
Score-Based Diffusion Modeling for Nonparametric Empirical Bayes in Heteroscedastic Gaussian Mixtures

Gongyu Chen

Department of IEOR
UC Berkeley
Berkeley, CA 94720
gongyu@berkeley.edu

Ying Cui

Department of IEOR
UC Berkeley
Berkeley, CA 94720
yingcui@berkeley.edu

Abstract

We propose a generalized score-based diffusion framework for learning multivariate Gaussian mixture models with homoscedastic or heteroscedastic noise. Our goal is to nonparametrically estimate the latent location distribution and denoise the observations. Departing from the conventional maximum likelihood approach, we reinterpret each observation as a temporal slice of a family of stochastic diffusion processes. This modeling choice enables a principled characterization of the additive noise structure and supports a multi-step denoising procedure grounded in reverse-time dynamics. We introduce a score-based objective that explicitly models the latent distribution and accommodates observation-specific noise covariances. Theoretically, we establish that the score estimation error with n independent observations achieves a near-parametric error rate of $\text{polylog}(n)/n$, improving upon existing results in the diffusion literature. Empirically, our method outperforms the nonparametric maximum likelihood estimator in both density estimation and denoising fidelity, especially in high-dimensional settings. These findings suggest a promising direction for integrating nonparametric empirical Bayes with diffusion-based generative modeling for latent structure recovery.

1 Introduction

In many large-scale scientific domains, such as the analysis of high-dimensional astronomical data, measurements are often corrupted by *heteroscedastic* noise, where error variances vary across dimensions and differ across observations [Kelly, 2012, Akritas and Bershad, 1996, Anderson et al., 2018]. While this noise complicates downstream inference, it is frequently well-characterized through the data collection process. A central goal in such settings is to recover the true latent signals and their underlying distribution from noisy observations by leveraging structural information inherent in the data. This problem can be naturally modeled using heteroscedastic Gaussian location mixtures, described by:

$$X_i = \theta_i^* + Z_i, \quad \text{with } \theta_i^* \stackrel{\text{i.i.d.}}{\sim} G^* \text{ and } Z_i \stackrel{\text{ind}}{\sim} \mathcal{N}(0, \Sigma_i), \quad \text{for } i = 1, \dots, n, \quad (1)$$

where $\{X_i\}_{i=1}^n \subseteq \mathbb{R}^d$ are the observed data with distinct error distributions, $\{\theta_i^*\}_{i=1}^n$ are the unobserved latent signals drawn from a common unknown prior measure G^* , and $\{Z_i\}_{i=1}^n$ are independent Gaussian noise with known positive-definite covariance matrices $\{\Sigma_i\}_{i=1}^n$.

Model (1) falls within the empirical Bayes (EB) framework [Robbins, 1951, 1983], where the goal is to estimate the prior distribution G^* and recover the latent vectors $\{\theta_i^*\}_{i=1}^n$. One way to achieve the goal is through the *G-modeling* approach [Efron, 2014], which is directly modeling the prior on the parameter scale and then plugged into the posterior calculation. A principled and widely used approach to estimate G^* , without imposing any parametric assumption on it, is the *nonparametric*

maximum likelihood estimator (NPMLE) [Robbins, 1950, Kiefer and Wolfowitz, 1956, Jiang and Zhang, 2009, Gu and Koenker, 2016, Polyanskiy and Wu, 2020]. The NPMLE seeks a distribution $\hat{G} \in \mathcal{P}(\mathbb{R}^d)$ that maximizes the marginal likelihood of the observed data $\{X_i\}_{i=1}^n$:

$$\hat{G}_{\text{MLE}} \in \arg \max_{G \in \mathcal{P}(\mathbb{R}^d)} \frac{1}{n} \sum_{i=1}^n \log f_{G, \Sigma_i}(X_i), \quad (2)$$

where $\mathcal{P}(\mathbb{R}^d)$ is the family of all probability measures on $\mathbb{R}^d$ and $f_{G, \Sigma_i}(x) := \int \varphi_{\Sigma_i}(x - \theta) dG(\theta)$ is the marginal density of X_i under G with $\varphi_{\Sigma_i}(y) := (\det(2\pi\Sigma_i))^{-1/2} \exp(-\frac{1}{2}y^\top \Sigma_i^{-1}y)$. Once $\hat{G}$ is obtained, each latent signal θ_i can be estimated using the *empirical Bayes* posterior mean:

$$\hat{\theta}_i := \mathbb{E}_{\hat{G}}[\theta \mid X_i, \Sigma_i] \stackrel{(*)}{=} X_i + \Sigma_i \nabla_x \log f_{\hat{G}, \Sigma_i}(X_i), \quad (3)$$

where $(*)$ is due to the Tweedie's formula [Efron, 2011], linking posterior means to $\nabla_x \log f_{\hat{G}, \Sigma_i}(X_i)$, the score function of the marginal density. This nonparametric approach flexibly adapts to latent structures without strong modeling assumptions on G^* . Theoretical guarantees for the NPMLE in this model are well established in the homoscedastic setting $\Sigma_i = \Sigma$ or the univariate setting $d = 1$: it achieves a *near-parametric* sample complexity rate $O(n^{-1}(\log n)^{O(d)})$, i.e., parametric up to logarithmic multiplicative factors, for both density estimation and Bayes regret risk [Jiang and Zhang, 2009, Saha and Guntuboyina, 2020]. Recently, the results have further been extended to multivariate and heteroscedastic setting in Soloff et al. [2025].

Despite these guarantees, the NPMLE and its associated Tweedie-based denoising procedure suffer from inherent *statistical and computational limitations* that render their practical effectiveness in high-dimensional settings: (a) Statistically, Tweedie's posterior mean estimator often has an *over-shrinkage bias*. As an example, consider a joint Gaussian model for (X, θ^*) : $\begin{bmatrix} X \\ \theta^* \end{bmatrix} \sim$

$\mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2I_d & I_d \\ I_d & I_d \end{bmatrix}\right)$. It follows that marginally $\theta^* \sim \mathcal{N}(0, I_d)$, and $X \mid \theta^* \sim \mathcal{N}(\theta^*, I_d)$.

However, the posterior mean of θ^* is $\mathbb{E}[\theta^* \mid X] = X/2 \sim \mathcal{N}(0, I_d/2)$, thus underestimating the variance of the latent signal. (b) Computationally, although problem (2) is a convex optimization in G , it is infinite-dimensional and difficult to solve in practice. Since it is known that there exists a solution $\hat{G}_{\text{MLE}}$ with support on at most n atoms, one common way to address the issue is to restrict G to be a discrete measure of the form $\sum_{i=1}^m w_i \delta_{\mu_i}$, with $\sum_{i=1}^m w_i = 1$ for some $m \leq n$. Optimizing over $\{(w_i, \mu_i)\}_{i=1}^m \subseteq \mathbb{R}_+ \times \mathbb{R}^d$ leads to a finite dimensional but nonconvex optimization, often tackled via the expectation-maximization (EM) algorithm [Laird, 1978]. However, EM suffers from slow convergence and may converge to local maxima due to the nonconvexity. To mitigate these issues, prior work has suggested fixing the atoms $\{\mu_i\}_{i=1}^m$ using either grid points over a compact region [Koenker and Mizera, 2014, Feng and Dicker, 2018] or the observed data itself [Lashkari and Golland, 2007]. Yet the grid-based approach scales poorly with dimension, and the data-driven strategy is suboptimal under heteroscedastic noise [Soloff et al., 2025]. Figure 1 further illustrates the over-shrinkage and computational challenges that arise from these limitations.

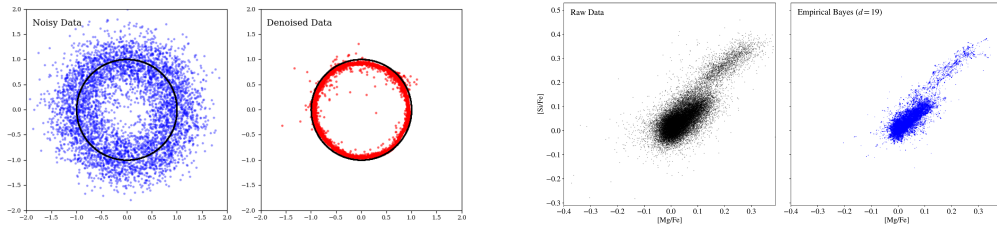


Figure 1: **Left**: 5,000 two-dimensional observations X_i (in blue) sampled from $X_i \mid \theta_i^* \sim \mathcal{N}(\theta_i^*, (0.3)^2 I_2)$ with θ^* uniformly distributed over a unit circle (in black). The posterior means (in red), obtained via the NPMLE and (3), are over-shrunk relative to the true unit circle. **Right** (adapted from Soloff et al. [2025, Figure 7]): a real-world 19-dimensional astronomy data (in black). The estimated posterior means (in blue) from NPMLE poorly reveal the underlying chemical structure (see Section 5 for the denoising results from our methods).

