquad \|\nabla^3 \log \Pi_t^y(x)\| = \max_{1\le i\le n} |h'''_{t,y_i}(x_i)|,$$

and therefore

$$\mathsf{L}_2 \le \sup_{t\in[\tau,T]} \sup_{z\in\mathbb{R}} \sup_{x\in\mathbb{R}} t\,|h''_{t,z}(x)| \lesssim \log(n\tau^{-1}),$$

$$\mathsf{L}_3 \le \sup_{t\in[\tau,T]} \sup_{z\in\mathbb{R}} \sup_{x\in\mathbb{R}} t^{3/2}|h'''_{t,z}(x)| \lesssim \log^{3/2}(n\tau^{-1}),$$

by (B.14) and (B.15).

For the Fisher information term, Fubini and the product structure yield

$$\begin{aligned} t\int_{\mathbb{R}^n} \|\nabla \log \Pi_t^y(x)\|^2\, \Pi_t^y(dx) &= t\sum_{i=1}^n \int_{\mathbb{R}^n} |h'_{t,y_i}(x_i)|^2\, \Pi_t^y(dx) \\ &= t\sum_{i=1}^n \int_{\mathbb{R}} |h'_{t,y_i}(u)|^2\, \rho_{t,y_i}(du) \\ &\le 2n \end{aligned}$$

by (B.16). Taking the supremum over $t \in [\tau, T]$ gives $\mathsf{M}_1 \le 2n$.

Similarly,

$$\begin{aligned} t^3\int_{\mathbb{R}^n} \|\partial_t \nabla \log \Pi_t^y(x)\|^2\, \Pi_t^y(dx) &= t^3\sum_{i=1}^n \int_{\mathbb{R}} |\partial_t h'_{t,y_i}(u)|^2\, \rho_{t,y_i}(du) \\ &\lesssim \sum_{i=1}^n \Big(\frac{|y_i|^2}{\tau} + \frac{\sigma^2}{\tau}\log n + \log^3(n\tau^{-1})\Big) \\ &= \frac{\|y\|_2^2}{\tau} + \frac{n\sigma^2}{\tau}\log n + n\log^3(n\tau^{-1}), \end{aligned}$$

by (B.17). Taking the supremum over $t \in [\tau, T]$ proves the bound on H_1. □

B.3 Proof of initialization error

Proof of Lemma 7. Let

$$a_T := 2T + 16\sigma^2 \log n, \qquad b_0(T) = T, \qquad b_1(T) = T + \frac{1}{s}.$$

Recall that Π_T^y factorizes across coordinates:

$$\Pi_T^y = \bigotimes_{i=1}^{n} \rho_{T,y_i},$$

where for each $z \in \mathbb{R}$,

$$\rho_{T,z} = (1 - \omega_{T,z})\, g_{0,z} + \omega_{T,z}\, g_{1,z},$$

and

$$g_{j,z} = \mathcal{N}(m_j z, r_j^2), \qquad m_j = \frac{b_j(T)}{a_T + b_j(T)}, \qquad r_j^2 = \frac{a_T b_j(T)}{a_T + b_j(T)}, \qquad j \in \{0, 1\}.$$

On the other hand, by definition of the initialization,

$$x_0 \sim \mathcal{N}\left(0, \frac{2T}{3} I_n\right).$$

If we denote

$$\gamma := \mathcal{N}\left(0, \frac{2T}{3}\right),$$

then $\mathsf{Law}(x_0) = \gamma^{\otimes n}$. By tensorization of relative entropy,

$$\mathsf{KL}(\Pi_T^y \| \mathsf{Law}(x_0)) = \sum_{i=1}^{n} \mathsf{KL}(\rho_{T,y_i} \| \gamma).$$

It therefore remains to show that, uniformly for every $z \in \mathbb{R}$,

$$\mathsf{KL}(\rho_{T,z} \| \gamma) \lesssim \frac{1}{T^2} + \frac{z^2}{T}.$$

Since $\rho_{T,z}$ is a convex combination of $g_{0,z}$ and $g_{1,z}$, convexity of KL divergence in the first argument gives

$$\mathsf{KL}(\rho_{T,z} \| \gamma) \le (1 - \omega_{T,z}) \mathsf{KL}(g_{0,z} \| \gamma) + \omega_{T,z} \mathsf{KL}(g_{1,z} \| \gamma).$$

Thus it is enough to bound $\mathsf{KL}(g_{j,z} \| \gamma)$ for $j \in \{0, 1\}$.

Under the assumptions of Theorem 3, the quantity $16\sigma^2 \log n$ is bounded by a

universal constant. Hence

$$a_T = 2T + O(1), \qquad b_j(T) = T + O(1), \qquad j \in \{0, 1\},$$

with constants independent of j. It follows that

$$a_T + b_j(T) = 3T + O(1),$$

and therefore

$$m_j = \frac{b_j(T)}{a_T + b_j(T)} = \frac{T + O(1)}{3T + O(1)} = \frac{1}{3} + O(T^{-1}),$$

while

$$r_j^2 = \frac{a_T b_j(T)}{a_T + b_j(T)} = \frac{(2T + O(1))(T + O(1))}{3T + O(1)} = \frac{2T}{3} + O(1).$$

Now recall the formula for the relative entropy between one-dimensional Gaussians:

$$\mathsf{KL}\big(\mathcal{N}(\mu, v^2) \,\|\, \mathcal{N}(0, \sigma^2)\big) = \frac{1}{2}\left(\frac{v^2}{\sigma^2} - 1 - \log\frac{v^2}{\sigma^2} + \frac{\mu^2}{\sigma^2}\right).$$

Applying this with $\mu = m_j z$, $v^2 = r_j^2$, and $\sigma^2 = 2T/3$, we obtain

$$\mathsf{KL}(g_{j,z} \| \gamma) = \frac{1}{2}\left(\frac{r_j^2}{2T/3} - 1 - \log\frac{r_j^2}{2T/3} + \frac{m_j^2 z^2}{2T/3}\right).$$

Because $r_j^2 = \frac{2T}{3} + O(1)$, we have

$$\frac{r_j^2}{2T/3} = 1 + O(T^{-1}).$$

For quantities of the form $1 + u$ with u small, one has

$$(1 + u) - 1 - \log(1 + u) = u - \log(1 + u) = O(u^2),$$

hence

$$\frac{r_j^2}{2T/3} - 1 - \log\frac{r_j^2}{2T/3} \lesssim \frac{1}{T^2}.$$

Similarly, since $m_j = \frac{1}{3} + O(T^{-1})$, we have $m_j^2 \lesssim 1$, and therefore

$$\frac{m_j^2 z^2}{2T/3} \lesssim \frac{z^2}{T}.$$

Combining the last two estimates yields

$$\mathsf{KL}(g_{j,z} \| \gamma) \lesssim \frac{1}{T^2} + \frac{z^2}{T}, \qquad j \in \{0, 1\}.$$

Returning to the mixture and using the convexity bound,

$$\mathsf{KL}(\rho_{T,z} \| \gamma) \lesssim \frac{1}{T^2} + \frac{z^2}{T}.$$

Finally, summing over coordinates gives

$$\mathsf{KL}(\Pi_T^y \| \mathsf{Law}(x_0)) = \sum_{i=1}^n \mathsf{KL}(\rho_{T,y_i} \| \gamma) \lesssim \sum_{i=1}^n \left(\frac{1}{T^2} + \frac{y_i^2}{T} \right) = \frac{n}{T^2} + \frac{\|y\|_2^2}{T}.$$

This proves the lemma. □

B.4 Proof of localization

Proof of Lemma 8. Recall

$$a_\tau := 2\tau + 16\sigma^2 \log n.$$

Since $\tau \le \sigma^2$, we have

$$a_\tau \asymp \sigma^2 \log n, \qquad \sqrt{a_\tau} \asymp \sigma \sqrt{\log n}, \qquad a_\tau \ge 2\tau.$$

We define

$$\mathcal{G} := \left\{ \|\zeta\|_\infty \le C\sqrt{\log n}, \quad \|x_\star\|_\infty \le C\sqrt{\tfrac{\log n}{s}}, \quad \min_{i \in \mathrm{supp}(x_\star)} |x_\star(i)| \ge \frac{c}{s^{3/2}} \right\}.$$

We first verify that the event $\mathcal{G}$ has overwhelming probability. The noise vector ζ has i.i.d. standard Gaussian coordinates, so by a union bound and the Gaussian tail

estimate,

$$\mathbb{P}\big(\|\zeta\|_\infty > C\sqrt{\log n}\big) \le 2n\exp\Big(-\frac{C^2\log n}{2}\Big),$$

which is at most 10^{-4} for all $n \ge 2$ if $C > 0$ is chosen sufficiently large.

Next, by the model for $x_\star$, each coordinate is either zero or Gaussian with variance $1/s$. Hence again by a union bound,

$$\mathbb{P}\Big(\|x_\star\|_\infty > C\sqrt{\tfrac{\log n}{s}}\Big) \le 2n\exp\Big(-\frac{C^2\log n}{2}\Big),$$

which is also at most 10^{-4} for C large enough.

Finally, for the lower bound on the nonzero coordinates, note that

$$\mathbb{P}\Big(\exists i \in \mathrm{supp}(x_\star) :\ |x_\star(i)| < \frac{c}{s^{3/2}}\Big) \le \sum_{i=1}^n \mathbb{P}\Big(x_\star(i) \neq 0,\ |x_\star(i)| < \frac{c}{s^{3/2}}\Big).$$

Since $\mathbb{P}(x_\star(i) \neq 0) = \lambda = s/n$ and, conditional on $x_\star(i) \neq 0$, $x_\star(i) \sim \mathcal{N}(0, 1/s)$, we get

$$\mathbb{P}\Big(x_\star(i) \neq 0,\ |x_\star(i)| < \frac{c}{s^{3/2}}\Big) = \frac{s}{n}\,\mathbb{P}\Big(|G| < \frac{c}{s}\Big), \qquad G \sim \mathcal{N}(0,1).$$

Using $\mathbb{P}(|G| < u) \lesssim u$ for $u \in (0,1)$, we obtain

$$\mathbb{P}\Big(\exists i \in \mathrm{supp}(x_\star) :\ |x_\star(i)| < \frac{c}{s^{3/2}}\Big) \lesssim c.$$

Thus, after choosing $c > 0$ sufficiently small and $C > 0$ sufficiently large, we indeed have

$$\mathbb{P}(\mathcal{G}) \ge 0.999.$$

We now work on the event $\mathcal{G}$. Since

$$\Pi_\tau^y = \bigotimes_{i=1}^n \rho_{\tau,y_i},$$

it is enough to analyze a single coordinate $X_i \sim \rho_{\tau,y_i}$.

For each $z \in \mathbb{R}$, the one-dimensional posterior at time τ admits the Gaussian-mixture representation established earlier:

$$\rho_{\tau,z} = \omega_0(\tau,z)\,g_{0,\tau,z} + \omega_1(\tau,z)\,g_{1,\tau,z},$$

where $g_{j,\tau,z}$ is the density of a Gaussian $\mathcal{N}(\mu_j(z), v_j)$, with

$$\mu_j(z) = \frac{b_j(\tau)}{a_\tau + b_j(\tau)}\, z, \qquad v_j = \frac{a_\tau b_j(\tau)}{a_\tau + b_j(\tau)}, \qquad j \in \{0, 1\},$$

and

$$b_0(\tau) = \tau, \qquad b_1(\tau) = \tau + \frac{1}{s}.$$

Moreover, the mixture weights satisfy

$$\frac{\omega_1(\tau, z)}{\omega_0(\tau, z)} = \frac{\lambda}{1-\lambda}\sqrt{\frac{a_\tau + b_0(\tau)}{a_\tau + b_1(\tau)}}\exp\left(\frac{z^2}{2}\Big(\frac{1}{a_\tau + b_0(\tau)} - \frac{1}{a_\tau + b_1(\tau)}\Big)\right).$$

We treat the off-support and on-support coordinates separately.

Case 1: $i \notin \mathrm{supp}(x_\star)$**.** Then $x_\star(i) = 0$, and therefore on $\mathcal{G}$,

$$|y_i| = |\sigma\zeta_i| \le C\sigma\sqrt{\log n} \lesssim \sqrt{a_\tau}.$$

Also,

$$a_\tau + b_0(\tau) = a_\tau + \tau \asymp a_\tau, \qquad a_\tau + b_1(\tau) = a_\tau + \tau + \frac{1}{s} \ge \frac{1}{s}.$$

Since the exponential factor is increasing in $|z|$ and $|y_i|^2 \lesssim a_\tau$, we have

$$0 \le \frac{y_i^2}{2}\Big(\frac{1}{a_\tau + b_0(\tau)} - \frac{1}{a_\tau + b_1(\tau)}\Big) \le \frac{y_i^2}{2(a_\tau + b_0(\tau))} \lesssim 1.$$

Consequently,

$$\frac{\omega_1(\tau, y_i)}{\omega_0(\tau, y_i)} \lesssim \frac{\lambda}{1-\lambda}\sqrt{\frac{a_\tau + b_0(\tau)}{a_\tau + b_1(\tau)}} \lesssim \frac{s}{n}\sqrt{s a_\tau}.$$

Under the small-noise assumption (4.9), we further have (by taking c_σ sufficiently small)

$$\omega_1(\tau, y_i) \le \frac{1}{3000\, n}.$$

Now consider the $j = 0$ component. Its variance is

$$v_0 = \frac{\tau a_\tau}{\tau + a_\tau} \le \tau,$$

and its mean satisfies

$$|\mu_0(y_i)| = \frac{\tau}{\tau + a_\tau}|y_i| \lesssim \frac{\tau}{\sqrt{a_\tau}} \le \sqrt{\tau},$$

because $a_\tau \ge \tau$. Therefore, for a sufficiently large universal constant M, a standard Gaussian tail bound gives

$$\mathbb{P}\Big(|X_i| > M\sqrt{\tau \log n} \;\Big|\; X_i \sim g_{0,\tau,y_i}\Big) \le \frac{1}{3000\, n}.$$

Combining this with the bound on the exceptional mixture weight, we obtain

$$\mathbb{P}\Big(|X_i| > M\sqrt{\tau \log n} \;\Big|\; \mathcal{G}\Big) \le \frac{1}{1500\, n}, \qquad i \notin S.$$

Case 2: $i \in \mathrm{supp}(x_\star)$**.** On $\mathcal{G}$ we have

$$|x_\star(i)| \ge \frac{c}{s^{3/2}}, \qquad |y_i - x_\star(i)| = |\sigma\zeta_i| \lesssim \sqrt{a_\tau}.$$

By the standing small-noise assumption (4.9), $\sqrt{a_\tau}$ is much smaller than $s^{-3/2}$; hence, after possibly shrinking the universal constant c_σ there, we may assume

$$|y_i| \ge \frac{c}{2s^{3/2}}.$$

We now show that the posterior overwhelmingly favors the $j = 1$ component. Using

$$a_\tau + b_0(\tau) \asymp a_\tau, \qquad a_\tau + b_1(\tau) \asymp \frac{1}{s},$$

we get

$$\sqrt{\frac{a_\tau + b_0(\tau)}{a_\tau + b_1(\tau)}} \asymp \sqrt{s a_\tau}.$$

Moreover,

$$\frac{1}{a_\tau + b_0(\tau)} - \frac{1}{a_\tau + b_1(\tau)} \gtrsim \frac{1}{a_\tau},$$

because $a_\tau \ll 1/s$ in the small-noise regime (4.9). Since $|y_i| \gtrsim s^{-3/2}$, it follows that

$$\frac{y_i^2}{2}\Big(\frac{1}{a_\tau + b_0(\tau)} - \frac{1}{a_\tau + b_1(\tau)}\Big) \gtrsim \frac{1}{s^3 a_\tau}.$$

Thus

$$\frac{\omega_1(\tau, y_i)}{\omega_0(\tau, y_i)} \gtrsim \frac{s}{n}\sqrt{sa_\tau}\,\exp\Big(\frac{c'}{s^3 a_\tau}\Big).$$

Again by the small-noise assumption (4.9), the exponential factor dominates the prefactor by an arbitrarily large power of n. In particular, after shrinking the universal constant c_σ in that assumption if necessary, we may ensure

$$\omega_0(\tau, y_i) \le \frac{1}{(20n)^3}.$$

Under the dominant $j = 1$ component, we have

$$X_i \sim \mathcal{N}(\mu_1(y_i), v_1), \qquad v_1 = \frac{(\tau + 1/s)a_\tau}{a_\tau + \tau + 1/s} \le a_\tau.$$

For the mean, we compute

$$|\mu_1(y_i) - y_i| = \frac{a_\tau}{a_\tau + b_1(\tau)}\,|y_i| \le sa_\tau |y_i|.$$

On $\mathcal{G}$ we also have

$$|y_i| \le |x_\star(i)| + |\sigma\zeta_i| \lesssim \sqrt{\frac{\log n}{s}} + \sqrt{a_\tau}.$$

Hence

$$|\mu_1(y_i) - y_i| \lesssim sa_\tau\sqrt{\frac{\log n}{s}} + sa_\tau^{3/2} = \sqrt{a_\tau}\,\sqrt{sa_\tau \log n} + sa_\tau^{3/2}.$$

Both terms are $O(\sqrt{a_\tau})$ in the small-noise regime (4.9), so

$$|\mu_1(y_i) - y_i| \lesssim \sqrt{a_\tau}.$$

Since also $y_i = x_\star(i) + O(\sqrt{a_\tau})$ on $\mathcal{G}$, we conclude that

$$\mu_1(y_i) = x_\star(i) + O(\sqrt{a_\tau}).$$

Because $v_1 \le a_\tau$, another Gaussian tail estimate yields, for a sufficiently large

universal constant M,

$$\mathbb{P}\Big(|X_i - x_\star(i)| > M\sqrt{a_\tau \log n} \;\Big|\; X_i \sim g_{1,\tau,y_i}\Big) \le \frac{1}{(20n)^3}.$$

Taking into account the negligible weight of the $j = 0$ component, we obtain

$$\mathbb{P}\Big(|X_i - x_\star(i)| > M\sqrt{a_\tau \log n} \;\Big|\; \mathcal{G}\Big) \le \frac{2}{(20n)^3}, \qquad i \in S.$$

Finally, since $\sqrt{a_\tau \log n} \lesssim \sigma \log n$, this becomes

$$\mathbb{P}\Big(|X_i - x_\star(i)| \lesssim \sigma \log n \;\Big|\; \mathcal{G}\Big) \ge 1 - \frac{2}{(20n)^3}, \qquad i \in S.$$

We now combine the two coordinatewise bounds. For inactive coordinates,

$$\mathbb{P}\Big(\exists i \notin S:\ |X_i| > M\sqrt{\tau \log n} \;\Big|\; \mathcal{G}\Big) \le n \cdot \frac{1}{1500\, n} \le \frac{1}{1500}.$$

For active coordinates,

$$\mathbb{P}\Big(\exists i \in S:\ |X_i - x_\star(i)| > M\sqrt{a_\tau \log n} \;\Big|\; \mathcal{G}\Big) \le n \cdot \frac{2}{(20n)^3} \le \frac{2}{8000\, n^2}.$$

Therefore, on $\mathcal{G}$, with probability at least

$$1 - \frac{1}{1500} - \frac{2}{8000\, n^2} \ge 0.999,$$

we have simultaneously for all $i \in [n]$,

$$|X(i) - x_\star(i)| \lesssim \begin{cases} \sigma \log n, & i \in S, \\ \sqrt{\tau \log n}, & i \notin S. \end{cases}$$

This proves the lemma. □

B.5 Proof of auxiliary lemmas

The following lemma is well-known; we provide a proof here for sake of completeness.