1.1 Our Methods and Contributions

To address the statistical and computational limitations of the classical NPMLE and Tweedie-based posterior denoising, we propose a novel perspective that reinterprets the data generation mechanism in model (1) through the lens of stochastic diffusion processes. Our method is built on two core ideas: replacing the NPMLE with score matching to estimate the prior within the G-modeling approach, and applying multi-step posterior denoising to recover latent signals. Below, we present an overview of our approach in the simplified homoscedastic setting; the full treatment of the heteroscedastic case is developed in detail in Section 2.

Modeling the data measurement process. Consider the homoscedastic version of model (1), where $\Sigma_i = \Sigma$ for all $i = 1, \dots, n$. In this special case, the observed data $\{X_i\}_{i=1}^n$ are i.i.d. samples from the convolution of the latent distribution G^* with the Gaussian kernel $\mathcal{N}(0, \Sigma)$:

$$X_i \sim G^* * \mathcal{N}(0, \Sigma), \quad \text{for } i = 1, \dots, n. \quad (4)$$

Rather than treating each observed data X_i as the result of a static additive noise model, we propose to interpret $\{X_i\}_{i=1}^n$ as snapshots from a forward stochastic diffusion process. Specifically, we model the measurement process as a zero-drift stochastic differential equation (SDE) that continuously perturbs samples from the latent prior G^* :

$$dX^{(t)} = g(t) \Sigma^{1/2} dB_t, \quad \text{with } X^{(0)} \stackrel{d}{=} \{\theta_i^*\}_{i=1}^n \sim G^*, \quad (5)$$

where $B_t \in \mathbb{R}^d$ denotes a standard Brownian motion and $g : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is the diffusion coefficient controlling the noise scaling. The function $g(t)$ can be adapted to reflect assumptions on the measurement process; for simplicity, one natural choice is $g(t) = \sqrt{\frac{2}{e^2-1}} e^t \Rightarrow \alpha(t) := \left(\int_0^t g(s)^2 ds \right) = \frac{e^{2t}-1}{e^2-1}$. Under this model, the marginal distribution of $X^{(t)}$ evolves according to:

$$X^{(t)} \sim G^* * \mathcal{N}\left(0, \Sigma^{(t)}\right), \quad \text{where } \Sigma^{(t)} := \alpha(t) \Sigma = \frac{e^{2t}-1}{e^2-1} \Sigma.$$

In particular, at time $t = 1$, we have $\Sigma^{(1)} = \Sigma$. Therefore, we can indeed view each X_i as sampled from the diffusion process (5) at $t = 1$, i.e., $\{X_i\}_{i=1}^n \sim X^{(1)}$.

Denoising the observations. The forward-time process in (5) is associated with a reverse-time diffusion process, which can be approximated by a deterministic flow that shares the same marginal distributions over time [Anderson, 1982, Maoutsa et al., 2020, Song et al., 2021a]. This flow is governed by the following ordinary differential equation (ODE):

$$d\bar{X}^{(t)} = -\frac{1}{2} g(t)^2 \Sigma \nabla_x \log f_{G^*, \Sigma^{(t)}}(\bar{X}^{(t)}) dt. \quad (6)$$

Given an estimator $\hat{G}$ of the prior G^* , this reverse time ODE naturally defines a generalized *multi-step* Tweedie-type empirical Bayes denoising procedure. Specifically, replacing G^* with $\hat{G}$ and discretizing the ODE backward in time from $t = 1$ to $t = 0$ with step size Δt yields the iterate:

$$\hat{\theta}_i^{(t-\Delta t)} = \hat{\theta}_i^{(t)} + \frac{\Delta t}{2} g(t)^2 \Sigma \nabla_x \log f_{\hat{G}, \Sigma^{(t)}}(\hat{\theta}_i^{(t)}) \quad \text{with initial condition } \hat{\theta}_i^{(1)} = X_i. \quad (7)$$

The final denoised estimate is taken as $\hat{\theta}_i := \hat{\theta}_i^{(0)}$ for $i = 1, \dots, n$. Notably, this method generalizes Tweedie’s formula in (3), which corresponds to a single-step approximation with $\Delta t = 1$. Our multi-step denoising procedure mirrors optimization dynamics, particularly Langevin-type flows over probability spaces, where smaller step sizes yield more controlled updates. This approach allows gradual refinement of the denoised signal, helping mitigate the over-shrinkage bias observed in one-step Tweedie denoising methods.

Learning the latent prior G^* . Our stochastic framework also naturally induces a denoising score matching (DSM) objective [Vincent, 2011] for learning the prior G^* . We propose to further perturb the noisy observations X_i using the forward SDE (5) to generate conditional samples from the kernel $p_{1t}(\cdot | X_i) = \varphi_{\Sigma^{(t)} - \Sigma}(\cdot - X_i)$, up to some stopping time T . We then match the score of the model’s marginal density to the conditional score induced by this perturbation:

$$\hat{G} := \arg \min_{G \in \mathcal{P}(\mathbb{R}^d)} \frac{1}{n} \sum_{i=1}^n \int_1^T \mathbb{E}_{p_{1t}(\cdot | X_i)} \left[g(t)^2 \left\| \nabla \log f_{G, \Sigma^{(t)}}(x) - \nabla \log p_{1t}(x | X_i) \right\|_\Sigma^2 \right] dt, \quad (8)$$

where $\|x\|_{\Sigma}^2 := x^\top \Sigma x$ denotes a weighted ℓ_2 -norm, extending the likelihood-weighting idea of Song et al. [2021b]. Intuitively, under our stochastic view, the observations at $t = 1$ already contain calibrated noise relative to G^* , and additional perturbation simulates conditional transitions. Matching model scores to these local conditional scores allows global recovery of the latent prior. Unlike standard diffusion models with the DSM objective that directly parameterize the score function with neural networks [Song et al., 2021a], we model the distribution G explicitly in (8). Such G -modeling is crucial in deriving new theoretical guarantees for the score estimation risk. Computationally, although the objective 8 is nonconvex, diffusion-type squared loss objectives have been empirically observed to exhibit smoother optimization landscapes and avoid poor local minima [Xu et al., 2024]. This empirical tractability is further supported by theory on the benign optimization behavior of overparameterized neural networks trained over mean-squared losses that can typically satisfy the Polyak–Łojasiewicz condition or behave linearly in the neural tangent kernel regime [Liu et al., 2022].

In summary, we propose a novel integration of score-based diffusion models with the classical EB framework via G -modeling. The rest of the paper is organized as follows: In Section 2, we present a stochastic process framework to model heteroscedastic Gaussian mixtures, where each observation is treated as a marginal realization from a distinct diffusion process with its own score function. Section 3 establishes the theoretical advantages of our approach, showing that G -modeling combined with diffusion yields near-parametric rates for score estimation. We further discuss how to numerically solve the G -modeling based score matching objective in Section 4. In Section 5, we present experimental results on both synthetic and real astronomical data, demonstrating substantial improvements in denoising and prior recovery over NPMLE and one-step Tweedie-based methods, supporting the improved computational scalability of our framework in high-dimensional settings.

1.2 Related Work

Score-based models for mixtures. Recent theory has provided provable guarantees for diffusion and score-based models in learning Gaussian mixtures, yet these results neither handle the heteroscedastic measurement setting we study nor aim at recovering the latent prior itself. Works such as Gatmiry et al. [2024], Chen et al. [2024], Shah et al. [2023] propose diffusion-based algorithms for mixture models under either finite discrete structures or continuous latent mixtures with homoscedastic noise. These approaches typically assume access to an infinite supply of i.i.d. samples and yield slow convergence guarantees of order $O((\log n)^{-1/p})$ in score or total variation distance, for some exponent $p > 1$. Importantly, these works directly parametrize the score function, using either neural networks or piecewise polynomial surrogates, without attempting nonparametric recovery of the latent mixing measure G^* or incorporate explicit G -modeling into the objective. Because of this direct score parametrization, these methods cannot be readily applied to our core task of latent recovery and are incompatible with our de-biased multi-step denoising procedure. A notable exception is Ghosh et al. [2025], which presents the first analysis of combining score matching with G -modeling, achieving a near-parametric rate $O(\log^6 n/n)$ for score estimation risk. However, their results are limited to the univariate, homoscedastic setting ($d = 1$). In contrast, our work tackles the more challenging regime of multivariate, heteroscedastic Gaussian location mixtures with arbitrary dimension $d \geq 1$. By coupling G -modeling with a stochastic process perspective, we develop an explicit estimator of the latent prior G^* and establish near-parametric rates for score estimation (up to small smoothing), along with finite-sample guarantees for recovering the latent signals.

Image denoising in computer vision. While our problem and the proposed method may appear superficially related to score-based diffusion approaches for image denoising, the settings and statistical challenges are fundamentally different. Most notably, our task lacks access to clean ground truth data, precluding the use of pre-trained score models, and all inference must be performed from noisy observations alone. In contrast, supervised denoising approaches such as BM3D [Dabov et al., 2006] requires access to a large number of paired noisy-clean images, while self-supervised methods like Noise2Score [Kim and Ye, 2021] rely on one-step Tweedie approximations that ignore the unknown prior G^* and leverage only partial model information Σ_i . These methods are either unrealistic or statistically inefficient in scientific domains where latent structure is present but clean data are unavailable, such as the astronomy measurement data. Even in moderate dimensions (e.g., $d = 20$), empirical Bayes is a highly nontrivial statistical task due to the data sparsity: each observation X_i corresponds to a unique latent parameter θ_i^* and an individual covariance Σ_i , yielding just one single noisy sample to infer each θ_i^* .

2 SDGM: Score-based Diffusion with G-Modeling

In this section, we present our general treatment of the score-based diffusion approach with G-modeling under the heteroscedastic setting. The key idea is still to reinterpret each noisy observation as a time-1 snapshot from a latent stochastic diffusion process. While our homoscedastic formulation serves as a motivating special case, the heteroscedastic setting introduces a crucial distinction: each observation X_i corresponds to a different SDE, despite all sharing a common latent initialization. Specifically, we consider the following family of zero-drift linear SDEs:

$$dX_i^{(t)} = g_i(t)\Sigma_i^{1/2}dB_t, \quad \text{with } X_i^{(0)} \sim G^*, \quad \text{for } i = 1, \dots, n, \quad (9)$$

where $\{g_i(t)\}_{i=1}^n \subset \mathbb{R}$ are the time-dependent diffusion coefficients. Without loss of generality, we assume $\alpha_i(t) := \int_0^t g_i(s)^2 ds$ satisfies $\alpha_i(1) = 1$ for all i . Then the time-dependent covariances are given by $\Sigma_i^{(t)} := \alpha_i(t)\Sigma_i$, and the marginal law of $X_i^{(t)}$ becomes

$$X_i^{(t)} \sim G^* * \mathcal{N}(0, \Sigma_i^{(t)}), \quad \text{for } t > 0, i = 1, \dots, n. \quad (10)$$

Thus, each observation X_i in the original mixture model (1) can be viewed as a marginal sample from this diffusion process at $t = 1$, i.e., $X_i \stackrel{d}{=} X_i^{(1)} \sim G^* * \mathcal{N}(0, \Sigma_i)$ for all $i = 1, \dots, n$.

As in the homoscedastic case, each forward SDE in (9) is coupled with the reverse-time ODE:

$$d\bar{X}_i^{(t)} = -\frac{1}{2}g_i(t)^2 \cdot \Sigma_i \cdot \nabla_x \log f_{G^*, \Sigma_i^{(t)}}(\bar{X}_i) dt, \quad \text{for } i = 1, \dots, n.$$

Given any plug-in estimator $\hat{G}$ for G^* , substituting $\Sigma_i^{(t)} = \Sigma_i^{(t)}$ in (7) defines a multi-step score-based denoising procedure that remains consistent with the EB framework. Furthermore, as $\Delta t \rightarrow 0$, the dynamics in (7) induces a deterministic transport map from the noisy empirical measure toward a latent measure $\hat{G} \approx G^*$. This perspective aligns with the design principles behind recent optimal transport approaches for empirical Bayes denoising [Soloff et al., 2025, Zhang et al., 2024], mitigating the over-shrinkage limitation of one-step Tweedie-type procedures.

Importantly, in our setting, the SDEs in (9) serve as a *belief model* for the true data-generating process, and the noisy observations in (1) are treated as marginal realizations from latent diffusion paths. This fundamentally differs from the generative score-based diffusion models like Song et al. [2021a], where SDEs are artificially constructed for the purpose of noise injection and score network training. In an idealized setting where the true latent measure G^* is known or accessible via unlimited samples, one could apply the framework of Song et al. [2021b,a] to learn a time-dependent score network $s_\theta(\cdot, t)$, mapping G^* to pure noise through an invertible path. However, our setting departs significantly from this setup. First, we observe only a finite collection of noisy samples at forward time $t = 1$, each generated by a unique, heteroscedastic SDE rather than homogeneous, artificially perturbed noise. Second, our goal is to explicitly estimate the latent distribution G^* , as opposed to modeling it implicitly via score functions. Finally, because no data is available at intermediate times $t \in (0, 1)$, the standard time-indexed score network $s_\theta(\cdot, t)$ cannot be trained in our setting.

To address these challenges, we propose a new objective that aggregates information across all diffusion paths to estimate G^* . Our method, which we term *score-based diffusion with G-modeling* (SDGM), obtains the estimator $\hat{G}$ by:

$$\hat{G} \in \arg \min_{G \in \mathcal{P}(\mathbb{R}^d)} \mathcal{J}_n(G) := \frac{1}{n} \sum_{i=1}^n \int_{t_0}^T g_i(t)^2 \mathbb{E}_{x|i,t} \left[\left\| \nabla \log q_G^{(t)}(x) - \nabla \log q_n^{(t)}(x) \right\|_{\Sigma_i}^2 \right] dt, \quad (11)$$

where, $t_0 \geq 1$ is the minimum diffusion time, and to simulate the homoscedastic behavior, we consider the *rescaled* Gaussian transition kernel: $x | i, t \sim K_t(\cdot | i) := \mathcal{N}([\alpha_i(t)\Sigma_i]^{-1/2}X_i, (1 - \alpha_i(t)^{-1})I_d)$ defined for $t \geq t_0 \geq 1$, and define the mixture densities $q_n^{(t)}$ and $q_G^{(t)}$ to be:

$$(\text{empirical}) \quad q_n^{(t)} := \frac{1}{n} \sum_{i=1}^n K_t(\cdot | i), \quad (G\text{-induced}) \quad q_G^{(t)} := \frac{1}{n} \sum_{i=1}^n f_{(\Sigma_i^{(t)})^{-1/2} \# G, I_d},$$

where $T \# G$ denotes the pushforward measure of G under map T such that for every set $A \subseteq \mathbb{R}^d$, $(T \# G)(A) = G(T^{-1}A)$. The defined mixtures satisfy that $\mathbb{E}q_n^{(t)} = q_{G^*}^{(t)}$. Intuitively, $q_G^{(t)}$ aggregates

the global behavior of the latent measure G across all SDEs, while $q_n^{(t)}$ serves as its empirical counterpart based on noisy observations. By matching the score functions of these two mixtures, we guide $\hat{G}$ toward an accurate approximation of G^* . In the homoscedastic special case with $\Sigma_i = \Sigma$ and $g_i = g$, the mixtures simplify to a rescaled version of the objective (11) that is equivalent to the score matching formulation in (8) up to constants.

3 Theoretical Guarantees

In this section, we provide guarantees on the statistical performance for both homoscedastic and heteroscedastic noise models. We primarily study three statistical properties to quantify how well the SDGM estimator $\hat{G}$ of (11) performs as a plug-in estimator for the true prior G^* .

The successful convergence of diffusion-based generative models is underlaid by a low score estimation error [Yakovlev and Puchkin, 2025, Bortoli, 2022]. Define the following *time-cumulative average Fisher divergence* for all path measures μ and ν :

$$\bar{\mathfrak{F}}_{[t_0, T]}(\mu || \nu) := \frac{1}{n} \sum_{i=1}^n \int_{t_0}^T \mathbb{E}_{x \sim \mu_i^{(t)}} \|\nabla \log \mu_i^{(t)} - \nabla \log \nu_i^{(t)}\|_2^2 dt.$$

Then, the specialized metric $\bar{\mathfrak{F}}_{[t_0, T]}(f_{G^*, \Sigma^{(t)}} || f_{\hat{G}, \Sigma^{(t)}})$ measures the time-cumulative risk on how $\nabla \log f_{\hat{G}, \Sigma^{(t)}}$ approximates $\nabla \log f_{G^*, \Sigma^{(t)}}$ along the forward-time SDE path beyond the observable time. It is consistent with the generalization risk concerned in Yakovlev and Puchkin [2025].

For density estimation, $\hat{G}$ also provides a natural estimate for the mixture marginal law $q_{G^*}^{(t)}$ aggregated from the paths of (9). We measure the quality of density estimation by the standard squared Hellinger accuracy, $\mathfrak{H}^2(p, q) := \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2$ between densities p and q . In particular, we are interested in the following squared Hellinger distances:

$$(\text{homoscedastic}) \quad \mathfrak{H}^2(f_{G^*, \Sigma^{(t)}}, f_{\hat{G}, \Sigma^{(t)}}), \quad (\text{heteroscedastic}) \quad \mathfrak{H}^2(q_{G^*}^{(t)}, q_{\hat{G}}^{(t)}),$$

which captures the global fit of the plug-in estimator $\hat{G}$ for estimating the joint law of $\{X_i^{(t)}\}_{i \geq 1}$.