Lemma 20 (One-dimensional Gaussian Hardy inequality)**.** *Let*

$$q_{m,v}(x) \coloneqq \frac{1}{\sqrt{2\pi v}} \exp\Big(-\frac{(x-m)^2}{2v}\Big), \qquad x \in \mathbb{R},$$

be the density of $\mathcal{N}(m, v)$*, where* $m \in \mathbb{R}$ *and* $v > 0$*. Let* $H : \mathbb{R} \to \mathbb{R}$ *be absolutely continuous, assume that*

$$H(\pm\infty) = 0,$$

and

$$\int_{\mathbb{R}} \frac{H(x)^2}{q_{m,v}(x)}\,dx < \infty, \qquad \int_{\mathbb{R}} \frac{H'(x)^2}{q_{m,v}(x)}\,dx < \infty.$$

Then

$$\int_{\mathbb{R}} \frac{H(x)^2}{q_{m,v}(x)}\,dx \le v \int_{\mathbb{R}} \frac{H'(x)^2}{q_{m,v}(x)}\,dx.$$

In particular, if

$$H(x) = \int_{-\infty}^{x} h(y)\,dy, \qquad \int_{\mathbb{R}} h(y)\,dy = 0,$$

and

$$\int_{\mathbb{R}} \frac{h(x)^2}{q_{m,v}(x)}\,dx < \infty,$$

then

$$\int_{\mathbb{R}} \frac{H(x)^2}{q_{m,v}(x)}\,dx \le v \int_{\mathbb{R}} \frac{h(x)^2}{q_{m,v}(x)}\,dx.$$

Proof. Write $q \coloneqq q_{m,v}$. Define

$$G(x) \coloneqq \frac{H(x)}{q(x)}.$$

Then $H = qG$, and therefore

$$H' = q'G + qG' = q\Big(G' + \frac{q'}{q}G\Big).$$

Since

$$\frac{q'(x)}{q(x)} = -\frac{x-m}{v},$$

we obtain

$$H'(x) = q(x)\Big(G'(x) - \frac{x-m}{v}G(x)\Big).$$

Hence

$$\frac{H'(x)^2}{q(x)} = q(x)\Big(G'(x) - \frac{x-m}{v}G(x)\Big)^2.$$

Integrating over $\mathbb{R}$ and expanding the square yields

$$\int_{\mathbb{R}} \frac{H'(x)^2}{q(x)}\,dx = \int_{\mathbb{R}} q(x)\,G'(x)^2\,dx - 2\int_{\mathbb{R}} q(x)\frac{x-m}{v}G(x)G'(x)\,dx + \int_{\mathbb{R}} q(x)\frac{(x-m)^2}{v^2}G(x)^2\,dx.$$

Using $2GG' = (G^2)'$, the middle term becomes

$$-2\int_{\mathbb{R}} q(x)\frac{x-m}{v}G(x)G'(x)\,dx = -\int_{\mathbb{R}} q(x)\frac{x-m}{v}(G(x)^2)'\,dx.$$

We now integrate by parts. The boundary term vanishes because

$$q(x)\frac{x-m}{v}G(x)^2 = \frac{x-m}{v}\,\frac{H(x)^2}{q(x)},$$

and the assumptions imply that this quantity tends to 0 as $x \to \pm\infty$ along a sequence, hence the usual truncation argument is valid. Therefore

$$-\int_{\mathbb{R}} q(x)\frac{x-m}{v}(G(x)^2)'\,dx = \int_{\mathbb{R}}\Big(q(x)\frac{x-m}{v}\Big)'G(x)^2\,dx.$$

A direct computation gives

$$\Big(q(x)\frac{x-m}{v}\Big)' = \frac{q(x)}{v} + \frac{x-m}{v}q'(x) = \frac{q(x)}{v} - \frac{(x-m)^2}{v^2}q(x).$$

Substituting this identity above, we obtain

$$-2\int_{\mathbb{R}} q(x)\frac{x-m}{v}G(x)G'(x)\,dx = \int_{\mathbb{R}}\left(\frac{q(x)}{v} - \frac{(x-m)^2}{v^2}q(x)\right)G(x)^2\,dx.$$

Combining the last two displays, the $(x-m)^2$ terms cancel and we arrive at

$$\int_{\mathbb{R}} \frac{H'(x)^2}{q(x)}\,dx = \int_{\mathbb{R}} q(x)\,G'(x)^2\,dx + \frac{1}{v}\int_{\mathbb{R}} q(x)\,G(x)^2\,dx.$$

Since the first term on the right-hand side is nonnegative,

$$\int_{\mathbb{R}} \frac{H'(x)^2}{q(x)}\,dx \geq \frac{1}{v}\int_{\mathbb{R}} q(x)\,G(x)^2\,dx.$$

Finally, because $H = qG$, we have

$$q(x)\,G(x)^2 = \frac{H(x)^2}{q(x)},$$

and therefore

$$\int_{\mathbb{R}} \frac{H(x)^2}{q(x)}\,dx \leq v\int_{\mathbb{R}} \frac{H'(x)^2}{q(x)}\,dx.$$

This proves the first claim.

For the second claim, if

$$H(x) = \int_{-\infty}^{x} h(y)\,dy \qquad \text{and} \qquad \int_{\mathbb{R}} h(y)\,dy = 0,$$

then H is absolutely continuous, $H' = h$ a.e., and

$$H(+\infty) = \int_{\mathbb{R}} h(y)\,dy = 0, \qquad H(-\infty) = 0.$$

Applying the first part with $H' = h$ gives

$$\int_{\mathbb{R}} \frac{H(x)^2}{q_{m,v}(x)}\,dx \leq v\int_{\mathbb{R}} \frac{h(x)^2}{q_{m,v}(x)}\,dx.$$

The proof is complete. □

Appendix C

Proofs for Chapter 5

We present the proofs of the results in Chapter 5.

C.1 Preliminaries

We first introduce some tools we use in the rest of the proof.

Lemma 21 (Pinsker's inequality, Polyanskiy and Wu [93])**.** *For any two probability distributions p, q on $\mathcal{M}$, we have*

$$\mathsf{TV}(p, q) \le \sqrt{2\mathsf{KL}(p \parallel q)}.$$

Lemma 22. *Let v be a vector field on $\mathcal{M}$. In a local coordinate on $\mathcal{M}$, we have*

$$\partial_\alpha v_\beta = (\nabla_\alpha v)^\beta - \Gamma^\beta_{\alpha\gamma} v_\gamma.$$

Here $\Gamma^\beta_{\alpha\gamma}$ is the Christoffel symbol, defined as

$$\Gamma^\beta_{\alpha\gamma} = \frac{1}{2} g^{\beta\delta}(\partial_\alpha g_{\gamma\delta} + \partial_\gamma g_{\alpha\delta} - \partial_\delta g_{\alpha\gamma}).$$

Lemma 23 (Metric distortion in normal coordinates)**.** *There exist coefficients $c, C > 0$ polynomial in d and constant in other parameters, such that the following holds. Let $x \in \mathcal{M}$. In the normal coordinates (∂_α) at x, for any $y \in \mathcal{M}$ such that*

$\rho(x,y) \le c/K$, *we have*

$$\begin{aligned} \|g(y) - I\| &\le CKd^2(x,y), \\ \|\partial_\alpha g_{\beta\gamma}\| &\le CK\rho(x,y), \\ \|\partial_{\alpha\beta} g_{\gamma\xi}\| &\le CK. \end{aligned}$$

Proof. This is a quantitative version of the well-known Taylor expansion of g in normal coordinates (cf. Berline et al. [8, Proposition 1.28]):

$$g_{\alpha\beta}(\exp_x(u)) = \delta_{\alpha\beta} - \frac{1}{3} R_{\alpha\gamma\beta\xi}(x) u^\gamma u^\xi + O\big((\|\mathrm{Rm}\| + \|\nabla \mathrm{Rm}\|)\|u\|^3\big), \quad \|u\| \le c/K.$$

Let $(e_\alpha)_{\alpha=1}^d$ be an orthonormal basis of $T_x\mathcal{M}$; normal coordinates at x are defined by identifying a point $\exp_x(z^\alpha e_\alpha)$ with its coordinate vector (z^α). Denote by ∂_α the coordinate vector fields.

Step I: Representation by Jacobi fields. Define the geodesic segment with *unit parameter* $s \in [0,1]$:

$$\gamma(s) := \exp_x(sv), \qquad \gamma(0) = x, \quad \gamma(1) = y, \quad \dot\gamma(s) = \frac{d}{ds}\gamma(s).$$

Then $\nabla_s \dot\gamma = 0$ and $|\dot\gamma(s)| \equiv |v| = r$.

Fix an orthonormal basis $(e_\alpha)_{\alpha=1}^d$ of $T_x\mathcal{M}$ and parallel transport it along γ to obtain an orthonormal frame $(E_\alpha(s))$ along γ:

$$\nabla_s E_\alpha(s) = 0, \qquad E_\alpha(0) = e_\alpha.$$

Let $J_\beta(s)$ be the Jacobi field along γ corresponding to varying the initial point in direction e_β in the normal coordinate chart, i.e.

$$J_\beta(s) := \mathrm{d}(\exp_x)_{sv}(se_\beta) \in T_{\gamma(s)}\mathcal{M}.$$

Equivalently, J_β is the unique Jacobi field solving

$$\nabla_s^2 J_\beta + \mathrm{Rm}(J_\beta, \dot\gamma)\dot\gamma = 0, \qquad J_\beta(0) = 0, \quad \nabla_s J_\beta(0) = e_\beta. \tag{C.1}$$

Note that in normal coordinates, $\partial_\beta|_x = e_\beta$ and the geodesic variation $\exp_x(s(v+\varepsilon e_\beta))$ yields (C.1).

Write $J_\beta(s)$ in the parallel frame:

$$J_\beta(s) = \sum_{\alpha=1}^{d} \mathsf{J}_{\alpha\beta}(s)\, E_\alpha(s),$$

and let $\mathsf{J}(s) \in \mathbb{R}^{d\times d}$ be the matrix with entries $\mathsf{J}_{\alpha\beta}(s)$. Since E_α is parallel, (C.1) becomes the *matrix Jacobi equation*

$$\mathsf{J}''(s) + \mathsf{R}(s)\,\mathsf{J}(s) = 0, \qquad \mathsf{J}(0) = 0, \quad \mathsf{J}'(0) = I, \tag{C.2}$$

where the curvature matrix $\mathsf{R}(s)$ is defined by

$$\big(\mathsf{R}(s)\,u\big)_\alpha := \Big\langle \mathrm{Rm}\Big(\sum_{\mu} u_\mu E_\mu(s), \dot\gamma(s)\Big)\dot\gamma(s),\, E_\alpha(s)\Big\rangle.$$

From $\|\mathrm{Rm}\| \le K$ and $|\dot\gamma| = r$ we have

$$\|\mathsf{R}(s)\| \le K\,|\dot\gamma(s)|^2 = Kr^2, \qquad s\in[0,1]. \tag{C.3}$$

In normal coordinates, the coordinate vector fields at $y = \gamma(1)$ are

$$\partial_\beta|_y = \mathrm{d}(\exp_x)_v(e_\beta) = J_\beta(1).$$

Since the frame at $s=1$ is orthonormal, the metric coefficients are

$$g_{\beta\gamma}(y) = \langle \partial_\beta|_y, \partial_\gamma|_y\rangle = \langle J_\beta(1), J_\gamma(1)\rangle = \sum_{\alpha=1}^{d} \mathsf{J}_{\alpha\beta}(1)\,\mathsf{J}_{\alpha\gamma}(1) = \big(\mathsf{J}(1)^{\top}\mathsf{J}(1)\big)_{\beta\gamma}.$$

Hence, as matrices,

$$g(y) = \mathsf{J}(1)^{\top}\mathsf{J}(1). \tag{C.4}$$

Step II: Control of $\mathsf{J}(1) - I$ via Grönwall inequality. From (C.2), integrating twice and using $\mathsf{J}(0) = 0,\ \mathsf{J}'(0) = I$, we get the exact Volterra equation

$$\mathsf{J}(s) = sI - \int_0^s (s-\tau)\,\mathsf{R}(\tau)\,\mathsf{J}(\tau)\,d\tau, \qquad s \in [0,1]. \tag{C.5}$$

Taking operator norms and using (C.3) gives for $s \in [0,1]$:

$$\|\mathsf{J}(s)\| \le s + Kr^2 \int_0^s (s-\tau)\,\|\mathsf{J}(\tau)\|\,d\tau.$$

A standard Grönwall argument yields

$$\|\mathsf{J}(s)\| \le Cs \quad \text{and} \quad \|\mathsf{J}'(s)\| \le C \qquad \text{for all } s \in [0,1],\ \text{provided } r \le c/\sqrt{K}. \tag{C.6}$$

Now subtract sI in (C.5):

$$\mathsf{J}(s) - sI = -\int_0^s (s-\tau)\,\mathsf{R}(\tau)\,\mathsf{J}(\tau)\,d\tau.$$

Using (C.3) and (C.6),

$$\|\mathsf{J}(1) - I\| \le \int_0^1 (1-\tau)\,\|\mathsf{R}(\tau)\|\,\|\mathsf{J}(\tau)\|\,d\tau \le C\int_0^1 (1-\tau)\,(Kr^2)\,\tau\,d\tau \le CKr^2.$$

Combine with (C.4):

$$g(y) - I = \big(\mathsf{J}(1)^\top \mathsf{J}(1) - I\big) = \big(\mathsf{J}(1) - I\big)^\top + \big(\mathsf{J}(1) - I\big) + \big(\mathsf{J}(1) - I\big)^\top \big(\mathsf{J}(1) - I\big),$$

so

$$\|g(y) - I\| \le C\|\mathsf{J}(1) - I\| \le CKr^2. \tag{C.7}$$

Step III: Control of first derivatives. We first control $\partial_\alpha \mathsf{J}(1)$ as a function of the coordinate v. Let $v \mapsto \mathsf{J}_v(s)$ denote the Jacobi matrix for the geodesic $\gamma_v(s) = \exp_x(sv)$. Differentiate the ODE (C.2) w.r.t. v^α:

$$\partial_\alpha \mathsf{J}'' + \mathsf{R}\,\partial_\alpha \mathsf{J} = -(\partial_\alpha \mathsf{R})\,\mathsf{J}, \qquad \partial_\alpha \mathsf{J}(0) = 0, \quad \partial_\alpha \mathsf{J}'(0) = 0. \tag{C.8}$$

Bound on $\partial_\alpha \mathsf{R}$. Recall $\mathsf{R}(s)$ represents the operator $u \mapsto \mathrm{Rm}(u, \dot\gamma)\dot\gamma$ in the parallel

frame. Varying v changes both γ and $\dot\gamma$; the corresponding variation field $V_\alpha(s) := \partial_\alpha \gamma_v(s)$ along γ is itself a Jacobi field with $V_\alpha(0) = 0,\ \nabla_s V_\alpha(0) = e_\alpha$, hence by the same estimate as (C.6)

$$\|V_\alpha(s)\| \le Cs, \qquad \|\nabla_s V_\alpha(s)\| \le C. \tag{C.9}$$

Using the product rule on $\mathrm{Rm}(\cdot,\dot\gamma)\dot\gamma$ and our Assumption 2, together with $|\dot\gamma| = r$ and (C.9), one obtains the uniform operator bound

$$\begin{aligned}\|\partial_\alpha \mathsf{R}(s)\| &\le C\Big(\|\nabla \mathrm{Rm}\|\,\|V_\alpha(s)\|\,|\dot\gamma|^2 + \|\mathrm{Rm}\|\,|\dot\gamma|\,\|\partial_\alpha \dot\gamma(s)\|\Big)\\ &\le C\big(K\cdot s\cdot r^2 + K\cdot r\cdot 1\big)\\ &\le CKr,\end{aligned} \tag{C.10}$$

for all $s \in [0,1]$ (since $s \le 1$). Here we used $\partial_\alpha \dot\gamma = \nabla_s V_\alpha$.

Now solve (C.8) by the same Duhamel principle: integrating twice with zero initial data gives

$$\partial_\alpha \mathsf{J}(s) = -\int_0^s (s-\tau)\Big(\mathsf{R}(\tau)\,\partial_\alpha \mathsf{J}(\tau) + (\partial_\alpha \mathsf{R})(\tau)\,\mathsf{J}(\tau)\Big)\,d\tau. \tag{C.11}$$

Using (C.3), (C.10), and (C.6), we obtain

$$\begin{aligned}\|\partial_\alpha \mathsf{J}(s)\| \le& Kr^2\int_0^s (s-\tau)\,\|\partial_\alpha \mathsf{J}(\tau)\|\,\mathrm{d}\tau + CKr\int_0^s (s-\tau)\,\|\mathsf{J}(\tau)\|\,d\tau\\ \le& Kr^2\int_0^s (s-\tau)\,\|\partial_\alpha \mathsf{J}(\tau)\|\,\mathrm{d}\tau + CKr\,s^3.\end{aligned}$$

Apply the same Grönwall comparison as before (now with a forcing term $CKr\,s^3$) to conclude, for $r \le c/\sqrt{K}$,

$$\|\partial_\alpha \mathsf{J}(1)\| \le CKr. \tag{C.12}$$

Finally differentiate $g = \mathsf{J}^\top \mathsf{J}$:

$$\partial_\alpha g = (\partial_\alpha \mathsf{J})^\top \mathsf{J} + \mathsf{J}^\top(\partial_\alpha \mathsf{J}),$$

so by (C.6) (at $s = 1$) and (C.12),

$$\|\partial_\alpha g(y)\| \le 2\|\partial_\alpha \mathsf{J}(1)\|\,\|\mathsf{J}(1)\| \le C(Kr)\cdot 1 \le CKr. \tag{C.13}$$

Step IV: Control of second derivatives. Differentiate (C.8) once more:

$$\partial_{\alpha\beta}\mathsf{J}'' + \mathsf{R}\,\partial_{\alpha\beta}\mathsf{J} = -(\partial_{\alpha\beta}\mathsf{R})\,\mathsf{J} - (\partial_\alpha \mathsf{R})\,\partial_\beta \mathsf{J} - (\partial_\beta \mathsf{R})\,\partial_\alpha \mathsf{J}, \tag{C.14}$$

with zero initial data at $s = 0$.