The third metric of interest we establish is the squared Wasserstein distance (see e.g., Villani [2008]), which is defined for all $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ that $W_2^2(\mu, \nu) := \inf_{\gamma \in \Pi(\mu, \nu)} \mathbb{E}_{(x, y) \sim \gamma} [\|x - y\|_2^2]$ with $\Pi(\mu, \nu)$ denotes the set of all joint measures over μ, ν . The distance $W_2^2(\hat{G}, G^*)$ naturally characterizes the direct consistency of $\hat{G}$ for the true G^* .

We make the following assumptions on the latent structure and the data generation process (9). In the assumptions, the support conditions are widely considered in literature [Jiang and Zhang, 2009, Saha and Guntuboyina, 2020, Ghosh et al., 2025]. They encompass a wide range of realistic scenarios, including discrete mixtures, compactly supported manifold-supported latent structures, and representative experiment settings in Section 5. Furthermore, well-conditioned assumptions on noise covariances are ubiquitous in prior works [Soloff et al., 2025] and are essential to ensure smoothly-defined score everywhere and to enable information-theoretically meaningful latent recovery.

Assumption 1. *The true latent measure in model (1) is compactly supported, i.e., $G^* \in \mathcal{P}([-M, M]^d)$ for some finite radius $M > 0$. There exists $0 < \underline{\sigma} \leq \bar{\sigma} < \infty$ such that the covariances satisfy $\underline{\sigma} I_d \preceq \Sigma_i \preceq \bar{\sigma} I_d$ for all $i = 1, \dots, n$. The diffusion coefficient satisfies $\alpha_i(1) = 1$ (such that $\Sigma_i^{(1)} = \Sigma_i$) and other mild regularity conditions specified in the supplemental.*

Our first theorem characterizes the performance of our SDGM framework for the canonical homoscedastic model (4), which is specialized from our guarantee for the heteroscedastic model.

Theorem 1 (Homoscedastic noise). *Under Assumption 1 and further assume that $\Sigma_i = \Sigma$, $g_i = g$ for all $i = 1, \dots, n$. Let $\hat{G}$ minimizes objective (8) constrained over the family $\mathcal{P}([-M, M]^d)$ and with parameters t_0 and T satisfying $1 + \frac{1}{\log n} \leq \alpha(t_0) \leq \alpha(T) \leq 2$. Then provided n sufficiently large, the following bounds hold with probability at least $1 - n^{-2}$:*

Score and density estimation:

$$\mathbb{E}\tilde{\mathfrak{F}}_{[t_0, T]}(f_{G^*, \Sigma^{(t)}} \| f_{\hat{G}, \Sigma^{(t)}}) \lesssim_{M, d, (\underline{\sigma}, \bar{\sigma}), g} \frac{1}{n} (\log n)^{2d+3}, \quad (12)$$

$$\mathbb{E}\mathfrak{H}^2(f_{G^*, \Sigma^{(t_0)}}, f_{\hat{G}, \Sigma^{(t_0)}}) \lesssim_{M, d, (\underline{\sigma}, \bar{\sigma}), g} \frac{1}{n} (\log n)^{2d+3}. \quad (13)$$

Deconvolution:

$$\mathbb{E}W_2^2(G^*, \hat{G}) \lesssim \frac{1}{\log n}. \quad (14)$$

Here, we write $f(n) \lesssim g(n)$ if there exists a universal constant C independent of n such that $f(n) \leq Cg(n)$, $\forall n$. We write $\lesssim_{\theta_1, \theta_2, \dots}$ to emphasize that the constant $C = C_\theta$ also depends on the fixed specification parameters θ_i 's. In the following, we discuss implications of this result. Under a latent manifold setting with homoscedastic isotropic noise and neural parameterization of the score function, recent work [Yakovlev and Puchkin, 2025] proves the rate $\mathbb{E}\tilde{\mathfrak{F}}_{[t_0, T]}(f_{G^*, \Sigma^{(t)}} \| f_{\hat{G}, \Sigma^{(t)}}) = O(n^{-2\beta/(4\beta+d_0)} \text{polylog}(n))$, where d_0 is the intrinsic dimension of the manifold and β denotes a smoothness parameter. To our knowledge, this is the sharpest rate known for diffusion-based score estimation under such assumptions. By leveraging explicit G -modeling to impose the structure on the estimated scores, our result achieves an improved $O(n^{-1})$ leading dependence on sample size n (up to logarithmic factors), under more general assumptions. Moreover, our Hellinger accuracy bound (13), modulo a small distributional shift from $t = 1$ to t_0 , matches the rate achieved in NPMLE literature [Saha and Guntuboyina, 2020]. If the target is to estimate $f_{G^*, \Sigma}$ at the observational noise level, an unavoidable smoothing cost due to diffusion arises. In this regime, the task incurs the inevitable minimax rate $O(n^{-2/(d+4)})$ that cannot compete with NPMLE [Wibisono et al., 2024]. Finally, in the Gaussian deconvolution setting with fixed noise design, the minimax rate deteriorates to the logarithmic regime $O(1/\log n)$ due to inherent ill-posedness induced by Gaussian smoothing and it is the best rate possibly achieved by both NPMLE and our framework [Soloff et al., 2025, Dedecker and Michel, 2013]. We further discuss limitations of the current theory in Section 6.

Our next theorem generalizes the results to the heteroscedastic settings. The core contents of Theorem 2 parallels that of Theorem 1, but additional care is needed to account for observation-specific noise covariances and SDEs. To make the readers easy to follow, we present the main result informally below and defer detailed technical derivations and proofs of Theorem 1 and 2 to the supplemental material.

Theorem 2 (Heteroscedastic noise; Informal). *Under Assumption 1, provided solution $\hat{G}$ to (11) constrained over the family $\mathcal{P}([-M, M]^d)$ with (t_0, T) satisfying well-behaved smoothing condition, then with high probability,*

$$\begin{aligned} \mathbb{E}\tilde{\mathfrak{F}}_{[t_0, T]}(q_{G^*}^{(t_0)} \| q_{\hat{G}}^{(t_0)}) &\lesssim \frac{(\log n)^{2d+3}}{n}, & \mathbb{E}\mathfrak{H}^2(q_{G^*}^{(t_0)}, q_{\hat{G}}^{(t_0)}) &\lesssim \frac{(\log n)^{2d+3}}{n}, \\ \mathbb{E}W_2^2\left(\frac{1}{n} \sum_{i=1}^n (\Sigma_i)^{-1/2} \#G^*, \frac{1}{n} \sum_{i=1}^n (\Sigma_i)^{-1/2} \#\hat{G}\right) &\lesssim \frac{1}{\log n}. \end{aligned}$$

4 Solution Methods to Solve (11)

The optimization problem in (11) is an *infinite dimensional nonconvex* problem that is challenging to solve numerically. We discuss practical strategies for approximately computing the SDGM solution.

Strategy one: discretize the support of G . A classical approach to deal with the infinite dimensional variable G , as commonly used in NPMLE, is to restrict the class of priors to discrete measures with a finite number of supports: $G = \sum_{j=1}^m w_j \delta_{\mu_j}$, with $\sum_{j=1}^m w_j = 1$, $w_j \geq 0$. This reduces the optimization to a finite-dimensional, albeit nonconvex, problem over $\{(w_j, \mu_j)\}_{j=1}^m$.

Strategy two: continuous modeling with normalizing flows. An alternative approach is to model G continuously using normalizing flows [Rezende and Mohamed, 2015, Dinh et al., 2017, Durkan et al., 2019]. A normalizing flow represents a flexible distribution as the pushforward of a simple base distribution (e.g., standard Gaussian) through an invertible neural network. We parameterize G

using a Neural Spline Flow [Durkan et al., 2019] implemented via the `nflows` library [Durkan et al., 2020], allowing efficient sampling and density evaluation. In our framework, we draw Monte Carlo samples from the flow to estimate f_G and its score during training. Sampling from a continuous prior is theoretically justified by empirical Bayes approximation results: given a continuous G , there exists a discrete approximation G' that closely matches f_G and its gradient [Saha and Guntuboyina, 2020, Lemma D.3]. Furthermore, continuous modeling enables richer recovery of the underlying structure of G , which is crucial for scientific interpretation and denoising applications [Shen and Wu, 2022].

Batched and conditional estimation of q_G . The full mixture $q_G(x)$ plays a central role in our theoretical guarantees, but evaluating its score $\nabla_x \log q_G$ is computationally intensive for large n . A natural relaxation is to approximate q_G at each iteration using only a mini-batch of observations. In particular, we focus on the case where the mini-batch size $B = 1$: at each training step, we align the rescaled conditional score $\nabla_x \log f_{G, \Sigma_i^{(t)}}(x)$ with the empirical conditional score $\nabla_x \log \mathcal{N}(x; X_i, (\alpha_i(t) - 1)\Sigma_i)$. This corresponds to a *conditional score matching* update, mirroring the classical intuition of denoising score matching [Vincent, 2011], while offering computational simplicity and stability in stochastic optimization. In fact, subsampled estimation for mixture inference has also been used in early works [Celeux et al., 2001, Ihler et al., 2003].

5 Numerical Experiments

In this section, we evaluate the empirical performance of our proposed method on synthetic heteroscedastic datasets and real-world astronomy data. We also investigate whether the empirical scaling behavior aligns with our theoretical guarantees, particularly as we vary the problem dimension. We refer the readers to the supplemental for more implementation details and results.

Baselines and metrics. We evaluate two variants of our proposed method: SDGM-C, which uses a continuous prior modeled via normalizing flows, and SDGM-D, which uses a discrete prior. We compare our approaches against the state-of-the-art NPMLE solvers for the problem implemented in literature. Specifically, we compare with (1) NPMLE-N, the discretized NPMLE with supports taken fixed at noisy data (thus the objective (2) is optimized over only weights and becomes convex), as advocated in Lashkari and Golland [2007]; we adopt the Newton-based augmented Lagrangian method implemented from [Zhang et al., 2024]. (2) PEM, the partial EM method which is a generalized EM method proposed in Zhang et al. [2022] that used tighter sequential relaxations.

In addition, we consider a hybrid heuristic called SDGM-P to further improve the likelihood value defined in (2). This method first trains the SDGM-C model to learn a continuous prior, from which we sample support points. These support points are then used for a *post-processing* step via the convex NPMLE-N solver, effectively refining the prior estimate to maximize the likelihood.

We evaluate the performance of different methods over following metrics: (1) Fisher divergence (FD) defined by $\mathfrak{F}_n(f_G, f_*) := \frac{1}{n} \sum_{i=1}^n \|\nabla \log f_{G^*, \Sigma_i}(X_i) - \nabla \log f_{\hat{G}, \Sigma_i}(X_i)\|_2^2$. (2) Negative log-likelihood (NLL) defined by $-\frac{1}{n} \sum_{i=1}^n \log f_{\hat{G}, \Sigma_i}(X_i)$, which serves as a proximal for the KL divergence between $f_{\hat{G}, \Sigma_i}$ and f_{G^*, Σ_i} . (3) Squared Wasserstein-2 distance (W_2^2) between estimated $\hat{G}$ and G^* , approximated by the entropy-regularized optimal transport distance [Cuturi, 2013, Genevay et al., 2018]. (4) Mean squared error (MSE) between the denoised and clean data $\frac{1}{n} \sum_{i=1}^n \|\hat{\theta}_i - \theta_i^*\|_2^2$.

Synthetic datasets. We evaluate our method on two synthetic datasets with low-dimensional manifold structure, both of which are widely used in the NPMLE literature [Saha and Guntuboyina, 2020, Zhang et al., 2024]. Specifically, for $d \geq 2$, we generate clean latent vectors θ_i^* that exhibit nontrivial structure only in the first two dimensions, with the remaining dimensions padded with zeros:

Example 1 (circle): The first two coordinates of each θ_i^* are drawn uniformly from a circle of radius 6 centered at the origin.

Example 2 (uppercase letter ‘‘A’’): The first two coordinates of each θ_i^* are drawn uniformly from one of five line segments that form the shape of the capital letter ‘‘A.’’ These segments connect the points $(-4, 6)$, $(-2, 0)$, $(0, 6)$, $(2, 0)$, and $(4, 6)$, with each segment selected with equal probability $1/5$.

We then corrupt each θ_i^* with independently generated heteroscedastic Gaussian noise $Z_i \sim \mathcal{N}(0, \Sigma_i)$, where each Σ_i is a diagonal matrix. The diagonal entries of Σ_i are sampled independently and uniformly from the interval $[(0.8)^2, (1.2)^2]$.

Table 1: Models are trained with dimension $d = 16$ and sample size $n = 2 \times 10^4$. The $\downarrow$ indicates that lower values are better. We report generalization risks evaluated on a separate test dataset of size 10^4 . Denoising of the baseline methods NPMLE-N and PEM is performed via the one-step Tweedie’s formula, while our methods use the ODE solver. Results for stochastic methods are averaged over 5 independent runs.

G^*	circle				uppercase letter “A”			
metric	FD $\downarrow$	NLL $\downarrow$	$W_2^2 \downarrow$	MSE $\downarrow$	FD $\downarrow$	NLL $\downarrow$	$W_2^2 \downarrow$	MSE $\downarrow$
NPMLE-N	46.82	25.88	14.43	3.57	34.88	25.10	12.41	3.04
PEM	4.71	24.89	4.52	1.26	3.78	24.40	4.09	1.35
SDGM-C	1.27 ± 0.04	24.82 ± 0.01	2.07 ± 0.77	1.13 ± 0.01	1.23 ± 0.11	24.33 ± 0.01	1.68 ± 0.31	1.32 ± 0.03
SDGM-D	3.72 ± 0.05	24.88 ± 0.00	3.95 ± 0.03	1.64 ± 0.03	5.26 ± 0.05	24.41 ± 0.00	3.94 ± 0.00	1.73 ± 0.02
SDGM-P	0.68 ± 0.06	24.78 ± 0.00	1.07 ± 0.06	1.29 ± 0.02	0.49 ± 0.02	24.30 ± 0.00	1.25 ± 0.07	1.45 ± 0.02

The results are summarized in Table 1 and Figures 2 and 3, where $\hat{G}$ of our methods are plotted using the marginal density estimated through geodesic-KDE [Vincent and Bengio, 2002]. As shown in Table 1, SDGM consistently outperforms the NPMLE baselines across all metrics, with the additional post-processing step (SDGM-P) further improving the quality of prior estimation in most cases. The denoising effects of different methods are evident in Figures 2 and 3. Specifically, the estimated signals $\hat{\theta}_i$ from NPMLE and one-step Tweedie denoising appear overly blurred and shrunk, failing to recover the underlying manifold structure. In contrast, SDGM, equipped with multi-step ODE-based denoising, is able to clearly capture and preserve the latent geometry of the data.

Real-world astronomy data. We further demonstrate our SDGM-C on a challenging real-world scientific problem: estimating the latent distribution of stellar chemical abundances and denoising observations from the Apache Point Observatory Galactic Evolution Experiment (APOGEE) Data Release 14 [Abolfathi et al., 2018, Majewski et al., 2017]. With pre-processing in Ratcliffe et al. [2020], this dataset, comprising $n = 27,238$ stars with $d = 19$ chemical abundance measurements (e.g., [Fe/H], [Mg/Fe], [Si/Fe]), features heteroscedastic, element-correlated Gaussian noise for each star. The true underlying abundances θ_i^* are expected to reside on a lower-dimensional manifold governed by astrophysical nucleosynthetic processes [Ting et al., 2012, Weinberg et al., 2019]. Accurately revealing the underlying physical structures through the denoised abundances is crucial for reliable studies of Milky Way formation and evolution [Freeman and Bland-Hawthorn, 2002]. As depicted earlier in Figure 1, applying the NPMLE across the full 19-dimensional space is computationally prohibitive and produces unreliable estimates, as evidenced by Soloff et al. [2025], Zhang et al. [2024]. Notably, applying SDGM-C to the APOGEE data (Figure 4) reveals sharpened astrochemically significant structures. Taken together with our theoretical guarantees, these empirical results showcase that SDGM couples superior denoising fidelity with graceful high-dimensional scalability, making it a promising tool for robust inference in complex scientific datasets.

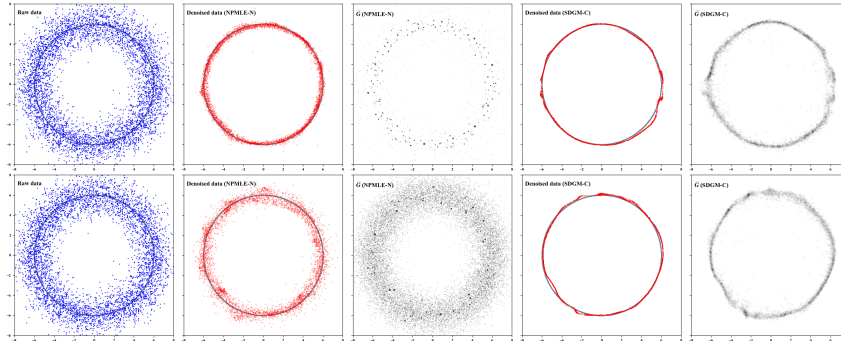


Figure 2: Projected plots of the circle experiment onto the first two dimensions. NPMLE-N (columns 2 & 3) performance significantly degrades when the dimension increases from $d = 8$ (top row) to $d = 16$ (bottom row). Our proposed SDGM-C (columns 4 & 5) stays efficient when d grows.