Bound on $\partial_{\alpha\beta}\mathsf{R}$. Under Assumption 2, $\partial_{\alpha\beta}\mathsf{R}$ can be bounded *uniformly by* CK on $[0,1]$ as follows: expanding the second parameter derivative of $\mathrm{Rm}(\cdot,\dot\gamma)\dot\gamma$ produces terms of the schematic form

$$(\nabla\mathrm{Rm})(V)\cdot\dot\gamma\cdot(\partial\dot\gamma), \qquad \mathrm{Rm}(\cdot,\partial\dot\gamma)\cdot(\partial\dot\gamma), \qquad (\nabla\mathrm{Rm})(\partial V)\cdot\dot\gamma\cdot\dot\gamma,$$

and also terms involving $\nabla_s(\partial V)$, all of which are controlled using $\|V\| \lesssim 1$, $\|\nabla_s V\| \lesssim 1$ and the fact that each appearance of $\dot\gamma$ contributes a factor r. Concretely, one shows (using (C.9) for both V_α, V_β and the same Jacobi estimates for their derivatives) that

$$\|\partial_{\alpha\beta}\mathsf{R}(s)\| \le CK \qquad \text{for all } s\in[0,1]. \tag{C.15}$$

Now apply Duhamel's principle to (C.14) with zero initial data:

$$\partial_{\alpha\beta}\mathsf{J}(s) = -\int_0^s (s-\tau)\Big(\mathsf{R}\,\partial_{\alpha\beta}\mathsf{J} + (\partial_{\alpha\beta}\mathsf{R})\,\mathsf{J} + (\partial_\alpha\mathsf{R})\,\partial_\beta\mathsf{J} + (\partial_\beta\mathsf{R})\,\partial_\alpha\mathsf{J}\Big)(\tau)\,d\tau.$$

Take norms and use (C.3), (C.15), (C.6), (C.10), (C.12):

$$\begin{aligned}\|\partial_{\alpha\beta}\mathsf{J}(s)\| \le{}& Kr^2\int_0^s (s-\tau)\,\|\partial_{\alpha\beta}\mathsf{J}(\tau)\|\,d\tau \\ &+ CK\int_0^s (s-\tau)\,\|\mathsf{J}(\tau)\|\,d\tau + C(Kr)(Kr)\int_0^s (s-\tau)\,d\tau.\end{aligned}$$

Since $\|\mathsf{J}(\tau)\| \le C\tau$, the second integral is bounded by CKs^3, and the third is bounded by $CK^2r^2s^2 \le CKs^2$ provided $r \le c/\sqrt{K}$. Thus for $s \le 1$,

$$\|\partial_{\alpha\beta}\mathsf{J}(s)\| \le Kr^2\int_0^s (s-\tau)\,\|\partial_{\alpha\beta}\mathsf{J}(\tau)\|\,d\tau + CK.$$

Applying Grönwall argument once more yields

$$\|\partial_{\alpha\beta}\mathsf{J}(1)\| \le CK. \tag{C.16}$$

Finally differentiate $g = \mathsf{J}^\top \mathsf{J}$ twice:

$$\partial_{\alpha\beta} g = (\partial_{\alpha\beta}\mathsf{J})^\top \mathsf{J} + \mathsf{J}^\top(\partial_{\alpha\beta}\mathsf{J}) + (\partial_\alpha \mathsf{J})^\top(\partial_\beta \mathsf{J}) + (\partial_\beta \mathsf{J})^\top(\partial_\alpha \mathsf{J}).$$

Hence by (C.6), (C.12), (C.16) (and $K^2 r^2 \le CK$ for $r \le c/\sqrt{K}$),

$$\|\partial_{\alpha\beta} g(y)\| \le C\|\partial_{\alpha\beta}\mathsf{J}(1)\| \cdot \|\mathsf{J}(1)\| + C\|\partial_\alpha \mathsf{J}(1)\|\,\|\partial_\beta \mathsf{J}(1)\| \le CK + C(Kr)^2 \le CK. \tag{C.17}$$

This completes the proof. □

Lemma 24. *Fix $x \in \mathcal{M}$. Define*

$$J(x,u) \;:=\; \big|\det \mathrm{d}\exp_x(u)\big| \;=\; \sqrt{\det g_{ij}(\exp_x u)}.$$

There exist universal constants $c, C > 0$, such that for $u \in T_x\mathcal{M}$ with $\|u\| \le \frac{c}{Kd}$, we have the following bound on $J(x,u)$:

$$\Big|J(x,u) - 1\Big| \le CdK\|u\|^2. \tag{C.18}$$

In particular, we have

$$\frac{1}{2} \le J(x,u) \le 2, \quad \|u\| \le \frac{c}{Kd}.$$

Proof. Work in normal coordinates at x so that $\exp_x : B_{\mathrm{euc}}(0, 1/K) \subset T_x\mathcal{M} \to B_{\mathrm{geo}}(x, 1/K)$ is a diffeomorphism and $g_{ij}(0) = \delta_{ij}$, $\Gamma^k_{ij}(0) = 0$.

From Lemma 23, we know that $\|g(\exp_x u) - I\| \le CK\|u\|^2$. In the region $\|u\| \le \frac{c}{Kd}$, we have

$$\|g(\exp_x u) - I\| \le \frac{c}{d}.$$

Therefore, by Taylor expansion of determinants, we know

$$\begin{aligned}|\det g(\exp_x u) - 1| &= |\det(I + g(\exp_x u) - I) - 1| \\ &\le C \operatorname{tr}(g(\exp_x u) - I) \\ &\le Cd \cdot \|g(\exp_x u) - I\| \\ &\le Cd \cdot CK\|u\|^2.\end{aligned}$$

This concludes the proof by adjusting C if necessary. □

The metric distortion bound implies that geodesic is almost a straight line, in a sufficiently small normal neighborhood. The following quantitative bound shall be useful.

Lemma 25 (Geodesics are almost straight in small balls)**.** *There exist coefficients $c, C > 0$ polynomial in d and constant in other parameters, such that the following holds. Fix any $x \in \mathcal{M}$ and let $0 < r \le c/K$. Let $y, z \in B_x(r)$ and γ be the unit-speed geodesic connecting y to z. Write*

$$y(s) \coloneqq \exp_x^{-1}(\gamma(s)) \in T_x\mathcal{M} \simeq \mathbb{R}^d$$

for its representation in normal coordinates at x. Then:

(i) (Almost constant velocity)

$$\sup_{s\in[0,\ell]} \left|\dot{y}(s) - \dot{y}(0)\right| \;\le\; CKr^2.$$

(ii) (Almost linear trajectory)

$$\sup_{s\in[0,\ell]} \left|y(s) - y(0) - s\,\dot{y}(0)\right| \;\le\; CKr^3.$$

In words, in normal coordinates at x, any geodesic segment contained in $B_x(r)$ deviates from the Euclidean line segment connecting its endpoints by at most $O(Kr^3)$ in position and $O(Kr^2)$ in direction.

Proof. Work in normal coordinates at x. By bounded geometry and the choice of r,

the metric coefficients satisfy, in view of Lemma 23, that

$$\|g(y) - I\| \le CK|y|^2, \qquad \|\partial g(y)\| \le CK|y|, \qquad |y| \le r,$$

which implies the Christoffel symbols obey

$$|\Gamma(y)| \le CK|y| \le CKr.$$

The coordinate representation $y(s)$ of the geodesic satisfies the geodesic equation

$$\ddot{y}^k(s) + \Gamma^k_{ij}(y(s))\, \dot{y}^i(s)\dot{y}^j(s) = 0.$$

Since γ is unit–speed and $g(y)$ is uniformly equivalent to the Euclidean metric on $|y| \le r$, we have $|\dot{y}(s)| \asymp 1$. Consequently,

$$|\ddot{y}(s)| \le CKr \qquad \text{for all } s \in [0, \ell].$$

Integrating once gives

$$|\dot{y}(s) - \dot{y}(0)| \le \int_0^s |\ddot{y}(u)|\, \mathrm{d}u \le CKr\, s \le CKr^2,$$

proving (i). Integrating again yields

$$|y(s) - y(0) - s\dot{y}(0)| \le \int_0^s \int_0^u |\ddot{y}(w)|\, \mathrm{d}w\, \mathrm{d}u \le CKr\, s^2 \le CKr^3,$$

which proves (ii). □

We spell out the constants in a few classical inequalities in geometric analysis.

Lemma 26 (Schoen and Yau [105, Thm. 4.6])**.** *Let $(\mathcal{M}, g)$ be a complete Riemannian manifold with $\mathrm{Ric}(\mathcal{M}) \ge -K$ for some $K \ge 0$. Let $H(x, y, t)$ be the heat kernel, i.e., the fundamental solution of $(\Delta - \frac{\partial}{\partial t})u(x,t) = 0$. Then, for every $\delta_{\mathrm{Sch}} > 0$ and $\alpha > 1$,*

$$H(t,x,y) \;\le\; C(\delta_{\mathrm{Sch}}, d, \alpha)\, V_x\big(\sqrt{t}\big)^{-1/2} V_y\big(\sqrt{t}\big)^{-1/2} \exp\left[-\frac{r^2(x,y)}{(4+\delta_{\mathrm{Sch}})t} + C_1\, \delta_{\mathrm{Sch}} K t\right],$$

where $V_x(R) = \mu(B_x(R))$, $C(\delta_{\mathrm{Sch}}, d, \alpha) = (1+\delta_{\mathrm{Sch}})^{d\alpha} \exp\big(\frac{1+\alpha}{\delta_{\mathrm{Sch}}}\big)$, and $C_1 = \frac{\alpha d}{\alpha - 1}$.

To unleash the power of Lemma 26, we need the following lower bound on volume of geodesic balls assuming bounded geometry.

Lemma 27 (Günther's comparison theorem)**.** *Under Assumption 2, we have*

$$V_x(r) \geq \frac{(2\pi)^{d/2}}{\Gamma(d/2)} \int_0^r \left(\frac{\sin(t\sqrt{K})}{\sqrt{K}} \right)^{d-1} \mathrm{d}t, \quad 0 \leq r \leq 1/K.$$

Here Γ is the Gamma function. In particular, when $r \leq c/K$ for some small universal constant $c > 0$, we have

$$V_x(r) \geq \frac{\pi^{d/2}}{d\Gamma(d/2)} r^d \geq \frac{1}{d^{d/2}} r^d.$$

Proof. We observe that $\|\mathrm{Rm}\| \leq K$ implies that the sectional curvature is upper bounded by K, since by definition $\sec(u,v) = \mathrm{Rm}(u,v,u,v)$. The first inequality follows from the classical form of Günther's comparison theorem, see for example Gray [40, Theorem 3.17]. The second inequality follows from the elementary bound that $\sin(x) \geq \frac{1}{2}x$ for $x \in [0, c]$, where c is a small universal constant, and the crude bound $\Gamma(x) \leq x^{x-1}$ for $x \geq 1$. □

A complementary lower bound to Lemma 26 is as follows.

Lemma 28 (Li and Xu [73, Thm. 1.5])**.** *Let $(\mathcal{M}, g)$ be complete, possibly with $\mathrm{Ric}(\mathcal{M}) \geq -K$. For the (Neumann) heat kernel $H(x,y,t)$ and all $x, y \in \mathcal{M}$, $t > 0$,*

$$\begin{aligned} H(t,x,y) &\geq (4\pi t)^{-d/2} \frac{(2Kt)^{d/2}}{(\mathrm{e}^{2Kt} - 2Kt - 1)^{d/4}} \exp\left[-\frac{\rho(x,y)^2}{4t} \left(1 + \frac{Kt\coth(Kt) - 1}{Kt} \right) \right], \\ H(t,x,y) &\geq (4\pi t)^{-d/2} \exp\left[-\frac{\rho(x,y)^2}{4t} \left(1 + \tfrac{1}{3}Kt \right) - \frac{d}{4}Kt \right]. \end{aligned} \tag{C.19}$$

The above bounds for heat kernel translates seamlessly to p_t, since p_t is a convolution of p_0 with $H(t,\cdot,\cdot)$. We formalize this in the following lemma.

Lemma 29. *We have*

$$\inf_{x,y\in\mathcal{M}} H(t,x,y) \leq \inf_{x\in\mathcal{M}} p_t(x) \leq \sup_{x\in\mathcal{M}} p_t(x) \leq \sup_{x,y\in\mathcal{M}} H(t,x,y).$$

Proof. This follows from taking infimum and supremum respectively in the formula

(Duhamel principle)

$$p_t(y) = \int_{\mathcal{M}} p_0(x) H(t, x, y) \mu(\mathrm{d}x).$$

□

The following lemma compiles a few follow-ups of Li-Yau estimates [45, 47] with constants made explicit.

Lemma 30. *Under Assumption 2, we have Han-Zhang's inequality*

$$\frac{\nabla^2 p_t}{p_t} \preceq C_{\mathsf{HZ}} \left(\frac{1}{t} + K \right) \left(1 + \log \frac{\sup p_{t/2}}{p_t} \right).$$

On the other hand, we also have Hamilton's Harnack inequality

$$\nabla^2 \log p_t = \frac{\nabla^2 p_t}{p_t} - \frac{(\nabla p_t)(\nabla p_t)^\top}{p_t^2} \succeq -\frac{1}{2t} g - C_{\mathsf{Ham}} \left(1 + \log \frac{\sup p_{t/2}}{p_t} \right) g$$

and

$$\|\nabla \log p_t\|^2 = \frac{\|\nabla p_t\|^2}{p_t^2} \leq C \left(\frac{1}{t} + K \right) \log \frac{\sup p_{t/2}}{p_t}.$$

Here $C > 0$ is a universal constant, $C_{\mathsf{HZ}} = CdK$, $C_{\mathsf{Ham}} = CdK^2$.

Proof. The last inequality follows from Hamilton [45, Theorem 1.1]. The proof of the rest two inequalities require tracing the proofs of Hamilton [45], Han and Zhang [47]. The details would be too tedious to reproduce here, so we leave pointers to relevant proofs for interested readers.

To prove the second inequality, we trace the proof of Hamilton [45, Theorem 4.3] to see that if $A > 0$ is such that

$$\frac{\Delta_{\mathcal{M}} p_t}{p_t} \leq \frac{A}{t} \left(d + \log \frac{\sup p_{t/2}}{p_t} \right)$$

and

$$\frac{\|\nabla p_t\|^2}{p_t^2} \leq \frac{A}{t} \left(d + \log \frac{\sup p_{t/2}}{p_t} \right),$$

then

$$\frac{\nabla^2 p_t}{p_t} - \frac{(\nabla p_t)(\nabla p_t)^\top}{p_t^2} \succeq -\frac{1}{2t} g - CA(\|\mathrm{Rm}\| + \|\nabla \mathrm{Rm}\|) g.$$

Here $C > 0$ is an absolute constant. By Assumption 2, this implies

$$C_{\mathsf{Ham}} \le CKC_{\mathsf{HZ}}.$$

Tracing the proof the main theorem in Han and Zhang [47, Page 9], we can see

$$C_{\mathsf{HZ}} \le C_{\mathsf{LY}}(1+K) = CdK,$$

where C_{LY} is the maximum of the coefficients before t^{-1} and K in Li and Yau [75, Theorem 1.2]. This was explicitly defined as Cd therein, by setting $\alpha = 2$ there. This completes the proof. □

Lemma 31. *For any three points $x, y, z \in \mathcal{M}$, we have*

$$\frac{\rho(x,z)^2}{1-t} + \frac{\rho(z,y)^2}{t} \ge \rho(x,y)^2, \qquad \forall t \in (0,1).$$

The equality is attainable at some point $z_\star$ on the minimum-length geodesic from x to y. Moreover, if x, y, z are within $\iota \le 1/\operatorname{poly}(d, K)$ distance to each other, then the function

$$\psi(z) \coloneqq \frac{\rho(x,z)^2}{1-t} + \frac{\rho(z,y)^2}{t} - \rho(x,y)^2$$

is $\frac{1-Cd^2K^2\iota}{t(1-t)}$-strongly convex in the normal coordinates at x (or y), where $C > 0$ is a universal constant.

Proof. The first inequality follows from Cauchy-Schwarz inequality and triangle inequality:

$$(1-t+t)\left(\frac{\rho(x,z)^2}{1-t} + \frac{\rho(z,y)^2}{t}\right) \ge (\rho(x,z)+\rho(z,y))^2 \ge \rho(x,y)^2.$$

Let $\gamma : [0,1] \to \mathcal{M}$ be the constant-speed, minimum-length geodesic from x to y. It is then straightforward to check that

$$\frac{\rho(x,z_\star)^2}{1-t} + \frac{\rho(z_\star,y)^2}{t} = \rho(x,y)^2, \quad z_\star \coloneqq \gamma(\lambda_\star),$$

where $\lambda_\star \in (0,1)$ solves the quadratic equation

$$\frac{\lambda^2}{1-t} + \frac{(1-\lambda)^2}{t} = 1.$$

Proving the strong convexity requires Lemma 23 and Lemma 24, which implies that $\rho(x,z)^2$ and $\rho(z,y)^2$ are both $(1 - Cd^2K^2\iota)$-strongly convex in the normal coordinates. The desired conclusion then follows from

$$\frac{1}{1-t} + \frac{1}{t} = \frac{1}{t(1-t)}.$$

The proof is completed. □

C.2 Initialization error

We require the following result from Urakawa [119, Proposition 2.6].

Lemma 32. *Denote $H(t,x,y)$ the heat kernel on $\mathcal{M}$. Assume*

$$A \coloneqq \sup_{t\le 1}\sup_{x\in\mathcal{M}} t^{d/2} H(t,x,x).$$

Then for any probability distribution p_0, its evolution along heat flow $\partial_t p_t = \frac{1}{2}\Delta_{\mathcal{M}}$ satisfies

$$\mathsf{TV}(p_t,\mu) \le \sqrt{A}\,\mathrm{e}^{-\frac{1}{2\lambda_1}(t-\frac{1}{2})}, \quad t \ge 1.$$

We combine this with the Li-Yau upper bound (Lemma 26) to obtain

Lemma 33 (First part of Lemma 9)**.** *Under Assumption 2, there exists a universal constant $C > 0$ such that*

$$\mathsf{TV}(p_N,\mu) \le \mathrm{e}^{C(K+d\log(Kd))}\mathrm{e}^{-\frac{1}{2\lambda_1}(T-\frac{1}{2})}.$$

Proof. Plug the bound in Lemma 27 into Lemma 26 and use the fact that the supremum of heat kernel $\sup_{x,y} H(t,x,y)$ is decreasing in t (by convolution inequality),

we obtain

$$\begin{aligned} A &\le (Cd/K)^d \mathrm{e}^{CK} \\ &\le \exp(C'K + C'd\log(Kd)), \end{aligned}$$

for some universal constants $C, C' > 0$. Note that we absorbed $d \log K$ into $K + d \log d$. The desired claim follows. □

C.3 Score matching error

We now prove the second inequality in Lemma 9.