6 Discussions, Limitations, and Future Work

This work develops a stochastic-process reinterpretation of Gaussian location mixtures that bridges classical empirical Bayes inference and modern diffusion-based modeling. Through this lens, we introduce a diffusion-type G -modeling objective and a multi-step generalized Tweedie denoising procedure that jointly estimate the latent measure and recover the underlying signals from het-

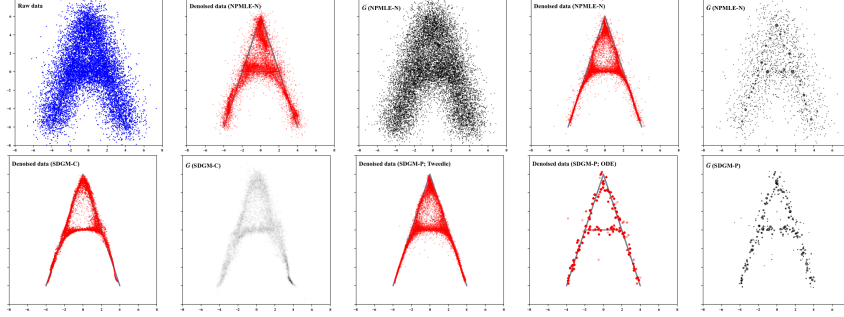


Figure 3: Projected plots of the letter A experiment in $d = 16$ onto the first two dimensions. NPMLE and PEM (top row) struggled by learning only a noisy latent prior, and causing denoised data to collapse within the region between the two legs of the A , illustrating over-shrinkage. The bottom row shows our SDGM variants. Notably, even $\hat{G}$ is estimated by SDGM, the one-step Tweedie denoising (middle of the bottom row) still over-shrinks θ_i . By contrast, the proposed denoising scheme better preserves the A structure, with fewer points confined to the center of the A .

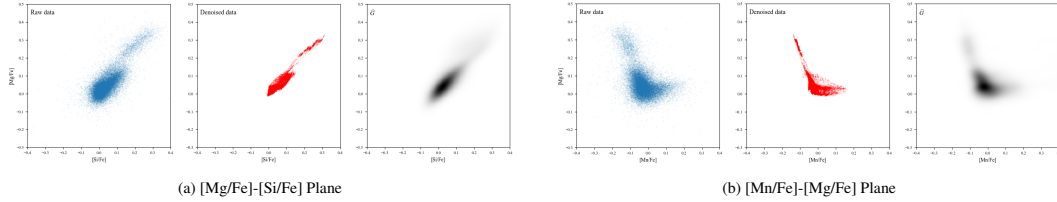


Figure 4: **Left:** Noisy APOGEE observations projected onto two 2-D planes. **Middle:** The denoised data reveals a fusiform-like or banana-like structure in the central region and a thin manifold structure in the upper region. **Right:** The marginal density of fitted $\hat{G}$ suggests strong abundance correlation among the chemicals. These results significantly improve the performance of NPMLE and Tweedie denoising in Figure 1.

eroscedastic mixtures. Unlike standard score-based diffusion approaches that directly parameterize the score function, our method explicitly models the latent prior distribution G , which is crucial for consistent recovery of the latent law. In contrast to traditional NPMLE solvers that suffer from slow convergence and over-shrinkage bias, our framework achieves comparable or improved denoising and prior recovery performance, with better empirical scalability in high-dimensional regimes.

Limitations. Our theoretical analysis builds on assumptions comparable to those in prior NPMLE studies [Saha and Guntuboyina, 2020, Soloff et al., 2025] and attains a near-parametric rate of $n^{-1}(\log n)^{O(d)}$. These results rely on standard regularity conditions, including compact support of the latent measure and well-conditioned covariance matrices, which are common in NPMLE theory but restrict general applicability in heavy-tailed or ill-conditioned settings. Although the rate matches the minimax lower bound in Hellinger distance for Gaussian mixture estimation [Kim and Guntuboyina, 2022], the dependence on the ambient dimension d remains unfavorable and limits its practicality in very high dimensions. This reflects the intrinsic difficulty of estimating a d -dimensional latent measure and recovering each θ_i from a single noisy sample X_i . Since our experiments show that the score-based G -modeling approach is more robust and computationally scalable than existing NPMLE solvers, and diffusion models have been observed to capture low-dimensional data manifolds [Pidstrigach, 2022], a promising future direction is to derive rates depending only on the intrinsic dimension $d_0 \ll d$ of the manifold supporting G^* , following assumptions similar to Yakovlev and Puchkin [2025]. Algorithmically, our current solution methods rely on heuristic sampling and approximations that do not guarantee convergence to global minima of the infinite-dimensional nonconvex objective. The formulation in Section 2 was primarily chosen to ensure theoretical tractability under the heteroscedastic setting and is not yet optimized for large-scale computation. Future work should aim to bridge this gap between theory and implementation by refining both the modeling formulation and algorithmic design, supported by further analysis of optimization and approximation errors within the empirical Bayes framework.

Broader implications. The line of work suggests that integrating diffusion-based generative modeling with nonparametric empirical Bayes principles can yield a unified approach to statistical deconvolution, manifold recovery, and inverse inference under heteroscedastic noise. Future investigations can further extend this framework toward manifold-aware score estimation and other related directions that enhance its theoretical scope and practical applicability.

Acknowledgment

The authors are partially supported by the National Science Foundation Division of Mathematical Sciences [Grant DMS-2416250] and the National Institutes of Health [Grant 1R01CA287413-01].

References

- Brandon C Kelly. Measurement error models in astronomy. In *Statistical challenges in modern astronomy V*, pages 147–162. Springer, 2012.
- Michael G Akritas and Matthew A Bershad. Linear regression for astronomical data with measurement errors and intrinsic scatter. *Astrophysical Journal*, 470:706, 1996.
- Lauren Anderson, David W Hogg, Boris Leistedt, Adrian M Price-Whelan, and Jo Bovy. Improving gaia parallax precision with a data-driven model of stars. *The Astronomical Journal*, 156(4):145, 2018.
- Herbert Robbins. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, volume 1, pages 131–148. University of California Press Berkeley, 1951.
- Herbert Robbins. Some thoughts on empirical Bayes estimation. *The Annals of Statistics*, pages 713–723, 1983.
- Bradley Efron. Two modeling strategies for empirical Bayes estimation. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(2):285, 2014.
- Herbert Robbins. A generalization of the method of maximum likelihood: Estimating a mixing distribution. *The Annals of Mathematical Statistics*, 21(2):314–315, 1950.
- Jack Kiefer and Jacob Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906, 1956.
- Wenhua Jiang and Cun-Hui Zhang. General maximum likelihood empirical bayes estimation of normal means. *The Annals of Statistics*, pages 1647–1684, 2009.
- Jiaying Gu and Roger Koenker. On a problem of robbins. *International Statistical Review*, 84(2): 224–244, 2016.
- Yury Polyanskiy and Yihong Wu. Self-regularizing property of nonparametric maximum likelihood estimator in mixture models. *arXiv preprint arXiv:2008.08244*, 2020.
- Bradley Efron. Tweedie’s Formula and Selection Bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.
- Sujayam Saha and Adityanand Guntuboyina. On the nonparametric maximum likelihood estimator for gaussian location mixture densities with application to gaussian denoising. *The Annals of Statistics*, 48(2):738–762, 2020.
- Jake A Soloff, Adityanand Guntuboyina, and Bodhisattva Sen. Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(1):1–32, 2025.
- Nan Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.
- Roger Koenker and Ivan Mizera. Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association*, 109(506):674–685, 2014.
- Long Feng and Lee H Dicker. Approximate nonparametric maximum likelihood for mixture models: A convex optimization approach to fitting arbitrary multivariate mixing distributions. *Computational Statistics & Data Analysis*, 122:80–91, 2018.

- Danial Lashkari and Polina Golland. Convex clustering with exemplar-based models. *Advances in neural information processing systems*, 20, 2007.
- Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.
- Dimitra Maoutsa, Sebastian Reich, and Manfred Opper. Interacting particle solutions of fokker-planck equations through gradient-log-density estimation. *Entropy*, 22(8):802, 2020.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021a.
- Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.
- Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34:1415–1428, 2021b.
- Tianshuo Xu, Peng Mi, Ruilin Wang, and Yingcong Chen. Towards faster training of diffusion models: An inspiration of a consistency phenomenon. *arXiv preprint arXiv:2404.07946*, 2024.
- Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59: 85–116, 2022.
- Khashayar Gatmiry, Jonathan Kelner, and Holden Lee. Learning mixtures of gaussians using diffusion models. *arXiv preprint arXiv:2404.18869*, 2024.
- Sitan Chen, Vasilis Kontonis, and Kulin Shah. Learning general gaussian mixtures with efficient score matching. *arXiv preprint arXiv:2404.18893*, 2024.
- Kulin Shah, Sitan Chen, and Adam Klivans. Learning mixtures of gaussians using the ddpm objective. *Advances in Neural Information Processing Systems*, 36:19636–19649, 2023.
- Sulagna Ghosh, Nikolaos Ignatiadis, Frederic Koehler, and Amber Lee. Stein’s unbiased risk estimate and Hyvärinen’s score matching. *arXiv preprint arXiv:2502.20123*, 2025.
- Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising with block-matching and 3d filtering. In *Image processing: algorithms and systems, neural networks, and machine learning*, volume 6064, pages 354–365. SPIE, 2006.
- Kwanyoung Kim and Jong Chul Ye. Noise2score: Tweedie’s approach to self-supervised image denoising without clean images. *Advances in Neural Information Processing Systems*, 34:864–874, 2021.
- Yangjing Zhang, Ying Cui, Bodhisattva Sen, and Kim-Chuan Toh. On efficient and scalable computation of the nonparametric maximum likelihood estimator in mixture models. *Journal of Machine Learning Research*, 25(8):1–46, 2024.
- Konstantin Yakovlev and Nikita Puchkin. Generalization error bound for denoising score matching under relaxed manifold assumption. *arXiv preprint arXiv:2502.13662*, 2025.
- Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL <https://openreview.net/forum?id=MhK5aXo3gB>. Expert Certification.
- Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2008.
- Andre Wibisono, Yihong Wu, and Kaylee Yingxi Yang. Optimal score estimation via empirical bayes smoothing. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4958–4991. PMLR, 2024.