Lemma 34. *Under the same assumptions as in Theorem 4, we have*

$$\sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \mathbb{E}\|\mathscr{S}_{t_k, Y_{t_k}}(Y_t) - \mathscr{S}^{\star}_{t_k, Y_{t_k}}(Y_t)\|^2 \mathrm{d}t \le 2\varepsilon^2_{\mathsf{score}}.$$

Proof. This is relatively straightforward. Notice that in normal coordinates, by Lemma 23, we have

$$\begin{aligned} \|\mathscr{S}_{t_k, Y_{t_k}}(Y_t) - \mathscr{S}^{\star}_{t_k, Y_{t_k}}(Y_t)\|^2 &= g_{\alpha\beta}(Y_t)(\hat{s}_t - \nabla \log p_t)^{\alpha}(\hat{s}_{t_k}(Y_{t_k}) - \nabla \log p_{t_k}(Y_{t_k}))^{\beta} \\ &\le \|g(Y_t)\| \cdot \|\hat{s}_{t_k}(Y_{t_k}) - \nabla \log p_{t_k}(Y_{t_k})\|^2 \\ &\le 2\|\hat{s}_{t_k}(Y_{t_k}) - \nabla \log p_{t_k}(Y_{t_k})\|^2 \end{aligned}$$

for $\rho(Y_t, Y_{t_k}) \le c/K$, and is 0 otherwise due to our cutoff η_ω. Therefore

$$\mathbb{E}\|\mathscr{S}_{t_k, Y_{t_k}}(Y_t) - \mathscr{S}^{\star}_{t_k, Y_{t_k}}(Y_t)\|^2 \le 2\mathbb{E}\|\hat{s}_{t_k}(Y_{t_k}) - \nabla \log p_{t_k}(Y_{t_k})\|^2,$$

and consequently,

$$\begin{aligned} &\sum_{k=1}^{N} \int_{t_{k-1}}^{t_k} \mathbb{E}\|\mathscr{S}_{t_k, Y_{t_k}}(Y_t) - \mathscr{S}^{\star}_{t_k, Y_{t_k}}(Y_t)\|^2 \mathrm{d}t \\ &\le 2\sum_{k=1}^{N} (t_k - t_{k-1}) \mathbb{E}\|\hat{s}_{t_k}(Y_{t_k}) - \nabla \log p_{t_k}(Y_{t_k})\|^2 = 2\varepsilon^2_{\mathsf{score}}. \end{aligned}$$

This proves the claim, as desired. □

C.4 Discretization error

Lemma 35. *Under the same assumptions as in Theorem 4 and assuming (5.3) without loss of generality, there is a universal constant $C > 0$ such that for $t_k - h \le t \le t_k$, we have*

$$\mathbb{E}\|\nabla \log p_t(Y_t) - \mathscr{S}^{\star}_{t_k, Y_{t_k}}(Y_t)\|^2 \le \frac{C d^6 K^8}{t^3}(t_k - t).$$

Proof. For convenience, set the reverse time

$$\tau := t_k - t.$$

The main challenge is that Li-Yau estimates provide sharp uniform control up to second-order derivatives of $\log p_t$, but a naïve calculation of the difference $\nabla \log p_t(Y_t) - \mathscr{S}^{\star}_{t_k, Y_{t_k}}(Y_t)$ involves third-order derivatives. More precisely, a straightforward Taylor expansion will introduce a factor of $\partial_\tau \nabla \log p_t$, which, by reverse-time heat equation $\partial_\tau p_t = -\frac{1}{2}\Delta_{\mathcal{M}} p_t$, contains third-order derivatives of p_t. We bypass this difficulty by making use of Itô's calculus to show that third-order derivatives cancel out; this is inspired by Benton et al. [7], where a similar strategy was employed to the Euclidean setting.

Step I. Applying Itô/Stratonovish formula. We first compute $\partial_\tau \nabla \log p_t$. Since the forward heat equation is $\partial_t p_t = \frac{1}{2}\Delta_{\mathcal{M}} p_t$, we have

$$\partial_\tau \nabla \log p_t \;=\; -\partial_t \nabla \log p_t = -\nabla\left(\frac{\partial_t p_t}{p_t}\right) = -\frac{1}{2}\nabla\left(\frac{\Delta_{\mathcal{M}} p_t}{p_t}\right).$$

Use the manifold quotient rule and the identity $\Delta_{\mathcal{M}} \log p_t = \frac{\Delta_{\mathcal{M}} p_t}{p_t} - \|\nabla \log p_t\|^2$ to rewrite

$$\frac{\Delta_{\mathcal{M}} p_t}{p_t} = \Delta_{\mathcal{M}} \log p_t + \|\nabla \log p_t\|^2.$$

Therefore, we have

$$\begin{aligned}\partial_\tau \nabla \log p_t &= -\frac{1}{2}\nabla\left(\frac{\Delta_{\mathcal{M}} p_t}{p_t}\right) = -\frac{1}{2}\nabla\Delta_{\mathcal{M}} \log p_t - \frac{1}{2}\nabla\|\nabla \log p_t\|^2 \\ &= -\frac{1}{2}\nabla\Delta_{\mathcal{M}} \log p_t - \nabla^2 \log p_t \cdot \nabla \log p_t.\end{aligned}$$

On the other hand, it is straightforward to calculate

$$\nabla^2 \log p_t = \frac{\nabla^2 p_t}{p_t} - \frac{(\nabla p_t)(\nabla p_t)^\top}{p_t^2}.$$

Now, from Itô's formula, we know

$$\begin{aligned}
\mathrm{d}\nabla \log p_t(Y_t) &= (\partial_\tau \nabla \log p_t)(Y_t)\mathrm{d}\tau + \nabla^2 \log p_t(Y_t)(\nabla \log p_t(Y_t)\mathrm{d}\tau + U_{Y_t} \circ \mathrm{d}W_t) \\
&\quad + \frac{1}{2}\Delta_{\mathcal{M}} \nabla \log p_t(Y_t)\mathrm{d}\tau \\
&= -\frac{1}{2}\nabla \Delta_{\mathcal{M}} \log p_t \mathrm{d}\tau - \nabla^2 \log p_t \cdot \nabla \log p_t \mathrm{d}\tau + \nabla^2 \log p_t \cdot \nabla \log p_t \mathrm{d}\tau \\
&\quad + \nabla^2 \log p_t \cdot U_{Y_t} \circ \mathrm{d}W_t + \frac{1}{2}\Delta_{\mathcal{M}} \nabla \log p_t \mathrm{d}\tau \\
&= \frac{1}{2}(\Delta_{\mathcal{M}} \nabla - \nabla \Delta_{\mathcal{M}}) \log p_t \mathrm{d}\tau + \nabla^2 \log p_t \cdot U_{Y_t} \circ \mathrm{d}W_t \\
&= \frac{1}{2}\operatorname{Ric}^\sharp(\nabla \log p_t, \cdot)\mathrm{d}\tau + \nabla^2 \log p_t \cdot U_{Y_t} \circ \mathrm{d}W_t,
\end{aligned}$$

where the last line follows from Bochner's identity $(\Delta_{\mathcal{M}} \nabla - \nabla \Delta_{\mathcal{M}}) f = \operatorname{Ric}^\sharp(\nabla f, \cdot)$, and $\operatorname{Ric}^\sharp$ denotes the $(1,1)$-tensor obtained by raising one index in Ricci curvature. Notice here the cancellation of third-order derivatives.

On the other hand,

$$\mathrm{d}\mathscr{S}^\star_{t_k, Y_{t_k}}(Y_t) = \nabla \mathscr{S}^\star_{t_k, Y_{t_k}}(Y_t) \cdot (\nabla \log p_t \mathrm{d}\tau + U_{Y_t} \circ \mathrm{d}W_t) + \frac{1}{2}\Delta_{\mathcal{M}} \mathscr{S}^\star_{t_k, Y_{t_k}}(Y_t)\mathrm{d}\tau.$$

In normal coordinates, $\mathscr{S}^\star_{t_k, Y_{t_k}}$ is a constant vector field inside $B(0, \omega/3)$, therefore we have (cf. Lemma 22):

$$\nabla_\alpha \mathscr{S}^\star_{t_k, Y_{t_k}}(Y_t)^\beta = \Gamma^\beta_{\alpha\gamma} \mathscr{S}^\star_{t_k, Y_{t_k}}(Y_t)^\gamma = \Gamma^\beta_{\alpha\gamma} \nabla^\gamma \log p_{t_k}(Y_{t_k}), \quad \rho(Y_t, Y_{t_k}) \le \omega/3,$$

and similarly, when $\rho(Y_t, Y_{t_k}) \le \omega/3$, we have

$$\begin{aligned}
\Delta_{\mathcal{M}} \mathscr{S}^\star_{t_k, Y_{t_k}}(Y_t)^\alpha = &- \operatorname{Ric}^\alpha{}_\beta \nabla^\beta \log p_{t_k}(Y_{t_k}) \\
&- g^{\beta\gamma}\left(\partial_\gamma \Gamma^\alpha_{\beta\xi} + \Gamma^\alpha_{\gamma\zeta}\Gamma^\zeta_{\beta\xi} + \Gamma^\alpha_{\zeta\xi}\Gamma^\zeta_{\beta\gamma}\right) \nabla^\xi \log p_{t_k}(Y_{t_k}).
\end{aligned}$$

Step II. Bounding the coefficients. Combine the above formulas with the estimates given in Lemma 23, we obtain for some universal constant $C > 0$:

$$\begin{aligned}
\left\|\mathrm{Ric}^\sharp(\nabla \log p_t, \cdot)\right\| &\le CK\|\nabla \log p_t(Y_t)\|, \\
\left\|\nabla \mathscr{S}^\star_{t_k, Y_{t_k}}\right\| &\le CK\|\nabla \log p_{t_k}(Y_{t_k})\|, \qquad \rho(Y_t, Y_{t_k}) \le \omega/2, \\
\left\|\Delta_{\mathcal{M}} \mathscr{S}^\star_{t_k, Y_{t_k}}\right\| &\le CK^2\|\nabla \log p_{t_k}(Y_{t_k})\|, \qquad \rho(Y_t, Y_{t_k}) \le \omega/2.
\end{aligned}$$

Outside the geodesic ball $B_{Y_{t_k}}(\omega/2)$, the field $\mathscr{S}^\star_{t_k, Y_{t_k}}$ is non-zero only inside $B_{Y_{t_k}}(\omega)$. Between these two balls, we have to take into account the radial derivative of the cutoff function η_ω, whose first order derivative is bounded by $C\omega^{-1}$ and second order derivative by $C\omega^{-2}$ by our construction of η_ω (recall that $|\eta'| + |\eta''| \le 100$). Apply Lemma 22 and Lemma 23 again, this time we bound

$$\begin{aligned}
\left\|\nabla \mathscr{S}^\star_{t_k, Y_{t_k}}\right\| &\le CK\omega^{-1}\|\nabla \log p_{t_k}(Y_{t_k})\|, \\
\left\|\Delta_{\mathcal{M}} \mathscr{S}^\star_{t_k, Y_{t_k}}\right\| &\le CK^2\omega^{-2}\|\nabla \log p_{t_k}(Y_{t_k})\|.
\end{aligned}$$

We now apply Itô's formula on manifold (i.e., take expectation and invoke the martingale property in (5.1)). We set up some shorthand notation for simplicity. Define

$$\mathcal{E}_t := \mathbb{E}\|\nabla \log p_t(Y_t)\|^2 + \mathbb{E}\|\nabla \log p_{t_k}(Y_{t_k})\|^2 + \mathbb{E}\|\nabla^2 \log p_t(Y_t)\|^2,$$

$$\mathcal{F}_t := \mathbb{E}\|\nabla \log p_t(Y_t)\|^4 + \mathbb{E}\|\nabla \log p_{t_k}(Y_{t_k})\|^4 + \mathbb{E}\|\nabla^2 \log p_t(Y_t)\|^4,$$

and

$$\mathcal{G}_t := \mathbb{E}\left[\left(\|\nabla \log p_t(Y_t)\|^2 + \|\nabla \log p_{t_k}(Y_{t_k})\|^2 + \|\nabla^2 \log p_t(Y_t)\|^2\right) \mathbf{1}_{\{\rho(Y_t, Y_{t_k}) > \omega/3\}}\right].$$

Then apply Itô's formula and collect the above bounds to see (cf. Benton et al. [7])

$$\begin{aligned}
&\left|\frac{\mathrm{d}}{\mathrm{d}\tau}\mathbb{E}\left\|\nabla\log p_t(Y_t)-\mathscr{S}^{\star}_{t_k,Y_{t_k}}(Y_t)\right\|^2\right| \\
&\le CK^2\mathcal{E}_t+CK^2\omega^{-2}\mathcal{G}_t \\
&\le CK^2\mathcal{E}_t+CK^2\omega^{-2}\sqrt{\mathcal{F}_t}\sqrt{\mathbb{P}(\rho(Y_t,Y_{t_k})>\omega/3)} \\
&\le CK^2d\left(\frac{1}{t}+dK^2\right)^2\sup_{t\le s\le t_k}\sqrt{\mathbb{E}\log^4\frac{\sup p_{s/2}}{p_s(Y_s)}}\left(1+\omega^{-2}\sqrt{\mathbb{P}(\rho(Y_t,Y_{t_k})>\omega/3)}\right).
\end{aligned} \tag{C.20}$$

Here the last line follows from Lemma 30.

Step III. Controlling expectations via Chebyshev and Li-Yau estimates. To bound $\mathbb{E}\log^4\frac{\sup p_{s/2}}{p_s(Y_s)}$, we note that

$$\mathbb{E}\left(\frac{1}{p_t(Y_t)}\right)=\int\frac{1}{p_t}p_t\mathrm{d}\mu=\int\mathrm{d}\mu=1.$$

By Chebyshev's inequality, we have

$$\mathbb{P}\left(\frac{1}{p_t(Y_t)}\ge\lambda\right)\le\lambda^{-1},\quad\lambda>0,$$

and then

$$\mathbb{P}\left(\log^4\frac{1}{p_t(Y_t)}\ge\lambda\right)\le\mathrm{e}^{-\sqrt[4]{\lambda}},\quad\lambda\ge0.$$

Integrate with respect to λ, we see

$$\mathbb{E}\log^4\frac{1}{p_t(Y_t)}\le C.$$

We then apply Li-Yau's estimate (Lemma 26) combined with Lemma 27, Lemma 29 to obtain $\sup\log p_{s/2}\le\sup\log H(s/2,x,y)\lesssim d\log\frac{d}{s}+Ks$, where the first inequality follows from $p_{s/2}$ being the convolution of p_0 with $H(s/2,x,y)$. These together shows

$$\mathbb{E}\log^4\frac{\sup p_{s/2}}{p_s(Y_s)}\le\left(Cd\log\frac{d}{s}+CKs\right)^4\le\frac{Cd^5K^4}{\delta}.$$

Here we used $\log \frac{d}{s} \le C\left(\frac{d}{s}\right)^{1/4}$, and $s \ge t \ge \delta$.

Step IV. Controlling exit probability via stopping time. It remains to bound the probability $\mathbb{P}(\rho(Y_t, Y_{t_k}) > \omega/3)$. This would follow from a stopping time argument. We claim that given $t_k - t \le h$, we have

$$\mathbb{P}(\rho(Y_t, Y_{t_k}) > \omega/3) \le \exp\left(-\frac{c\omega^2}{t_k - t}\right) \le \omega^4. \tag{C.21}$$

where the last inequality follows from (5.3). Plug this back into the desired conclusion of the lemma is proved.

We now prove (C.21). Let σ be the largest $t \le t_k$ such that $\rho(Y_t, Y_{t_k}) > \omega/3$. We have

$$\mathbb{P}(\rho(Y_t, Y_{t_k}) > \omega/3) \le \mathbb{P}(\sigma \ge t).$$

In the interval $[\sigma, t_k]$, Y_t stays in the geodesic ball $B_{Y_{t_k}}(\omega/3)$, and follows the SDE (5.2). In normal coordinates, this can be spelled out explicitly:

$$\begin{aligned} \mathrm{d}Y_t^\alpha &= \nabla^\alpha \log p_t(Y_t)\mathrm{d}t + A_\beta^\alpha(Y_t) \circ \mathrm{d}W_t^\beta \\ &= \left(\nabla^\alpha \log p_t(Y_t) + \frac{1}{2}(\partial_\gamma A_\beta^\alpha)A^{\beta\gamma}\right)\mathrm{d}t + A_\beta^\alpha \mathrm{d}W_t^\beta, \end{aligned} \tag{C.22}$$

where A is the square root of the matrix representing the coefficients of the Laplace-Beltrami operator

$$\frac{1}{\sqrt{\det g}}\partial_\alpha(\sqrt{\det g}\cdot g^{\alpha\beta}\partial_\beta).$$

It can be checked with the help of Lemma 23 that $\|A - I\| \le CdK\omega^2$, and $\|\partial_\alpha A\| \le Cd^2K\omega$; we omit the computation that has a similar pattern as many of the previous arguments. Furthermore, we have $\|\nabla \log p_t(Y_t)\| \lesssim (\delta^{-1} + K)\log \frac{\sup_{t/2} p_t}{p_t}$ by the same argument via Lemma 30 as before. This time we combine the uniform bound provided by Lemma 28 with Lemma 26 to conclude $\log \frac{\sup p_{t/2}}{p_t} \lesssim (\delta^{-1} + K + d\log d)^2 \operatorname{Diam}(\mathcal{M})^2 \lesssim (\delta^{-1} + K + d\log d)^2 K^2$ by Assumption 2. This shows

$$\sup \|\nabla \log p_t\| \lesssim (\delta^{-1} + K + d\log d)^3 K^2. \tag{C.23}$$

Therefore, in view of Lemma 23 to convert the above bound to normal coordinates, and together with the aforementioned bound for A and $\partial_\alpha A$, we see that the drift

term up to time σ will not exceed

$$\sup\left\|\nabla^\alpha \log p_t(Y_t) + \frac{1}{2}(\partial_\gamma A^\alpha_\beta)A^{\beta\gamma}\right\| \cdot (t_k - \sigma) \le C(\delta^{-1} + K + d\log d)^3 K^2 (t_k - \sigma) \le \frac{\omega}{12}, \tag{C.24}$$

where the last inequality used (5.3). On the other hand, the bound on A implies that the quadratic variation of the martingale part does not exceed

$$\int_\sigma^{t_k} A^\alpha_\gamma A^\gamma_\beta \mathrm{d}t \preceq 2(t_k - \sigma) I.$$

By Burkholder-Davis-Gundy inequality [97], the tail of $\int_\sigma^{t_k} A^\alpha_\beta \mathrm{d}W_t^\beta$ is $O(1)$-subgaussian, thus we have

$$\mathbb{P}\left(\sigma \ge t,\ \left\|\int_\sigma^{t_k} A^\alpha_\beta \mathrm{d}W_t^\beta\right\| > \frac{\omega}{12}\right) \le \exp\left(\frac{-c(\omega - 2\sqrt{\rho(t_k - t)})^2}{t_k - t}\right).$$

In view of (5.3), combine this with (C.24) and (C.22), we have proved (C.21) as claimed. □

Lemma 36. *Under the same assumptions as in Theorem 4 and assuming (5.3) without loss of generality, the discretization error obeys the following upper bound:*

$$\sum_{k=1}^N \int_{t_{k-1}}^{t_k} \mathbb{E}\|\nabla \log p_t(Y_t) - \mathscr{S}^\star_{t_k, Y_{t_k}}(Y_t)\|^2 \mathrm{d}t \le \frac{C d^6 K^8}{\delta^3} h^2 N,$$

where $C > 0$ is a universal constant.

Proof. This follows directly from Lemma 35. □

C.5 Brownian motion simulation error

In this section, we handle the Brownian motion simulation error using the machinery of Minakshisundaram-Pleijel parametrix. A complete introduction to this heavy machinery would require establish a whole system of notation and lemmas in geometric analysis, which is unduly burdensome. We instead refer the interested reader to Berline et al. [8] for a comprehensive treatment, and point to results there whenever needed.

C.5.1 Overview

Our aim is to prove the following lemma.

Lemma 37. *Under the same assumptions as in Theorem 4, and assuming (5.3) without loss of generality, we have*

$$\mathsf{TV}(p_0^{\mathsf{aux}}, q_0^\star) \le \sqrt{hT}\ \mathrm{poly}(d, K, \delta^{-1}).$$

To better explain the idea of the proof, we ignore the rejection sampling procedure in the construction of $\widehat{\mathcal{K}}_k$ temporarily. Our starting point is the observation that by Fokker-Planck equation, $\widehat{\mathcal{K}}_k$ is the heat kernel associated to the Euclidean Laplacian with drift $\mathscr{S}_{t_k, Y_{t_k}}$, in normal coordinates. On the other hand, $\mathcal{K}_k^{\mathsf{aux}}$ is also a heat kernel with the same drift, but associated to the manifold Laplace-Beltrami operator. The following lemma shows that the two solutions coincide up to first order in time, at least in a polynomially small neighborhood of initial point and in a polynomially short time.

Lemma 38. *Let $F^{\mathcal{H}}(t, x, y)$ be the (generalized) heat kernel for the operator $\mathcal{H} = \frac{1}{2}\Delta_{\mathcal{M}} + \langle \mathscr{S}_{t_k, Y_{t_k}}, \nabla \rangle$. Define the Euclidean density*

$$\varphi_t(u; x) \coloneqq \frac{1}{(2\pi t)^{d/2}} \exp\left(-\frac{\|u - \mathscr{S}_{t_k, Y_{t_k}}(x) t\|^2}{2t} \right), \quad u \in T_x\mathcal{M},$$

and let $\Phi(t, x, y)$ be the density of the push-forward by $\exp_x$ of $\eta_\omega \varphi_t(\cdot; x)$ with respect to the volume measure, where η_ω is the cutoff function defined in (5.4). Then there exists polynomial $\mathrm{poly}(d, K)$ with universally constant coefficients, such that for all $0 < t \le \frac{1}{\mathrm{poly}(d, K, \delta^{-1})}$ and for all $\rho(y, Y_{t_k}) \le t^{5/12}$, we have

$$\left| \frac{F^{\mathcal{H}}(t, Y_{t_k}, y)}{\Phi(t, Y_{t_k}, y)} - 1 \right| \le \mathrm{poly}(d, K, \delta^{-1}) t.$$

With Lemma 38 in hand, it is tempting to calculate the KL error with the

following heuristic:

$$\begin{aligned}\mathsf{KL}(p_k^{\mathsf{aux}}\mathcal{K}_k^{\mathsf{aux}} \parallel p_k^{\mathsf{aux}}\widehat{\mathcal{K}}_k) &\lesssim -\mathbb{E}\int F^{\mathcal{H}}(h, Y_{t_k}, \cdot)\log\frac{\Phi(h, Y_{t_k}, \cdot)}{F^{\mathcal{H}}(h, Y_{t_k}, \cdot)} \\ &\lesssim \mathbb{E}\int F^{\mathcal{H}}(h, Y_{t_k}, \cdot)\left(\frac{F^{\mathcal{H}}(h, Y_{t_k}, \cdot)}{\Phi(h, Y_{t_k}, \cdot)} - 1\right)^2 \\ &\lesssim \mathrm{poly}(d, K, \delta^{-1})h^2,\end{aligned}$$

where we ignore the fact that Lemma 38 holds only in a small neighborhood; the first line is post-processing inequality, and the second line stems from the fact that for two distributions p, q, we have

$$\begin{aligned}\int p\log\frac{q}{p} &= \int p\log\left(1 + \frac{q-p}{p}\right) \\ &\geq \int p\left(\frac{q-p}{p} - C\frac{(q-p)^2}{p^2}\right) \\ &= -C\int p\left(\frac{q}{p} - 1\right)^2,\end{aligned}$$

given $\frac{q}{p} - 1$ is sufficiently small, where the last line follows from $\int p = \int q = 1$. From this, we conclude that the accumulated error along N steps is bounded by $\mathrm{poly}(d, K)h^2N = \mathrm{poly}(d, K)hT$, and the desired bound follows from Pinsker's inequality.

Apart from Lemma 38, the above computation is the essence of this proof. The rest of this section is mainly devoted to proving Lemma 38, and then formalizing the above computation by handling exceptional events of exiting the polynomially small neighborhood.

C.5.2 Proof of Lemma 38: a parametrix estimate

We begin the proof of Lemma 38. For simplicity, denote by v^α the normal coordinate representation of $\hat{s}_{t_k}(Y_{t_k})$. Naturally, our initial test solution is the drifted heat kernel, as simulated by our discretized process:

$$\varphi_t(u) \coloneqq \frac{1}{(2\pi t)^{d/2}}\exp\left(-\frac{\|u - vt\|^2}{2t}\right), \quad u \in \mathbb{R}^d.$$

Before we compare this with the manifold heat kernel, there is one subtlety we need to keep in mind. The density φ_t is with respect to the *Lebesgue* measure on $T_x\mathcal{M}$, not with respect to the *volume* on $\mathcal{M}$. We compute and define the corresponding density on $\mathcal{M}$ as follows:

$$\Phi(t,x,y) = \varphi_t(\log_x y)\sqrt{\Delta(x,y)}, \quad \Delta(x,y) \coloneqq \frac{|\det \mathrm{d}\log_x y|^2}{\det g(y)}, \quad y \in B_x(\omega).$$

Here all quantities are computed in normal coordinates. The factor $\Delta(x,y)$ is known as the van Vleck-Morette determinant. From Lemma 23 and Lemma 24, we know that

$$\frac{1}{2} \le \Delta(x,y) \le 2, \quad \text{if } \rho(x,y) \le \frac{c}{Kd}. \tag{C.25}$$

We consider the generalized Laplacian

$$\mathcal{H} \coloneqq \frac{1}{2}\Delta_{\mathcal{M}} + \langle \mathscr{S}_{t_k, Y_{t_k}}, \nabla \rangle.$$

As in Lemma 38, denote by $F^{\mathcal{H}}$ the heat kernel of $\mathcal{H}$ at time $t_k - t_{k-1}$. We also propose an approximation of $F^{\mathcal{H}}$ by

$$\begin{aligned}
\Psi(t,x,y) &\coloneqq G(t,x,y)\exp(\psi(x,y))\sqrt{\Delta(x,y)}, \\
G(t,x,y) &\coloneqq \frac{1}{(2\pi t)^{d/2}}\exp\left(-\frac{\rho^2(x,y)}{2t}\right),
\end{aligned}$$

where for any two point $x, y \in \mathcal{M}$, letting $\gamma : [0,1] \to$ be a constant-speed geodesic connecting x to y, we define

$$\psi(x,y) \coloneqq \int_0^1 \left\langle \mathscr{S}_{t_k, Y_{t_k}}(\gamma(s)),\ \frac{\mathrm{d}}{\mathrm{d}s}\gamma(s) \right\rangle_g \mathrm{d}s.$$

The auxiliary function Ψ bridges $F^{\mathcal{H}}$ and Φ in the following sense. On the one hand, we relate Φ and Ψ with the following lemma:

Lemma 39. *There exists a polynomial* $\mathrm{poly}(d,K)$ *with universally constant coefficients such that the following holds. For any* $0 < r \le \frac{1}{\mathrm{poly}(d,K,\delta^{-1})}$, $0 < t \le \frac{1}{\mathrm{poly}(d,K,\delta^{-1})}$

and for all $x, y \in B_{Y_{t_k}}(r)$*, we have*

$$\left|\frac{\Phi(t,x,y)}{\Psi(t,x,y)} - 1\right| \le \mathrm{poly}(d, K, \delta^{-1})(r^3 + t).$$

On the other hand, we have the following asymptotic expansion:

$$F^{\mathcal{H}}(t,x,y) = \Psi(t,x,y) \cdot \left(1 + \sum_{i=1}^{\infty} t^i u_i(x,y)\right), \quad t \to 0^+,$$

where u_i are smooth functions that can be computed explicitly via a recursive formula [8]. We will not need the formula here, but instead require u_1 and the remainder terms to be bounded properly. Such bounds have been well-established, which we wrap up into the following lemma. Recall the cutoff function η_ω with radius ω defined in (5.4). It is clear we can replace ω with any $\iota > 0$ to define a cutoff η_ι of radius ι.

Lemma 40 (adapted from Berline et al. [8])**.** *Fix a positive* $\iota \le 1/\mathrm{poly}(d, K)$*. There exists a smooth function* $u_1(x,y)$ *on* $\mathcal{M} \times \mathcal{M}$ *such that*

$$\|u_1\|_\infty + \|\nabla_y u_1\|_\infty \le \mathrm{poly}(d, K, \delta^{-1}),$$

and for all $0 < t \le 1/\mathrm{poly}(d, K)$*,* $y \in B_x(\iota)$*, we have*

$$\left|(\partial_t - \mathcal{H})\left[\eta_\iota(\rho(x,y))\Psi(t,x,y)(1 + t u_1(x,y))\right]\right| \le r_\eta(t,x,y) + r_\psi(t,x,y), \tag{C.26}$$

where

$$r_\eta(t,x,y) \le \frac{1}{\iota t}\mathrm{poly}(d,K)\mathbb{1}_{\frac{\iota}{2} \le \rho(x,y) \le \iota}\, G(t,x,y), \tag{C.27a}$$

$$r_\psi(t,x,y) \le t \cdot \mathrm{poly}(d, K, \delta^{-1})\mathbb{1}_{\rho(x,y) \le \iota}\, G(t,x,y). \tag{C.27b}$$

Proof. The inequality on u_1 follows from Theorem 2.26 in Berline et al. [8], with $\mathcal{H}$ the same as our $\mathcal{H}$ and therefore $F = \langle \mathscr{S}_{t_k, Y_{t_k}}, \nabla \rangle$. Note that all the coefficients in $\mathcal{H}$ are bounded in C^2 by $\mathrm{poly}(d,K)(1 + \|\mathscr{S}_{t_k,Y_{t_k}}\|_{C^2(\mathcal{M})})$, which is further bounded by $\mathrm{poly}(d, K, \delta^{-1})$ as we will show momentarily. In fact, by Lemma 23 and Assumption 2,

(A3), we have

$$\begin{aligned}\|\mathscr{S}_{t_k,Y_{t_k}}\|_{C^2(\mathcal{M})} &= \|\mathscr{S}_{t_k,Y_{t_k}}\|_\infty + \|\nabla\mathscr{S}_{t_k,Y_{t_k}}\|_\infty + \|\nabla^2\mathscr{S}_{t_k,Y_{t_k}}\|_\infty \\ &\le \mathrm{poly}(d,K)\,\omega^{-2}(1+\|\nabla\log p_{t_k}(Y_{t_k})\|),\end{aligned}$$

and then we control $\|\nabla\log p_{t_k}(Y_{t_k})\| \le \mathrm{poly}(d,K,\delta^{-1})$ via (C.23), yielding the claimed bound (recall that $\omega^{-1} = \mathrm{poly}(d,K)$ by definition).

We proceed to prove (C.26). We follow the proof of Theorem 2.29, item (iii) in Berline et al. [8], and choose the cutoff function ψ there to be η_ι as defined in (5.4), and with the differential operator B defined in Berline et al. [8], we have

$$\begin{aligned}&\left|(\partial_t - \mathcal{H})\left[\Psi(t,x,y)(1+tu_1(x,y))\right]\right| \\ &\quad\le \underbrace{\frac{1}{\iota t}\,\mathrm{poly}(d)\,G(t,x,y)}_{=:r_\eta,\ \nabla\eta_\iota\ \text{related terms}} + \underbrace{t\cdot G(t,x,y)\cdot\left|(B_y u_2)(x,y)\right|}_{=:r_\psi}.\end{aligned}$$

Here B_y can be viewed as a coordinate-transformed version of $\mathcal{H}$, applied to the variable y (precise definition can be found in the reference), and u_2 is the second order term in the expansion. Similar to the argument we used to bound $\|u_1\|$, in virtue of Theorem 2.26 in Berline et al. [8], we have

$$\|\mathscr{B}_x u_2\|_{L^\infty(\mu)} \le \mathrm{poly}(d,K,\delta^{-1}).$$

The claimed bound follows from combining the above inequalities. □

We now state the Volterra series representation of heat kernel.

Lemma 41 (Volterra series, Theorem 2.23 in Berline et al. [8])**.** *Fix a $\iota > 0$. Let*

$$\begin{aligned}\Psi_1(t,x,y) &\coloneqq \eta_\iota(\rho(x,y))\Psi(t,x,y)(1+tu_1(x,y)),\\ r_1(t,x,y) &\coloneqq (\partial_t - \mathcal{H})\,\Psi_1(t,x,y).\end{aligned}$$

Define the time-space convolution operator $*$ as

$$(f*g)(t,x,y) = \int_0^t\int f(t-s,x,z)g(s,z,y)\mu(\mathrm{d}z)\mathrm{d}s.$$

Then we have

$$F^{\mathcal{H}} = \Psi_1 + \sum_{k=1}^{\infty} (-1)^k \, \Psi_1 * r_1^{*k}, \quad \textit{where} \quad r_1^{*k} := \underbrace{r_1 * \cdots * r_1}_{k \text{ times}},$$

on any domain such that the series on the right hand side converges absolutely uniformly.

Lemma 42 (Iterative bounds for Volterra series)**.** *There exists a polynomial denoted by* $\mathrm{poly}(d, K, \delta^{-1})$ *with universally constant coefficients such that the following holds. Assume* $0 < t \le 1/\mathrm{poly}(d, K.\delta^{-1})$*, take* $\iota = 4t^{5/12}d$ *in the definition of* Ψ*. Then we have, for all* $\rho(x, y) \le t^{5/12} = \iota/(4d)$*, that*

$$\sum_{k=1}^{\infty} \left| \Psi_1 * r_1^{*k} \right| (t, x, y) \le t \cdot \mathrm{poly}(d, K, \delta^{-1}) \, G(t, x, y).$$

Proof. Recall Lemma 40, and denote by P the polynomial factor $\mathrm{poly}(d, K)$ therein. Denote by $\lambda\Delta^k$ the dilated standard simplex

$$\lambda\Delta^k = \{(s_1, \cdots, s_{k+1}) : s_i \ge 0, \sum_{i=1}^{k+1} s_i = \lambda\}, \quad \lambda > 0.$$

Fix some $k \ge 1$. Set $z_0 = x$ and $z_k = y$, we have

$$\begin{aligned} \left| \Psi_1 * r_1^{*k} \right| (t, x, y) \le t^k P^k \int_{t\Delta^{k-1}} \mathrm{d}s \int_{\mathcal{M}^{k-1}} \prod_{i=1}^{k} \Bigg[& G(s_i, z_{i-1}, z_i) \\ & \times \left(\mathbf{1}_{\{\rho(z_{i-1}, z_i) \le \iota\}} + \frac{\mathbf{1}_{\{\rho(z_{i-1}, z_i) > \iota/2\}}}{\iota s_i} \right) \Bigg] \mu^{\otimes(k-1)}(\mathrm{d}z). \end{aligned} \tag{C.28}$$

We split the integral in (C.28) into a *local* part and an *outlier* part. Define a small "local" region

$$\mathcal{R} := \left\{ (z_1, \cdots, z_{k-1}) : \rho(z_i, x) \le 2\iota, \ \rho(z_{i-1}, z_i) \le \iota/2, \ i = 1, \cdots, k \right\}.$$

We further define

$$I_{\mathsf{loc}}(s) \coloneqq \int_{\mathcal{R}} \left(\prod_{i=1}^{k} G(s_i, z_{i-1}, z_i) \right) \mu^{\otimes(k-1)}(\mathrm{d}z),$$
$$I_{\mathsf{out}}(s) \coloneqq \int_{\mathcal{R}^c} \left(\prod_{i=1}^{k} G(s_i, z_{i-1}, z_i) \mathbb{1}_{\rho(z_{i-1},z_i)\le\iota} \left(1 + \frac{1}{\iota s_i} \mathbb{1}_{\rho(z_{i-1},z_i)>\iota/2}\right) \right) \mu^{\otimes(k-1)}(\mathrm{d}z).$$

It is clear that

$$\left|\Psi_1 * r_1^{*k}\right|(t,x,y) \le t^k P^k \int_{t\Delta^{k-1}} (I_{\mathsf{loc}}(s) + I_{\mathsf{out}}(s))\,\mathrm{d}s. \tag{C.29}$$

We will establish bounds for I_{loc} and I_{out} respectively.