- Jérôme Dedecker and Bertrand Michel. Minimax rates of convergence for wasserstein deconvolution with supersmooth errors in any dimension. *Journal of Multivariate Analysis*, 122:278–291, 2013.
- Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *International Conference on Machine Learning*, 2015.
- Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *International Conference on Learning Representations*, 2017.
- Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems*, 2019.
- Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. nflows: normalizing flows in PyTorch, November 2020. URL <https://doi.org/10.5281/zenodo.4296287>.
- Yandi Shen and Yihong Wu. Empirical bayes estimation: When does g -modeling beat f -modeling in theory (and in practice)? *arXiv preprint arXiv:2211.12692*, 2022.
- Gilles Celeux, Stéphane Chrétien, Florence Forbes, and Abdallah Mkhadri. A component-wise EM algorithm for mixtures. *Journal of Computational and Graphical Statistics*, 10(4):697–712, 2001.
- Alexander Ihler, Erik Sudderth, William Freeman, and Alan Willsky. Efficient multiscale sampling from products of gaussian mixtures. *Advances in Neural Information Processing Systems*, 16, 2003.
- Yangjing Zhang, Ying Cui, Bodhisattva Sen, and Kim-Chuan Toh. On efficient and scalable computation of the nonparametric maximum likelihood estimator in mixture models. *arXiv:2208.07514*, 2022.
- Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.
- Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.
- Pascal Vincent and Yoshua Bengio. Manifold parzen windows. *Advances in neural information processing systems*, 15, 2002.
- P. Abolfathi, D. S. Aguado, G. Aguilar, C. Allende Prieto, A. Almeida, T. T. Ananna, F. Anders, S. F. Anderson, B. H. Andrews, B. Anguiano, and et al. The Fourteenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the Extended Baryon Oscillation Spectroscopic Survey and from the Second Phase of the Apache Point Observatory Galactic Evolution Experiment. *The Astrophysical Journal Supplement Series*, 235(2):42, 2018. doi: 10.3847/1538-4365/aa9e8a.
- S. R. Majewski, R. P. Schiavon, P. M. Frinchaboy, C. Allende Prieto, and et al. The Apache Point Observatory Galactic Evolution Experiment (APOGEE). *The Astronomical Journal*, 154(3):94, 2017. doi: 10.3847/1538-3881/aa784d.
- Bridget L Ratcliffe, Melissa K Ness, Kathryn V Johnston, and Bodhisattva Sen. Tracing the assembly of the milky way’s disk through abundance clustering. *The Astrophysical Journal*, 900(2):165, 2020.
- Yuan-Sen Ting, K. C. Freeman, C. Kobayashi, G. M. De Silva, and J. Bland-Hawthorn. Chemical tagging with APOGEE: The role of dimensionality and metallicity. *Monthly Notices of the Royal Astronomical Society*, 427:882–893, 2012. doi: 10.1111/j.1365-2966.2012.22002.x.
- David H. Weinberg, Jon A. Holtzman, Sten Hasselquist, Melissa K. Ness, and et al. The Chemical Evolution of the Milky Way from APOGEE Abundances of α -Elements and Iron. *The Astrophysical Journal*, 876:65, 2019. doi: 10.3847/1538-4357/ab149a.
- Ken Freeman and Joss Bland-Hawthorn. The new galaxy: Signatures of its formation. *Annual Review of Astronomy and Astrophysics*, 40(1):487–537, 2002.

- Arlene KH Kim and Adityanand Guntuboyina. Minimax bounds for estimating multivariate gaussian location mixtures. *Electronic Journal of Statistics*, 16(1):1461–1484, 2022.
- Jakiw Pidstrigach. Score-based generative models detect manifolds. *Advances in Neural Information Processing Systems*, 35:35852–35865, 2022.
- Lanfeng Pan, Yehua Li, Kevin He, Yanming Li, and Yi Li. Generalized linear mixed models with gaussian mixture random effects: Inference and application. *Journal of multivariate analysis*, 175: 104555, 2020.
- Saharon Rosset and Ryan J Tibshirani. From fixed-x to random-x regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*, 2020.
- Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497 – 1537, 2005. doi: 10.1214/009053605000000282. URL <https://doi.org/10.1214/009053605000000282>.
- Yunwen Lei, Lixin Ding, and Yingzhou Bi. Local rademacher complexity bounds based on covering numbers. *Neurocomputing*, 218:320–330, 2016.
- Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.
- Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.
- Ricky T. Q. Chen. torchdiffeq, 2018. URL <https://github.com/rtqichen/torchdiffeq>.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Our abstract and introduction set out an overview of our methods, the theoretical guarantees, and empirical improvements that are detailed in Section 2, 3 and 5, respectively. We clearly introduce the scope of study and background context in the first few paragraphs of the introduction.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We discussed several non-positive implications with our theoretical results following Theorem 1. We also discussed computational challenges of our proposal in Section 4, with efforts to alleviate them.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: We detailed the most significant assumptions in the main text and presented a proof sketch. We aim to present complete technical derivations in our supplemental.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: Our major contribution is a statistical modeling perspective together with an optimization objective that can be implemented with various approaches. We discuss our approaches to approximately solve our proposed model. We detailed datasets, metrics, baselines and model variants in Section 5, and a pointer to the supplemental for implementation details, code, and more results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We detailed the complete synthetic data preparation process in the main text. For the APOGEE data we used, it is publicly available at <https://data.sdss.org/sas/dr14/apogee/vac/apogee-rc/cat/>. We provided the code together with usages in the supplemental.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Dataset generation procedures, noise levels, model variants and evaluation protocols are given in Section 5. We will put the architecture and hyperparameters for the continuous normalizing flow model in the supplemental.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Table 1 reports mean and standard deviation over 5 independent runs for each metric, if the method is run under stochastic optimization methods.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We plan to detail these details together with other implementation details in the supplemental.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: We acknowledge that we have reviewed the Code of ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discussed the impact of the problem in scientific domains. We demonstrated that our framework help to mitigate challenges in traditional methods, which has the potential to benefit future scientific inference tasks.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our work does not release a high-risk model or dataset.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cited the source of APOGEE data and relevant papers. The data is publicly available at <https://www.sdss4.org/>.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Our paper does not release any new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The work involves no human-subject data collection.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human-subject research is conducted.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core original method are developed without any LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

Supplemental Material for “Score-Based Diffusion Modeling for Nonparametric Empirical Bayes in Heteroscedastic Gaussian Mixtures”

A Code

Our code is available at <https://github.com/chgongyu/sdgm-neurips.git>.

B Proofs of Results in Section 3

B.1 Formal statements of Theorem 2

Assumption 2. For the SDEs defined in (9), assume

1. (latent measure) $G^* \in \mathcal{P}([-M, M]^d)$ for some radius $M \in [0, \infty)$.
2. (covariances) $\underline{\sigma}I_d \preceq \Sigma_i \preceq \bar{\sigma}I_d$ for all $i = 1, \dots, n$.
3. (diffusion coefficients) For all $i = 1, \dots, n$, $g_i : \mathbb{R}_+ \rightarrow \mathbb{R}$ is a continuous function. Let $\alpha_i(t) := \int_0^t g_i(s)^2 ds$. We require the normalization condition $\alpha_i(1) = 1$, and additionally the well-behavior condition: fix a $T > 1$, then for every $t \in [1, T]$, there exists $(\underline{g}, \bar{g}) > 0$ such that $\underline{g} \leq |g_i(t)|^2 \leq \bar{g}$.