Bounding the local integral. For ease of understanding, we begin by computing the first integral in I_{loc} with respect to z_1. Extracting the factors containing z_1, we need to calculate

$$\int_{\{\rho(z_1,x)\le 2\iota\}} \frac{1}{(2\pi s_1)^{d/2}} \frac{1}{(2\pi s_2)^{d/2}} \exp\left(-\frac{\rho(x,z_1)^2}{2s_1} - \frac{\rho(z_1,z_2)^2}{2s_2}\right) \mu(\mathrm{d}z_1). \tag{C.30}$$

To proceed, we will invoke Lemma 31. Let $z_\star$ be a minimizer of

$$V(z) = \frac{\rho(x,z)^2}{s_1(s_1+s_2)^{-1}} + \frac{\rho(z,z_2)^2}{s_2(s_1+s_2)^{-1}} - \rho(x,z_2)^2.$$

By Lemma 31, $V(z_\star) = 0$. Since $V(z) > 0$ whenever $\rho(x,z) \ge \rho(x,z_2)$, we know that $z_\star \in B_x(2\iota)$. Moreover, Lemma 31 and Lemma 23 together imply

$$V(z) \ge (1 - Cd^2K^2\iota) \cdot \frac{(s_1+s_2)^2}{s_1 s_2} \rho(z,z_\star)^2, \quad \forall z \in B_x(4\iota).$$

Here, the second inequality follows from strong convexity given by Lemma 31 and a comparison of geometric distance and Euclidean distance in normal coordinates fueled by Lemma 23. Denote for the moment that

$$\theta \coloneqq 1 - Cd^2K^2\iota.$$

Plugging this back into (C.30), we obtain

$$
\begin{aligned}
&\int_{\{\rho(z_1,x)\le 2\iota\}} \frac{1}{(2\pi s_1)^{d/2}} \frac{1}{(2\pi s_2)^{d/2}} \exp\left(-\frac{\rho(x,z_1)^2}{2s_1} - \frac{\rho(z_1,z_2)^2}{2s_2}\right) \mu(\mathrm{d}z) \\
&= \int_{\{\rho(z_1,x)\le 2\iota\}} \frac{1}{(2\pi s_1)^{d/2}} \frac{1}{(2\pi s_2)^{d/2}} \exp\left(-\frac{V(z_1)}{2(s_1+s_2)}\right) \mu(\mathrm{d}z_1) \\
&\le \int_{\{\rho(z_1,x)\le 2\iota\}} \frac{1}{(2\pi s_1)^{d/2}} \frac{1}{(2\pi s_2)^{d/2}} \exp\left(-\frac{\theta(s_1+s_2)\rho(z_1,z_\star)^2}{2s_1s_2} - \frac{\rho(x,z_2)^2}{2(s_1+s_2)}\right) \mu(\mathrm{d}z_1) \\
&\le 4\exp\left(-\frac{\rho(x,z_2)^2}{2(s_1+s_2)}\right) \cdot \int \frac{1}{(2\pi s_1)^{d/2}} \frac{1}{(2\pi s_2)^{d/2}} \exp\left(-\frac{\theta(s_1+s_2)\|Z\|^2}{2s_1s_2}\right) \mathrm{d}Z \\
&= 4(2\pi(s_1+s_2))^{d/2} G(s_1+s_2,x,z_2) \cdot \frac{1}{(2\pi s_1)^{d/2}} \frac{1}{(2\pi s_2)^{d/2}} \left(2\pi \cdot \frac{s_1s_2}{\theta(s_1+s_2)}\right)^{d/2} \\
&\le 4\theta^{-d/2} G(s_1+s_2,x,z_2),
\end{aligned}
$$

where the third-to-last line follows from change of variable to normal coordinates at $z_\star$ and from using (C.25) to bound the determinant; the penultimate line follows from Gaussian integration. Now, we note that

$$
\theta^{-d/2} = (1 - Cd^2K^2\iota)^{-d/2} \le \exp(2Cd^3K^2\iota) \le 2,
$$

give $\iota \le \frac{1}{100Cd^3K^2}$. Putting these pieces together, we proved

$$
\begin{aligned}
&\int_{\{\rho(z_1,x)\le 2\iota\}} \frac{1}{(2\pi s_1)^{d/2}} \frac{1}{(2\pi s_2)^{d/2}} \exp\left(-\frac{\rho(x,z_1)^2}{2s_1} - \frac{\rho(z_1,z_2)^2}{2s_2}\right) \mu(\mathrm{d}z_1) \\
&\quad \le 8G(s_1+s_2,x,z_2).
\end{aligned}
$$

Iterate the above argument for the integration over $z_2, \cdots, z_{k-1}$ to obtain

$$
I_{\mathsf{loc}}(s) \le 8^k \cdot G\left(\sum_{i=1}^k s_i, x, y\right) = 8^k \cdot G(t,x,y). \tag{C.31}
$$

Bounding the outlier integral. Next, we show how to control I_{out}. We first write

$$
\prod_{i=1}^k G(s_i, z_{i-1}, z_i) = \frac{1}{\prod_{i=1}^k (2\pi s_i)^{d/2}} \exp\left(-\sum_{i=1}^k \frac{\rho(z_{i-1},z_i)^2}{2s_i}\right). \tag{C.32}
$$

We claim that for any $z \in \mathcal{R}^c$, we have

$$T := \prod_{i=1}^{k} \exp\left(-\sum_{i=1}^{k} \frac{\rho(z_{i-1}, z_i)^2}{2s_i(d+1)}\right) \mathbb{1}_{\rho(z_{i-1},z_i)\le\iota}\Big(1+\frac{1}{\iota s_i}\mathbb{1}_{\rho(z_{i-1},z_i)>\iota/2}\Big) \le \exp\left(-\frac{\iota^2}{16td}\right). \tag{C.33}$$

The claim is proved at the end of this proof. It is tempting to plug this back into (C.32), and argue that when t is polynomially small, the integrand in I_{out} becomes exponentially small. However, this would not work since it does not resolve the singular factors $\prod_{i=1}^{k} s_i^{-d/2}$ in the integrand. For this purpose, we need the following crucial "freezing" trick, which follows trivially from $1 = \frac{1}{1+d^{-1}} + \frac{1}{d+1}$:

$$\begin{aligned}
&\prod_{i=1}^{k} G(s_i, z_{i-1}, z_i)\mathbb{1}_{\rho(z_{i-1},z_i)\le\iota}\Big(1+\frac{1}{\iota s_i}\mathbb{1}_{\rho(z_{i-1},z_i)>\iota/2}\Big) \\
&= \left(\prod_{i=1}^{k}(1+d^{-1})^{d/2} G\big((1+d^{-1})s_i, z_{i-1}, z_i\big)\mathbb{1}_{\rho(z_{i-1},z_i)\le\iota}\right) T.
\end{aligned}$$

The idea is to keep the Gaussian behavior to resolve the $s_i^{-d/2}$ factors, and only single out a very small proportion to demonstrate exponential smallness. Plug (C.33) into the above identity to obtain

$$\begin{aligned}
I_{\mathsf{out}}(s) \le{}& \exp\left(-\frac{\iota^2}{16td}\right) \\
&\times \int_{\mathcal{R}^c} \prod_{i=1}^{k} \left[(1+d^{-1})^{d/2} G\big((1+d^{-1})s_i, z_{i-1}, z_i\big)\mathbf{1}_{\{\rho(z_{i-1},z_i)\le\iota\}}\right] \mu^{\otimes(k-1)}(\mathrm{d}z).
\end{aligned}$$

Integrate successively for each variable, convert to normal coordinates, and apply Lemma 23, Eqn. (C.25), and Gaussian integration as we did in bounding I_{loc}, we obtain

$$\begin{aligned}
&\int_{\mathcal{R}^c} \left(\prod_{i=1}^{k} G\big((1+d^{-1})s_i, z_{i-1}, z_i\big)\mathbb{1}_{\rho(z_{i-1},z_i)\le\iota}\right) \mu^{\otimes(k-1)}(\mathrm{d}z) \\
&\quad \le 8^k(1+d^{-1})^{dk/2}(ct)^{-d/2} \le 32^k(ct)^{-d/2}.
\end{aligned}$$

Therefore

$$I_{\mathsf{out}}(s) \le 32^k(ct)^{-d/2} \cdot \exp\left(-\frac{\iota^2}{16td}\right) \le 32^k G(t, x, y), \tag{C.34}$$

where in the last inequality we used the assumption $\rho(x,y) \le t^{5/12} = \iota/(4d)$ and $t \le 1/\operatorname{poly}(d,K)$, so that $\exp(-\iota^2/(32td)) \le \exp(-\rho(x,y)^2/(2t))$ and $\exp(-\iota^2/(32td)) = \exp(-\frac{1}{2}dt^{-1/6}) \le (2\pi t)^{-d/2}$.

Putting things together. We plug the bounds (C.31) and (C.34) into (C.29) to obtain

$$\begin{aligned}\left|\Psi_1 * r_1^{*k}\right|(t,x,y)| &\le t^k P^k \int_{t\Delta^{k-1}} (8^k + 32^k) G(t,x,y)\mathrm{d}s \\ &\le \frac{40}{(k-1)!}(40t)^{2k-1} P^k G(t,x,y).\end{aligned}$$

The desired conclusion of Lemma 42 follows from the above inequality by summing over k and taking $t \le 1/\operatorname{poly}(d,K)$.

Proof of Claim (C.33). For $z \in \mathcal{R}^c$, let

$$J = \{i : \rho(z_{i-1}, z_i) > \iota/2\}.$$

By definition of $\mathcal{R}^c$, either J is nonempty, or there is i_0 such that $\rho(x, z_{i_0}) > 2\iota$. For $i \in J$, we note that

$$\exp\left(-\frac{\rho(z_{i-1},z_i)^2}{2s_i d}\right)\left(1+\frac{1}{\iota s_i}\right) \le \exp\left(-\frac{\iota^2}{8s_i d}\right)\left(1+\frac{1}{\iota s_i}\right) \le \exp\left(-\frac{\iota^2}{16td}\right), \tag{C.35}$$

where the last inequality follows from $s_i \le t$, $\iota = 4t^{5/12}d$, and that $t \le 1/\operatorname{poly}(d,K)$. When J is nonempty, we readily deduce (C.33) as all the other factors are ≤ 1.

When J is empty, let i_0 be such that $\rho(x, z_{i_0}) > 2\iota$. We apply Cauchy-Schwarz to obtain

$$\sum_{i=1}^k \frac{\rho(z_{i-1},z_i)^2}{s_i} \ge \frac{1}{\sum_{i=1}^k s_i}\left(\sum_{i=1}^k \rho(z_{i-1},z_i)\right)^2 = \frac{1}{t}\left(\sum_{i=1}^k \rho(z_{i-1},z_i)\right)^2.$$

Then, by triangle inequalities, we have

$$\sum_{i=1}^{i_0} \rho(z_{i-1},z_i) \ge \rho(z_0,z_{i_0}) = \rho(x,z_{i_0}) > 2\iota, \quad \text{therefore} \quad \sum_{i=1}^k \frac{\rho(z_{i-1},z_i)^2}{s_i} \ge \frac{2\iota^2}{t}.$$

The desired claim (C.33) follows immediately, given that J is empty. $\square$

We now have all the ingredients to prove Lemma 38.

Proof of Lemma 38. This follows immediately from Lemma 39, Lemma 41 and Lemma 42. Note that in applying Lemma 39, we used $r \le t^{5/12}$, thus $r^3 \le t^{5/4} \le t$. $\square$

C.5.3 Proof of Lemma 39

By definition, we can compute

$$\frac{\Phi(t,x,y)}{\Psi(t,x,y)} = \exp\left((\log_x y)\cdot v - \psi(x,y)\right)\exp\left(-\|v\|^2 t/2\right).$$

Note that $\|v\| \le C\|\nabla \log p_{t_k}(Y_{t_k})\|$ by Lemma 23, which in turn is bounded by $\mathrm{poly}(d,K,\delta^{-1})$ by (C.23). When $t \le \frac{1}{\mathrm{poly}(d,K,\delta^{-1})} \le \frac{1}{4\|v\|^2}$, we have $|\exp(-\|v\|^2 t/2) - 1| \le \|v\|^2 t \le \mathrm{poly}(d,K,\delta^{-1})t$. Therefore, it suffices to show

$$|(\log_x y)\cdot v - \psi(x,y)| \le \mathrm{poly}(d,K,\delta^{-1})r^3.$$

Recall the definition of ψ. Note that since $r \le \frac{1}{\mathrm{poly}(d,K,\delta^{-1})}$, when the polynomial $\mathrm{poly}(d,K,\delta^{-1})$ is sufficiently large, the geodesic γ from x to y is unique and is inside $B_{Y_{t_k}}(r)$. In the normal coordinate on Y_{t_k} within radius r, the vector field $\mathscr{S}_{t_k,Y_{t_k}}$ is represented by the constant vector v. We also recognize that in normal coordinate, $\gamma(1)-\gamma(0) = \log_x y$. We thus have

$$\begin{aligned}
|(\log_x y)\cdot v - \psi(x,y)| &= \left|\delta_{\alpha\beta}v^\alpha(\gamma^\beta(1)-\gamma^\beta(0)) - \int_0^1 g_{\alpha\beta}(\gamma(s))v^\alpha \frac{\mathrm{d}}{\mathrm{d}s}\gamma^\beta(s)\mathrm{d}s\right| \\
&= \left|\int_0^1 (g_{\alpha\beta}(\gamma(s)) - \delta_{\alpha\beta})v^\alpha \frac{\mathrm{d}}{\mathrm{d}s}\gamma^\beta(s)\mathrm{d}s\right| \\
&\le C\int_0^1 \|g(\gamma(s)) - I\|\cdot\|v\|\cdot\|\frac{\mathrm{d}}{\mathrm{d}s}\gamma(s)\|\mathrm{d}s \\
&\le C\|v\|\cdot\rho(x,y)\int_0^1 CK(s\rho(x,y))^2\mathrm{d}s \\
&\le \mathrm{poly}(d,K,\delta^{-1})\rho(x,y)^3,
\end{aligned}$$

as desired. Here the penultimate line follows from Lemma 23.

C.5.4 Proof of Lemma 37: handling exceptional events

Proof of Lemma 37. Recall that

$$p_0^{\mathsf{aux}} = p_N \mathcal{K}_N^{\mathsf{aux}} \mathcal{K}_{N-1}^{\mathsf{aux}} \cdots \mathcal{K}_1^{\mathsf{aux}}$$

and

$$q_0^\star = p_N \widehat{\mathcal{K}}_N \widehat{\mathcal{K}}_{N-1} \cdots \widehat{\mathcal{K}}_1.$$

We need to compare the kernel $\mathcal{K}_k^{\mathsf{aux}}$ with $\widehat{\mathcal{K}}_k$, $k = 1, \cdots, N$. To apply Lemma 38, we define two auxiliary kernels $\widetilde{\mathcal{K}}_k^{\mathsf{aux}}$, $\widetilde{\mathcal{K}}_k$ that are "localized" version of $\mathcal{K}_k^{\mathsf{aux}}$ and $\widehat{\mathcal{K}}_k$. We show the auxiliary kernels are close to $\mathcal{K}_k^{\mathsf{aux}}$ and $\widehat{\mathcal{K}}_k$ respectively in total variation, and establish bound on $\mathsf{KL}(\widetilde{\mathcal{K}}_k^{\mathsf{aux}} \,\|\, \widetilde{\mathcal{K}}_k)$. Denote

$$\mathcal{R}_x \coloneqq \{y \in \mathcal{M} : \rho(x, y) \le h^{5/12}\}, \quad R_x^c \coloneqq \mathcal{M} \setminus \mathcal{R}_x.$$

Recall the notation Φ in the proof of Lemma 42. To distinguish the kernels at different step, we denote Φ_k as the corresponding Φ at step k. Define $\widetilde{\mathcal{K}}_k^{\mathsf{aux}}, \widetilde{\mathcal{K}}_k$ by

$$\widetilde{\mathcal{K}}_k^{\mathsf{aux}}(x, \mathrm{d}y) = \mathcal{K}_k^{\mathsf{aux}}(x, \mathrm{d}y) \mathbb{1}_{\mathcal{R}_x}(y) + \frac{\mathcal{K}_k^{\mathsf{aux}}(x, \mathcal{R}_x^c)}{\mu(\mathcal{R}_x^c)} \mu(\mathrm{d}y) \mathbb{1}_{\mathcal{R}_x^c},$$
$$\widetilde{\mathcal{K}}_k(x, \mathrm{d}y) = \Phi_k(h, x, y) \mathbb{1}_{\mathcal{R}_x}(y) + \frac{\int_{\mathcal{R}_x^c} \Phi_k(h, x, z) \mu(\mathrm{d}z)}{\mu(\mathcal{R}_x^c)} \mu(\mathrm{d}y) \mathbb{1}_{\mathcal{R}_x^c},$$

By converting to normal coordinate and invoking Gaussian integration in the same way as in the proof of Lemma 42, we obtain

$$\exp\Big(-\frac{2}{h^{1/6}}\Big) \le \int_{\mathcal{R}_x^c} \Phi_k(h, x, z) \mu(\mathrm{d}z) \le \exp\Big(-\frac{1}{16 h^{1/6}}\Big). \tag{C.36}$$

When $h \le 1/\operatorname{poly}(d, K, \delta^{-1})$, it is apparent (e.g., follows from Gromov's volume comparison theorem) that $\mu(\mathcal{R}^x) \le 1/2$, thus

$$\frac{1}{2} \le \mu(\mathcal{R}_x^c) \le \mu(\mathcal{M}) = 1.$$

We observe that $\widehat{\mathcal{K}}_k$ differs from $\widetilde{\mathcal{K}}_k$ by a rejection sampling with radius $h^{1/4}$. With the above bounds and the same Gaussian integration technique, we see that the

probability of rejection is bounded by

$$\mathbb{P}(\text{rejection at step } k) \le \exp\Big(-\frac{(h^{1/4})^2}{16h}\Big) \le \exp\Big(-\frac{1}{16h^{1/2}}\Big). \tag{C.37}$$

Summing up, We readily obtain

$$\mathsf{TV}(\widehat{\mathcal{K}}_k, \widetilde{\mathcal{K}}_k) \le \exp\left(-\frac{1}{16h^{1/6}}\right). \tag{C.38}$$

On the other hand, by using the stopping time argument as in the proof of (C.21), we have

$$\mathcal{K}_k^{\mathsf{aux}}(x, \mathcal{R}_x^c) \le \exp\Big(-\frac{1}{16h^{1/6}}\Big). \tag{C.39}$$

Therefore, the following TV bound is obvious:

$$\mathsf{TV}(\mathcal{K}_k^{\mathsf{aux}}, \widetilde{\mathcal{K}}_k^{\mathsf{aux}}) \le \exp\left(-\frac{1}{16h^{1/6}}\right), \qquad . \tag{C.40}$$

Now we compute $\mathsf{KL}\big(\widetilde{\mathcal{K}}_k^{\mathsf{aux}}(x,\cdot) \parallel \widetilde{\mathcal{K}}_k(x,\cdot)\big)$. By definition, we have

$$\begin{aligned}\mathsf{KL}\big(\widetilde{\mathcal{K}}_k^{\mathsf{aux}}(x,\cdot) \parallel \widetilde{\mathcal{K}}_k(x,\cdot)\big) = &\underbrace{\int_{\mathcal{R}_x} \left(\log \frac{\widetilde{\mathcal{K}}_k^{\mathsf{aux}}(x,\mathrm{d}y)}{\widetilde{\mathcal{K}}_k(x,\mathrm{d}y)}\right) \widetilde{\mathcal{K}}_k^{\mathsf{aux}}(x,\mathrm{d}y)}_{=:T_1} \\ &+ \underbrace{\left(\log \frac{\mathcal{K}_k^{\mathsf{aux}}(x,\mathcal{R}_x^c)}{\int_{\mathcal{R}_x^c} \Phi_k(h,x,z)\mu(\mathrm{d}z)}\right) \widetilde{\mathcal{K}}_k(x,\mathcal{R}_x^c)}_{=:T_2}.\end{aligned}$$

We control the two terms separately.

Controlling T_1. We invoke Lemma 38 to see

$$\left|\frac{\widetilde{\mathcal{K}}_k^{\mathsf{aux}}(x,\mathrm{d}y)}{\widetilde{\mathcal{K}}_k(x,\mathrm{d}y)} - 1\right| \le \mathrm{poly}(d, K, \delta^{-1})h.$$

Therefore, we use the elementary fact that $\log(1+x) \geq x - 2x^2$ for $x \in [-1/2, 1/2]$ to obtain

$$\begin{aligned}
\log \frac{\widetilde{\mathcal{K}}_k^{\mathsf{aux}}(x, \mathrm{d}y)}{\widetilde{\mathcal{K}}_k(x, \mathrm{d}y)} &= -\log \frac{\widetilde{\mathcal{K}}_k(x, \mathrm{d}y)}{\widetilde{\mathcal{K}}_k^{\mathsf{aux}}(x, \mathrm{d}y)} \\
&\leq 1 - \frac{\widetilde{\mathcal{K}}_k(x, \mathrm{d}y)}{\widetilde{\mathcal{K}}_k^{\mathsf{aux}}(x, \mathrm{d}y)} + 2\left(\frac{\widetilde{\mathcal{K}}_k(x, \mathrm{d}y)}{\widetilde{\mathcal{K}}_k^{\mathsf{aux}}(x, \mathrm{d}y)} - 1\right)^2 \\
&\leq 1 - \frac{\widetilde{\mathcal{K}}_k(x, \mathrm{d}y)}{\widetilde{\mathcal{K}}_k^{\mathsf{aux}}(x, \mathrm{d}y)} + \mathrm{poly}(d, K, \delta^{-1})h^2,
\end{aligned}$$

provided $h \leq 1/\mathrm{poly}(d, K, \delta^{-1})$. Integrate with respect to $\widetilde{\mathcal{K}}_k^{\mathsf{aux}}(x, \mathrm{d}y)$ over $y \in \mathcal{R}_x$ to obtain

$$\begin{aligned}
T_1 &\leq \widetilde{\mathcal{K}}_k^{\mathsf{aux}}(x, \mathcal{R}_x) - \widetilde{\mathcal{K}}_k(x, \mathcal{R}_x) + \mathrm{poly}(d, K, \delta^{-1})h^2 \\
&\leq 2\exp\big(-\frac{1}{16h^{1/6}}\big) + \mathrm{poly}(d, K, \delta^{-1})h^2 \\
&\leq \mathrm{poly}(d, K, \delta^{-1})h^2,
\end{aligned}$$

where the second line follows from (C.36) and (C.39), and the last line follows from $h \leq 1/\mathrm{poly}(d, K, \delta^{-1})$ so that the exponential term is sufficiently small.

Controlling T_2. This is strightforward given (C.39) and (C.36). We obtain in the same way as above that

$$T_2 \leq \exp\big(-\frac{1}{32h^{1/6}}\big) \leq \mathrm{poly}(d, K, \delta^{-1}) \leq h^2.$$

Summarizing the above, we have shown that

$$\mathsf{KL}\big(\widetilde{\mathcal{K}}_k^{\mathsf{aux}}(x, \cdot) \parallel \widetilde{\mathcal{K}}_k(x, \cdot)\big) \leq \mathrm{poly}(d, K, \delta^{-1})h^2.$$

Accumulate the error over all N steps using post-processing inequality and apply Pinsker's inequality, we obtain

$$\mathsf{TV}(p_0^{\mathsf{aux}} \parallel q_0^\star) \leq \sqrt{\mathrm{poly}(d, K, \delta^{-1})h^2 N} \leq \sqrt{hT}\,\mathrm{poly}(d, K, \delta^{-1}),$$

since $hN = T - \delta \leq T$, as claimed. □

C.6 Proof of main results

Proof of Lemma 9. This follows from combining Lemma 33 and Lemma 34. □

Proof of Lemma 10. This follows from Lemma 36 and our choice of schedule $hN = T - \delta \leq T$. □

Proof of Theorem 4. This follows from Lemma 9, Lemma 10, and Lemma 37. □

Appendix D

Proofs for Chapter 6

D.1 Proof of the heat flow characterization

Proof of Theorem 5. We prove a general reverse-time representation for a single proximal step, and then apply it twice.

Step 1: a generic proximal step as reverse heat flow. Let r_0 be a probability density on $\mathbb{R}^d$, and let

$$r_t := r_0 * \varphi_t, \qquad \varphi_t(z) := (2\pi t)^{-d/2} \exp\left(-\frac{\|z\|^2}{2t}\right),$$

be its heat evolution. Fix $\tau > 0$, and consider the forward diffusion

$$\mathrm{d}Z_t = \mathrm{d}W_t, \qquad Z_0 \sim r_0, \qquad 0 \le t \le \tau.$$

Then Z_t has density r_t.

Now define the time-reversed process

$$\widehat{Z}_s := Z_{\tau - s}, \qquad 0 \le s \le \tau.$$

By the standard time-reversal formula for diffusions with unit diffusion matrix, the reversed process is again a diffusion, with drift given by the score of the time-$(\tau - s)$ marginal:

$$\mathrm{d}\widehat{Z}_s = \nabla \log r_{\tau - s}(\widehat{Z}_s)\,\mathrm{d}s + \mathrm{d}\widehat{W}_s, \qquad 0 \le s \le \tau,$$

for some Brownian motion $(\widehat{W}_s)_{0\le s\le\tau}$.

Condition now on the terminal value of the forward process, equivalently on the initial value of the reversed process:

$$\widehat{Z}_0 = Z_\tau = x.$$

Under this conditioning, $\widehat{Z}$ is precisely the reverse-time diffusion started from x,

$$\mathrm{d}X_s = \nabla \log r_{\tau-s}(X_s)\,\mathrm{d}s + \mathrm{d}B_s, \qquad X_0 = x, \qquad 0 \le s \le \tau.$$

Its terminal law is

$$\mathsf{Law}(X_\tau) = \mathsf{Law}(Z_0 \mid Z_\tau = x).$$

Since $Z_\tau = Z_0 + G$ with $G \sim \mathcal{N}(0, \tau I_d)$ independent of Z_0, Bayes' rule gives

$$\mathsf{Law}(Z_0 \in \mathrm{d}u \mid Z_\tau = x) \propto r_0(u)\,\varphi_\tau(x-u)\,\mathrm{d}u \propto r_0(u)\exp\left(-\frac{\|u-x\|^2}{2\tau}\right)\mathrm{d}u.$$

Therefore the endpoint X_τ of the reverse diffusion

$$\mathrm{d}X_s = \nabla \log r_{\tau-s}(X_s)\,\mathrm{d}s + \mathrm{d}B_s, \qquad X_0 = x,$$

has exactly the proximal law with base density r_0 and quadratic parameter τ.

Step II: application to the consistency substep. Take $r_0 = p^\star$ and $\tau = \eta^2$. Then the diffusion denoising sampler from input x produces exactly the terminal value at time η^2 of

$$\mathrm{d}X_t = \nabla \log q_{\eta^2-t}(X_t)\,\mathrm{d}t + \mathrm{d}B_t, \qquad X_0 = x, \qquad 0 \le t \le \eta^2.$$

Step III: application to the denoising substep. Let

$$q_0(u) \propto \exp(\mathcal{L}(u; y)),$$

and let q_t be its heat evolution. Applying the same argument with $r_0 = q_0$ and $\tau = \eta^2$, we find that, conditional on the intermediate state X_{η^2}, one proximal consistency

step is exactly the terminal value at time $2\eta^2$ of

$$\mathrm{d}X_t = \nabla \log p_{2\eta^2 - t}(X_t)\,\mathrm{d}t + \mathrm{d}B_t, \qquad \eta^2 < t \le 2\eta^2.$$

Step IV: concatenation. Composing the two Markov kernels yields the piecewise diffusion

$$\mathrm{d}X_t = \begin{cases} \nabla \log q_{\eta^2 - t}(X_t)\,\mathrm{d}t + \mathrm{d}B_t, & 0 \le t \le \eta^2, \\ \nabla \log p_{2\eta^2 - t}(X_t)\,\mathrm{d}t + \mathrm{d}B_t, & \eta^2 < t \le 2\eta^2, \end{cases} \qquad X_0 = x.$$

Its endpoint $X_{2\eta^2}$ is exactly the output of one full iteration of DPnP, in the order encoded by the displayed SDE. □

D.2 Proof of the theoretical guarantee

The following well-known lemma is crucial to our analysis. Let $H_t(x, y)$ be the heat kernel on $\mathcal{M}$, i.e., the distribution of the manifold Brownian motion starting from x.

Lemma 43 (Symmetry of heat kernel, cf. Berline et al. [8])**.** *We have*

$$H_t(x, y) = H_t(y, x), \quad \forall x, y \in \mathcal{M}.$$

We now prove the theorems stated in Chapter 6. The argument is identical to the proof of the corresponding Euclidean results in Appendix A, so we omit the repeated steps. We record only the manifold-specific ingredients: the lemmas below identify the transition kernels of the ideal manifold updates and characterize their stationary distribution.

Lemma 44 (Transition kernels of the manifold SDE)**.** *Let P_t be the heat semigroup on $\mathcal{M}$, with heat kernel $H_t(x, z)$ with respect to μ. Let $\varphi : \mathcal{M} \to [0, \infty)$ be continuous and not identically zero, and define*

$$h_t(x) \coloneqq P_t\varphi(x) = \int_{\mathcal{M}} H_t(x, z)\varphi(z)\,\mu(\mathrm{d}z).$$

Fix $T > 0$. The time-inhomogeneous diffusion

$$\mathrm{d}X_t = \nabla \log h_{T-t}(X_t)\,\mathrm{d}t + \mathrm{d}B_t^{\mathcal{M}}, \qquad 0 \le t < T, \qquad X_0 = x,$$

has transition kernel

$$K_{s,t}^{\varphi}(u, \mathrm{d}v) = \frac{h_{T-t}(v)}{h_{T-s}(u)} H_{t-s}(u, v)\,\mu(\mathrm{d}v), \qquad 0 \le s < t \le T.$$

In particular,

$$K_{0,T}^{\varphi}(x, \mathrm{d}z) = \frac{\varphi(z) H_T(x, z)}{h_T(x)}\,\mu(\mathrm{d}z).$$

Proof. This is the standard Doob h-transform of Brownian motion by the positive space-time harmonic function

$$(t, x) \mapsto h_{T-t}(x) = P_{T-t}\varphi(x).$$

Indeed, the heat semigroup property gives

$$\int_{\mathcal{M}} h_{T-t}(v) H_{t-s}(u, v)\,\mu(\mathrm{d}v) = P_{t-s} h_{T-t}(u) = h_{T-s}(u),$$

so $K_{s,t}^{\varphi}$ is a Markov kernel. The Doob transform therefore has transition kernel

$$K_{s,t}^{\varphi}(u, \mathrm{d}v) = \frac{h_{T-t}(v)}{h_{T-s}(u)} H_{t-s}(u, v)\,\mu(\mathrm{d}v),$$

and its infinitesimal generator is

$$\frac{1}{2}\Delta_{\mathcal{M}} + \langle \nabla \log h_{T-t}, \nabla \cdot \rangle,$$

which is precisely the generator of the displayed SDE. Taking $s = 0$ and $t = T$, and using $h_0 = \varphi$, gives the terminal kernel. □

Lemma 45 (Stationary distribution of manifold DPnP)**.** *For each $\eta > 0$, define*

$$\gamma_\eta(\mathrm{d}x, \mathrm{d}z) \propto p^\star(x)\, \mathrm{e}^{\mathcal{L}(z;y)}\, H_{\eta^2}(x, z)\,\mu(\mathrm{d}x)\mu(\mathrm{d}z).$$

Then the two conditional laws of γ_η are exactly the ideal data-consistency step and

the ideal diffusion denoising step of manifold DPnP*. In particular, the exact manifold update preserves the x-marginal*

$$\pi_\eta(\mathrm{d}x) \propto p^\star(x) \left(\int_{\mathcal{M}} \mathrm{e}^{\mathcal{L}(z;y)} H_{\eta^2}(x,z)\, \mu(\mathrm{d}z) \right) \mu(\mathrm{d}x).$$

In other words, π_η defined above is the stationary distribution of manifold DPnP *with constant annealing parameter η.*

Proof. Applying Lemma 44 with $\varphi(z) = \mathrm{e}^{\mathcal{L}(z;y)}$ and $T = \eta^2$, the SDE

$$\mathrm{d}X_t = \nabla \log q_{\eta^2 - t}(X_t)\, \mathrm{d}t + \mathrm{d}B_t, \qquad 0 \le t < \eta^2, \qquad X_0 = x,$$

has terminal transition kernel

$$K_{\mathrm{dc},\eta}(x, \mathrm{d}z) = \frac{\mathrm{e}^{\mathcal{L}(z;y)} H_{\eta^2}(x,z)}{q_{\eta^2}(x)}\, \mu(\mathrm{d}z).$$

Similarly, applying Lemma 44 with $\varphi(x) = p^\star(x)$ and $T = \eta^2$, the SDE

$$\mathrm{d}X_t = \nabla \log p_{2\eta^2 - t}(X_t)\, \mathrm{d}t + \mathrm{d}B_t, \qquad \eta^2 \le t < 2\eta^2, \qquad X_{\eta^2} = z,$$

has terminal transition kernel

$$K_{\mathrm{diff},\eta}(z, \mathrm{d}x) = \frac{p^\star(x) H_{\eta^2}(z,x)}{p_{\eta^2}(z)}\, \mu(\mathrm{d}x),$$

By the symmetry of the heat kernel, this may equivalently be written as

$$K_{\mathrm{diff},\eta}(z, \mathrm{d}x) = \frac{p^\star(x) H_{\eta^2}(x,z)}{p_{\eta^2}(z)}\, \mu(\mathrm{d}x).$$

Also, by the definition of heat kernel [8], we have

$$q_t(x) = \int_{\mathcal{M}} \mathrm{e}^{\mathcal{L}(z;y)} H_t(x,z)\, \mu(\mathrm{d}z)$$

and

$$p_t(z) = \int_{\mathcal{M}} p^\star(x) H_t(z,x)\, \mu(\mathrm{d}x).$$

Under γ_η, the conditional law of z given x is

$$\gamma_\eta(\mathrm{d}z \mid x) \propto \mathrm{e}^{\mathcal{L}(z;y)}\, H_{\eta^2}(x,z)\, \mu(\mathrm{d}z),$$

which is exactly the manifold data-consistency step.

On the other hand, the conditional law of x given z is

$$\gamma_\eta(\mathrm{d}x \mid z) \propto p^\star(x)\, H_{\eta^2}(x,z)\, \mu(\mathrm{d}x).$$

At this point, we invoke Lemma 43 to see

$$H_{\eta^2}(x,z) = H_{\eta^2}(z,x).$$

Hence

$$\gamma_\eta(\mathrm{d}x \mid z) \propto p^\star(x)\, H_{\eta^2}(z,x)\, \mu(\mathrm{d}x),$$

which is exactly the posterior denoising law targeted by the Riemannian score-based sampler.

Therefore, one exact step of manifold DPnP is exactly Gibbs sampling for γ_η, and so it preserves the x-marginal of γ_η. This marginal is

$$\gamma_\eta^X(\mathrm{d}x) \propto p^\star(x) \left(\int_{\mathcal{M}} \mathrm{e}^{\mathcal{L}(z;y)} H_{\eta^2}(x,z)\, \mu(\mathrm{d}z) \right) \mu(\mathrm{d}x) = \pi_\eta(\mathrm{d}x),$$

where

$$\pi_\eta(\mathrm{d}x) \propto p^\star(x)\, q_\eta(x)\, \mu(\mathrm{d}x), \qquad q_\eta(x) = \int_{\mathcal{M}} \mathrm{e}^{\mathcal{L}(z;y)} H_{\eta^2}(x,z)\, \mu(\mathrm{d}z).$$

It remains to pass to the limit $\eta \to 0$. Since $\mathcal{M}$ is compact and $\mathcal{L}(\cdot\,; y)$ is continuous, the heat semigroup converges uniformly to the identity on $C(\mathcal{M})$. Thus

$$q_\eta(x) = \int_{\mathcal{M}} \mathrm{e}^{\mathcal{L}(z;y)} H_{\eta^2}(x,z)\, \mu(\mathrm{d}z) \to \mathrm{e}^{\mathcal{L}(x;y)}$$

uniformly in $x \in \mathcal{M}$. Consequently,

$$\pi_\eta(\mathrm{d}x) \Rightarrow p^\star(\mathrm{d}x \mid y) \propto p^\star(x) \mathrm{e}^{\mathcal{L}(x;y)}\, \mu(\mathrm{d}x).$$

Combining this with the slowly diminishing annealing schedule yields the claimed convergence in distribution of $\hat{x}_{k_l}$ to the posterior $p^\star(\cdot \mid y)$. $\square$

Bibliography

[1] Asad Aali, Marius Arvinte, Sidharth Kumar, and Jonathan I Tamir. Solving inverse problems with score-based generative priors learned from noisy data. In *2023 57th Asilomar Conference on Signals, Systems, and Computers*, pages 837–843. IEEE, 2023. Cited on page 37.

[2] P.-A. Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008. Cited on page 68.

[3] Kwangjun Ahn and Sinho Chewi. Efficient constrained sampling via the mirror-Langevin algorithm. *Advances in Neural Information Processing Systems*, 34: 28405–28418, 2021. Cited on page 60.

[4] Jason M Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 2169–2176. IEEE, 2023. Cited on page 26.

[5] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005. Cited on pages 11, 12, and 42.

[6] Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. Cited on pages 1 and 11.

[7] Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024. Cited on pages 5, 39, 54, 56, 59, 60, 155, and 158.

[8] Nicole Berline, Ezra Getzler, and Michele Vergne. *Heat kernels and Dirac operators*. Springer Science & Business Media, 2003. Cited on pages 59, 142, 160, 164, 165, 179, and 181.

[9] Eliot Beyler and Francis Bach. Convergence of deterministic and stochastic diffusion-model samplers: A simple analysis in Wasserstein distance. *arXiv preprint arXiv:2508.03210*, 2025. Cited on page 60.

[10] Karthik Bharath, Alexander Lewis, Akash Sharma, and Michael V Tretyakov. Sampling and estimation on manifolds using the Langevin diffusion. In *Advances in Neural Information Processing Systems*, 2025. Cited on page 60.

[11] Charles A Bouman and Gregery T Buzzard. Generative plug and play: Posterior sampling for inverse problems. *arXiv preprint arXiv:2306.07233*, 2023. Cited on pages 21, 29, 33, and 90.

[12] Nicolas Bourbaki. *Théories spectrales: Chapitres 1 et 2*. Springer, 2023. Cited on page 97.

[13] Nicolas Bourbaki. *Théories spectrales: Chapitres 3 à 5*. Springer Nature, 2023. Cited on page 96.

[14] Joan Bruna and Jiequn Han. Posterior sampling with denoising oracles via tilted transport. *arXiv preprint arXiv:2407.00745*, 2024. Cited on page 47.

[15] Gregery T Buzzard, Stanley H Chan, Suhas Sreehari, and Charles A Bouman. Plug-and-play unplugged: Optimization-free reconstruction using consensus equilibrium. *SIAM Journal on Imaging Sciences*, 11(3):2001–2020, 2018. Cited on page 32.

[16] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015. Cited on page 27.

[17] Gabriel Cardoso, Yazid Janati El Idrissi, Sylvain Le Corff, and Eric Moulines. Monte Carlo guided diffusion for Bayesian linear inverse problems. *arXiv preprint arXiv:2308.07983*, 2023. Cited on pages 3, 33, and 47.

[18] Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. *arXiv preprint arXiv:2302.07194*, 2023. Cited on pages 51,

56, 57, 58, and 60.

[19] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022. Cited on page 10.

[20] Xiang Cheng, Jingzhao Zhang, and Suvrit Sra. Efficient sampling on Riemannian manifolds via Langevin MCMC. *Advances in Neural Information Processing Systems*, 35:5995–6006, 2022. Cited on pages 51 and 60.

[21] Xiang Cheng, Jingzhao Zhang, and Suvrit Sra. Theory and algorithms for diffusion processes on Riemannian manifolds. *arXiv preprint arXiv:2204.13665*, 2022. Cited on pages 5, 49, and 60.

[22] Sinho Chewi, Chen Lu, Kwangjun Ahn, Xiang Cheng, Thibaut Le Gouic, and Philippe Rigollet. Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm. In *Conference on Learning Theory*, pages 1260–1300. PMLR, 2021. Cited on page 25.

[23] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. Cited on pages xi, 3, 26, 29, 30, 31, 32, and 33.

[24] Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. Decomposed diffusion sampler for accelerating large-scale inverse problems. In *The Twelfth International Conference on Learning Representations*, 2024. Cited on page 33.

[25] Florentin Coeurdoux, Nicolas Dobigeon, and Pierre Chainais. Plug-and-play split Gibbs sampler: embedding deep generative priors in Bayesian inference. *IEEE Transactions on Image Processing*, 33:3496–3507, 2024. Cited on page 33.

[26] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021. Cited on page 60.

[27] Valentin De Bortoli, Emile Mathieu, Michael John Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative

modelling. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. Cited on pages 5, 49, 50, 51, 52, 53, 54, 57, 60, and 75.

[28] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. Cited on page 1.

[29] J. L. Doob. The Brownian movement and stochastic equations. *Annals of Mathematics*, 43(2):351–369, 1942. ISSN 0003486X. Cited on page 10.

[30] Zehao Dou and Yang Song. Diffusion posterior sampling for linear inverse problem solving: A filtering perspective. In *International Conference on Learning Representations*, 2024. Cited on pages 3, 33, and 47.

[31] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. Cited on page 32.

[32] Lawrence C Evans. *An introduction to stochastic differential equations*, volume 82. American Mathematical Soc., 2012. Cited on page 10.

[33] Zhenghan Fang, Sam Buchanan, and Jeremias Sulam. What's in a prior? Learned proximal networks for inverse problems. In *The Twelfth International Conference on Learning Representations*, 2024. Cited on page 32.

[34] Berthy T. Feng, Jamie Smith, Michael Rubinstein, Huiwen Chang, Katherine L. Bouman, and William T. Freeman. Score-based diffusion models as principled priors for inverse imaging. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10486–10497, 2023. Cited on page 3.

[35] Nic Fishman, Leo Klarner, Valentin De Bortoli, Emile Mathieu, and Michael Hutchinson. Diffusion models for constrained domains. *arXiv preprint arXiv:2304.05364*, 2023. Cited on page 5.

[36] Khashayar Gatmiry and Santosh S. Vempala. Convergence of the Riemannian Langevin algorithm. *ArXiv*, abs/2204.10818, 2022. Cited on page 5.

[37] Edward I George and Robert E McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993. Cited on page 47.

[38] Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:123 – 214, 03 2011. doi: 10.1111/j.1467-9868.2010.00765.x. Cited on page 5.

[39] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 35:14715–14728, 2022. Cited on page 33.

[40] Alfred Gray. *Tubes*, volume 221. Springer Science & Business Media, 2003. Cited on page 150.

[41] Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th international conference on international conference on machine learning*, pages 399–406, 2010. Cited on page 32.

[42] Yunrui Guan, Krishnakumar Balasubramanian, and Shiqian Ma. Riemannian proximal sampler for high-accuracy sampling on manifolds. *arXiv preprint arXiv:2502.07265*, 2025. Cited on pages 5 and 60.

[43] Wei Guo, Molei Tao, and Yongxin Chen. Provable benefit of annealed Langevin Monte Carlo for non-log-concave sampling. *arXiv preprint arXiv:2407.16936*, 2024. Cited on page 38.

[44] Shivam Gupta, Ajil Jalal, Aditya Parulekar, Eric Price, and Zhiyang Xun. Diffusion posterior sampling is computationally intractable. *arXiv preprint arXiv:2402.12727*, 2024. Cited on pages 26, 33, and 35.

[45] Richard S Hamilton. Matrix Harnack estimate for the heat equation. *Communications in analysis and geometry*, 1(1):113–126, 1993. Cited on page 151.

[46] Jihun Hamm and Daniel D. Lee. Grassmann discriminant analysis: A unifying view on subspace-based learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 376–383, 2008. Cited on page 68.

[47] Qing Han and Qi S Zhang. An upper bound for Hessian matrices of positive solutions of heat equations. *The Journal of Geometric Analysis*, 26(2):715–749, 2016. Cited on pages 151 and 152.

[48] Ulrich G. Haussmann and Étienne Pardoux. Time reversal of diffusions. *Annals*

of Probability, 14:1188–1205, 1986. Cited on page 1.

[49] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. Cited on pages 1, 11, 12, 33, and 60.

[50] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022. Cited on page 1.

[51] E.P. Hsu. *Stochastic Analysis on Manifolds*. Graduate studies in mathematics. American Mathematical Society, 2002. ISBN 9780821808023. Cited on page 57.

[52] Chin-Wei Huang, Milad Aghajohari, Joey Bose, Prakash Panangaden, and Aaron C Courville. Riemannian diffusion models. *Advances in Neural Information Processing Systems*, 35:2750–2761, 2022. Cited on page 5.

[53] Zhihan Huang, Yuting Wei, and Yuxin Chen. Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality. *arXiv preprint arXiv:2410.18784*, 2024. Cited on page 60.

[54] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005. Cited on pages 1, 5, and 11.

[55] Hemant Ishwaran and J Sunil Rao. Spike and slab variable selection: Frequentist and Bayesian strategies. *Annals of Statistics*, 33(2):730–773, 2005. Cited on page 47.

[56] Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. Diff-TTS: A denoising diffusion model for text-to-speech. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 3566–3570, 2021. doi: 10.21437/Interspeech.2021-469. Cited on page 1.

[57] Qijia Jiang. From estimation to sampling for Bayesian linear regression with spike-and-slab prior. *arXiv preprint arXiv:2307.05558*, 2023. Cited on page 47.

[58] Jürgen Jost. *Riemannian Geometry and Geometric Analysis*. Universitext.

Springer, Cham, 7 edition, 2017. ISBN 978-3-319-61859-3. doi: 10.1007/978-3-319-61860-9. Cited on page 7.

[59] Zahra Kadkhodaie and Eero Simoncelli. Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. *Advances in Neural Information Processing Systems*, 34:13242–13254, 2021. Cited on page 47.

[60] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. Cited on page 29.

[61] Bahjat Kawar, Gregory Vaksman, and Michael Elad. Stochastic image denoising by sampling from the posterior distribution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1866–1875, 2021. Cited on page 32.

[62] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022. Cited on pages 3 and 33.

[63] Robert Keys. Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing*, 29(6):1153–1160, 1981. Cited on page 28.

[64] Syamantak Kumar, Purnamrita Sarkar, Kevin Tian, and Yusong Zhu. Spike-and-slab posterior sampling in high dimensions. In Nika Haghtalab and Ankur Moitra, editors, *Proceedings of the Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 3407–3462. PMLR, 2025. Cited on page 47.

[65] Rémi Laumont, Valentin De Bortoli, Andrés Almansa, Julie Delon, Alain Durmus, and Marcelo Pereyra. Bayesian imaging using plug & play priors: when Langevin meets Tweedie. *SIAM Journal on Imaging Sciences*, 15(2): 701–737, 2022. Cited on page 33.

[66] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985, 2023. Cited on page 60.

[67] Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling

with a restricted Gaussian oracle. In *Conference on Learning Theory*, pages 2993–3050. PMLR, 2021. Cited on page 21.

[68] Oscar Leong and Yann Traonmilin. A recovery theory for diffusion priors: Deterministic analysis of the implicit prior algorithm. In *Proceedings of the Twenty-Ninth International Conference on Artificial Intelligence and Statistics*, 2026. Cited on pages 41 and 47.

[69] Gen Li and Yuchen Jiao. Improved convergence rate for diffusion probabilistic models. In *The Thirteenth International Conference on Learning Representations*, 2024. Cited on pages 54 and 60.

[70] Gen Li and Yuling Yan. Adapting to unknown low-dimensional structures in score-based diffusion models. *Advances in Neural Information Processing Systems*, 37:126297–126331, 2024. Cited on page 60.

[71] Gen Li and Yuling Yan. O(d/t) convergence theory for diffusion probabilistic models under minimal assumptions. In *The Thirteenth International Conference on Learning Representations*, 2025. Cited on pages 5, 54, and 60.

[72] Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Towards non-asymptotic convergence for diffusion-based generative models. In *International Conference on Learning Representations (ICLR)*, 2024. Cited on pages 5, 25, 39, 54, and 60.

[73] Junfang Li and Xiangjin Xu. Differential harnack inequalities on Riemannian manifolds i: Linear heat equation. *Advances in Mathematics*, 226(5):4456–4491, 2011. ISSN 0001-8708. doi: https://doi.org/10.1016/j.aim.2010.12.009. Cited on page 150.

[74] Mufan Bill Li and Murat A. Erdogdu. Riemannian Langevin algorithm for solving semidefinite programs. *Bernoulli*, 29:3093 – 3113, 2023. Cited on page 5.

[75] Peter Li and Shing Yau. On the parabolic kernel of the Schrödinger operator. *Acta Mathematica*, 156:153–201, 07 1986. doi: 10.1007/BF02399203. Cited on page 152.

[76] Jiadong Liang, Zhihan Huang, and Yuxin Chen. Low-dimensional adaptation of diffusion models: Convergence in total variation (extended abstract). In Nika Haghtalab and Ankur Moitra, editors, *Proceedings of the Thirty Eighth*

Conference on Learning Theory, volume 291 of *Proceedings of Machine Learning Research*, pages 3723–3729. PMLR, 2025. Cited on page 60.

[77] Yuchen Liang, Peizhong Ju, Yingbin Liang, and Ness Shroff. Broadening target distributions for accelerated diffusion models via a novel analysis approach. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025. Cited on page 60.

[78] Guan-Horng Liu, Tianrong Chen, Evangelos Theodorou, and Molei Tao. Mirror diffusion models for constrained and watermarked generation. *Advances in Neural Information Processing Systems*, 36:42898–42917, 2023. Cited on page 5.

[79] Aaron Lou, Minkai Xu, Adam Farris, and Stefano Ermon. Scaling Riemannian diffusion models. *Advances in Neural Information Processing Systems*, 36: 80291–80305, 2023. Cited on page 5.

[80] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35: 5775–5787, 2022. Cited on pages 12 and 34.

[81] Yi-An Ma, Niladri S Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L Bartlett, and Michael I Jordan. Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3):1942–1992, 2021. doi: 10.3150/20-BEJ1297. Cited on page 34.

[82] Oren Mangoubi and Nisheeth K Vishnoi. Nonconvex sampling with the Metropolis-adjusted Langevin algorithm. In *Conference on learning theory*, pages 2259–2293. PMLR, 2019. Cited on page 25.

[83] Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. Cited on pages 3 and 33.

[84] Vishal Monga, Yuelong Li, and Yonina C Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021. Cited on page 32.

[85] Andrea Montanari and Yuchen Wu. Provably efficient posterior sampling for

sparse linear regression via measure decomposition. *Journal of the American Statistical Association*, pages 1–19, 2026. Cited on page 47.

[86] Michelle Muniz, Matthias Ehrhardt, Michael Günther, and Renate Winkler. Higher strong order methods for linear Itô SDEs on matrix lie groups. *BIT Numerical Mathematics*, 62, 01 2022. doi: 10.1007/s10543-021-00905-9. Cited on page 2.

[87] National Geophysical Data Center / World Data Service. NCEI/WDS Global Significant Earthquake Database. doi: 10.7289/V5TD9V7K. Accessed: 2026-05-02. Cited on page 74.

[88] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171, 2021. Cited on page 1.

[89] Bernt Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer, Berlin, Heidelberg, 6 edition, 2003. doi: 10.1007/978-3-642-14394-6. Cited on page 1.

[90] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014. Cited on page 21.

[91] Peter Petersen. *Riemannian Geometry*, volume 171 of *Graduate Texts in Mathematics*. Springer, New York, 2 edition, 2006. ISBN 978-0-387-29246-5. doi: 10.1007/978-0-387-29403-2. Cited on page 7.

[92] M. Piggott and V. Solo. Geometric Euler–Maruyama schemes for stochastic differential equations in SO(n) and SE(n). *SIAM Journal on Numerical Analysis*, 54:2490–2516, 08 2016. doi: 10.1137/15M1019726. Cited on page 2.

[93] Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge university press, 2024. Cited on pages 92, 95, and 141.

[94] Peter Potaptchik, Iskander Azangulov, and George Deligiannidis. Linear convergence of diffusion models under the manifold hypothesis. In Nika Haghtalab and Ankur Moitra, editors, *Proceedings of the Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 4668–4685. PMLR, 2025. Cited on page 60.

[95] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen.

Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022. Cited on page 1.

[96] Edward T Reehorst and Philip Schniter. Regularization by denoising: Clarifications and new interpretations. *IEEE transactions on computational imaging*, 5(1):52–67, 2018. Cited on page 32.

[97] Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293. Springer Science & Business Media, 2013. Cited on page 160.

[98] Gareth O Roberts and Jeffrey S Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998. Cited on pages 22, 23, and 88.

[99] Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (RED). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017. Cited on page 32.

[100] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. Cited on page 1.

[101] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. Cited on page 29.

[102] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. Cited on page 1.

[103] Laurent Saloff-Coste. *Lectures on finite Markov chains*, pages 301–413. Springer Berlin Heidelberg, Berlin, Heidelberg, 1997. ISBN 978-3-540-69210-2. doi: 10.1007/BFb0092621. Cited on pages 96 and 97.

[104] HH Schaefer. *Banach Lattices and Positive Operators*, volume 215. Springer

Science & Business Media, 2012. Cited on page 96.

[105] Richard M Schoen and Shing-Tung Yau. *Lectures on differential geometry*, volume 1. International press Cambridge, MA, 1994. Cited on page 149.

[106] Amit Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and computational harmonic analysis*, 30(1):20–36, 2011. Cited on page 68.

[107] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265, 2015. Cited on page 1.

[108] Bowen Song, Soo Min Kwon, Zecheng Zhang, Xinyu Hu, Qing Qu, and Liyue Shen. Solving inverse problems with latent diffusion models via hard data consistency. In *The Twelfth International Conference on Learning Representations*, 2024. Cited on page 33.

[109] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. Cited on pages 11 and 12.

[110] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2022. Cited on page 33.

[111] Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning*, pages 32483–32498. PMLR, 2023. Cited on pages xi, 14, 29, 30, 31, 32, 33, and 34.

[112] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Neural Information Processing Systems*, 2019. Cited on page 1.

[113] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2021. Cited on page 3.

[114] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021. Cited on pages 1, 10, 11, and 12.

[115] Vishwak Srinivasan, Andre Wibisono, and Ashia Wilson. Fast sampling from constrained spaces using the Metropolis-adjusted mirror Langevin algorithm. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4593–4635. PMLR, 2024. Cited on page 60.

[116] Yu Sun, Zihui Wu, Yifan Chen, Berthy T Feng, and Katherine L Bouman. Provable probabilistic imaging using score-based generative priors. *IEEE Transactions on Computational Imaging*, 10:1290–1305, 2024. Cited on page 33.

[117] Luke Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701 – 1728, 1994. doi: 10.1214/aos/1176325750. Cited on pages 23 and 93.

[118] Brian L. Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023. Cited on page 33.

[119] Hajime Urakawa. Convergence rates to equilibrium of the heat kernels on compact Riemannian manifolds. *Indiana University mathematics journal*, pages 259–288, 2006. Cited on pages 56, 58, and 153.

[120] Singanallur V Venkatakrishnan, Charles A Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *IEEE Global Conference on Signal and Information Processing*, pages 945–948. IEEE, 2013. Cited on pages 21 and 32.

[121] Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009. Cited on page 102.

[122] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. Cited on pages 10 and 32.

[123] Maxime Vono, Nicolas Dobigeon, and Pierre Chainais. Split-and-augmented

Gibbs sampler-application to large-scale inference problems. *IEEE Transactions on Signal Processing*, 67(6):1648–1661, 2019. Cited on pages 21 and 33.

[124] Maxime Vono, Daniel Paulin, and Arnaud Doucet. Efficient MCMC sampling with dimension-free convergence rate using ADMM-type splitting. *Journal of Machine Learning Research*, 23(25):1–69, 2022. Cited on page 26.

[125] Andre Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR, 2018. Cited on pages 20 and 33.

[126] Luhuan Wu, Brian Trippe, Christian Naesseth, David Blei, and John P Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. *Advances in Neural Information Processing Systems*, 36, 2023. Cited on pages 3, 33, and 47.

[127] Xingyu Xu and Yuejie Chi. Provably robust score-based diffusion posterior sampling for plug-and-play image reconstruction. *Advances in Neural Information Processing Systems*, 37:36148–36184, 2024. Cited on pages 6, 13, and 47.

[128] Xingyu Xu, Ziyi Zhang, Yorie Nakahira, Guannan Qu, and Yuejie Chi. Polynomial convergence of riemannian diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026. Cited on pages 7 and 49.

[129] Zhiyang Xun, Shivam Gupta, and Eric Price. Posterior sampling by combining diffusion models with annealed Langevin dynamics. In *Advances in Neural Information Processing Systems*, 2025. Cited on pages 41 and 47.

[130] Yun Yang, Martin J Wainwright, and Michael I Jordan. On the computational complexity of high-dimensional Bayesian variable selection. *Annals of Statistics*, 44(6):2497–2532, 2016. Cited on page 47.

[131] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *The Eleventh International Conference on Learning Representations*, 2022. Cited on pages 12, 19, and 84.