Theorem 3. Provided conditions in Assumption 2, let (t_0, T) satisfy $t_0 \leq T$ and

$$\frac{1}{\log n} \leq \int_1^{t_0} g_i^2(t) dt \leq \int_1^T g_i^2(t) dt \leq 1.$$

Let $\hat{G}_n$ be an optimal solution to Objective (11) constrained over the measure class $\mathcal{P}([-M, M]^d)$, then provided n sufficiently large, with probability at least $1 - n^{-2}$,

- for score estimation:

$$\begin{aligned} \mathbb{E} \widetilde{\mathfrak{F}}_{[t_0, T]}(q_{G^*}^{(t)} \| q_{\hat{G}_n}^{(t)}) &:= \int_{t_0}^T \mathbb{E}_{x \sim q_{G^*}^{(t)}} \|\nabla \log q_{G^*}^{(t)}(x) - \nabla \log q_{\hat{G}_n}^{(t)}(x)\|_2^2 dt \\ &\leq C_{d, M, (\bar{\sigma}, \underline{\sigma}), (\underline{g}, \bar{g})} \frac{1}{n} (\log n)^{2d+3}; \end{aligned}$$

- for density estimation at $t = t_0$:

$$\mathbb{E} \mathfrak{S}^2(q_{G^*}^{(t_0)} \| q_{\hat{G}_n}^{(t_0)}) \leq C'_{d, M, (\bar{\sigma}, \underline{\sigma}), (\underline{g}, \bar{g})} \frac{1}{n} (\log n)^{2d+3},$$

- for the deconvolution risk,

$$\mathbb{E} W_2^2 \left(\frac{1}{n} \sum_{i=1}^n (\Sigma_i)^{-1/2} \# G^*, \frac{1}{n} \sum_{i=1}^n (\Sigma_i)^{-1/2} \# \hat{G}_n \right) \lesssim \frac{1}{\log n}.$$

B.2 Control of Fisher risk

Our arguments are built upon the “random- X ” regime with fixed design model (i.e., sample size n , latent prior G^* , and covariances $\{\Sigma_i\}_{i=1}^n$ are fixed). Conditioned on the fixed design, we can resample the dataset $\mathbf{X}_n = \{X_i\}_{i=1}^n$ from the stochastic processes $\{G^* * \mathcal{N}(0, \Sigma_i)\}_{i=1}^n$. This view is similar to, e.g., Pan et al. [2020] where the observed data as treated random realizations from a fixed design model; see also Rosset and Tibshirani [2020] for discussions on random versus fixed- X statistical viewpoints.

We recall the following definitions of the rescaled mixture laws and its empirical estimator:

$$q_G^{(t)} := \frac{1}{n} \sum_{i=1}^n f_{G_{i,t}, I_d}, \quad q_n^{(t)} = \frac{1}{n} \sum_{i=1}^n K_t(\cdot | i);$$

$$\mathbb{E}_{\mathbf{X}_n}[q_n^{(t)}] = q_G^{(t)},$$

where $G_{i,t} := (\Sigma_i^{(t)})^{-1/2} \# G$ denotes the pushforward measure under linear map induced by the covariance structure $\Sigma_i^{(t)}$ at time t , and the rescaled transition kernel that perturbs the i -th data pair (X_i, Σ_i) is:

$$K_t(\cdot | i) = \mathcal{N}\left([\alpha_i(t)\Sigma_i]^{-1/2}X_i, (1 - \alpha_i(t)^{-1})I_d\right).$$

We abbreviate the score function for $q_G^{(t)}$ and $q_n^{(t)}$ respectively as:

$$s_G^{(t)} := \nabla \log q_G^{(t)}, \quad s_n^{(t)} := \nabla \log q_n^{(t)}.$$

By definition, the cumulated Fisher divergence $\mathbb{E}_{\mathfrak{F}_{[t_0, T]}}(q_{G^*}^{(t)} || q_G^{(t)})$ between G^* and an arbitrary deterministic measure G is naturally related to the population score matching risk defined by:

$$\mathcal{J}_*(G) := \frac{1}{n} \sum_{i=1}^n \int_{t_0}^T g_i^2(t) \mathbb{E}_{x \sim f_{G_{i,t}^*, I_d}} \left[\left\| s_G^{(t)}(x) - s_{G^*}^{(t)}(x) \right\|_{\Sigma_i}^2 \right] dt.$$

It is clear that the Fisher risk can be upper bounded by $\mathcal{J}_*$ via rescaling of diffusion and noise coefficients:

$$\mathfrak{F}_{[t_0, T]}(q_{G^*}^{(t)} || q_G^{(t)}) \leq \sup_{i \in [n], t \in [t_0, T]} \{[g_i^2(t)\lambda_{\min}(\Sigma_i)]^{-1}\} \mathcal{J}_*(G).$$

Notice however that $\mathcal{J}_*(G)$ needs not to be the expectation of the finite- n empirical risk defined in eq.(11), i.e., $\mathbb{E}_{\mathbf{X}_n} \mathcal{J}_n(G) \neq \mathcal{J}_*(G)$ in general. This is due to that (1) the score estimator is a biased ratio estimator (though the density q_{G^*} and its gradient is unbiasedly estimated): $\mathbb{E}_{\mathbf{X}_n} s_n \neq s_{G^*}$, and (2) both the integrand functional s_n and the sampling law K_t are $\mathbf{X}_n$ -dependent thus they cannot be trivially decoupled.

Thus, we further define a decoupled risk with respect to population-level score s_{G^*} and data-dependent sampling law $K_i(\cdot | X_i)$:

$$\hat{\mathcal{J}}_*(G) := \frac{1}{n} \sum_{i=1}^n \int_{t_0}^T g_i^2(t) \mathbb{E}_{x \sim K_t(\cdot | i)} \left[\left\| s_G^{(t)}(x) - s_{G^*}^{(t)}(x) \right\|_{\Sigma_i}^2 \right] dt.$$

Then it is clear that $\mathbb{E}_{\mathbf{X}_n} \hat{\mathcal{J}}_*(G) = \mathcal{J}_*(G)$ for every deterministic G , because $\mathbb{E}[K_t(\cdot | i)] \equiv f_{G_{i,t}^*, I_d}$.

For the random estimator $\hat{G}_n$, the optimality condition of $\mathcal{J}_n(\hat{G}_n)$ implies a basic inequality that allows us to link $\mathcal{J}_*$ and $\mathcal{J}_n$ through $\hat{\mathcal{J}}_*$, as we detail below.

Lemma 1 (Basic inequality). *Every optimal solution $\hat{G}_n$ to (11) satisfies:*

$$\hat{\mathcal{J}}_*(\hat{G}_n) \leq 4 \mathcal{J}_n(G^*). \quad (15)$$

Proof of Lemma 1. We start with the following definition that presents a joint view over the tuple (i, t, x) to ease the integral representation in $\mathcal{J}_n$.

Definition 1 (Joint law over index, time and sample). *Fix the data $\{(X_i, \Sigma_i)\}_{i=1}^n$ and set*

$$Z = \sum_{i=1}^n \int_{t_0}^T g_i^2(t) dt.$$

Define the probability measure $\bar{\mathbb{P}}_n$ on $\{1, \dots, n\} \times [t_0, T] \times \mathbb{R}^d$ by

$$\bar{\mathbb{P}}_n\{i = k, t \in dt, x \in dx\} = \frac{g_k^2(t)}{nZ} \varphi_{(1-\alpha_k(t)^{-1})I_d}(x - [\alpha_k(t)\Sigma_k]^{-1/2}X_k) dt dx.$$

Equivalently:

$$\begin{cases} \Pr\{i = k\} = \frac{1}{n}, \\ (t \mid i = k) \text{ has density } \frac{g_k^2(t)}{\int_{t_0}^T g_k^2(u) du}, \\ (x \mid i = k, t) \sim K_t(\cdot \mid k). \end{cases}$$

With the above definitions, we first note that we can rewrite the objective (11) by:

$$\begin{aligned} \mathcal{J}_n(G) &= \frac{1}{n} \sum_{i=1}^n \int_{t_0}^T g_i^2(t) \mathbb{E}_{x \sim K_t(\cdot \mid i)} \|s_G^{(t)}(x) - s_n^{(t)}(x)\|_{\Sigma_i}^2 dt \\ &= Z \mathbb{E}_{(i,t,x) \sim \bar{\mathbb{P}}_n} \|s_G^{(t)}(x) - s_n^{(t)}(x)\|_{\Sigma_i}^2. \end{aligned}$$

By the optimality condition, provided $G \in \mathcal{P}(\mathbb{R}^d)$ is a minimizer of (11) and G^* is feasible, we have $\mathcal{J}_n(G) \leq \mathcal{J}_n(G^*)$. Using the quadratic expansion:

$$\|a - c\|_{\Sigma}^2 - \|b - c\|_{\Sigma}^2 = \|a - b\|_{\Sigma}^2 - 2(b - c)^{\top} \Sigma (a - b) \leq 0,$$

where we let $a = s_G^{(t)}$, $b = s_{G^*}^{(t)}$, and $c = s_n^{(t)}$, we obtain:

$$\mathbb{E}_{(i,t,x) \sim \bar{\mathbb{P}}_n} \|s_G^{(t)} - s_{G^*}^{(t)}\|_{\Sigma_i}^2 \leq 2 \mathbb{E}_{(i,t,x) \sim \bar{\mathbb{P}}_n} \left[(s_{G^*}^{(t)} - s_n^{(t)})^{\top} \Sigma_i (s_{G^*}^{(t)} - s_n^{(t)}) \right]. \quad (16)$$

By Cauchy-Schwarz inequality applied to the inner product term (16) and rearranging terms,

$$\mathbb{E}_{(i,t,x) \sim \bar{\mathbb{P}}_n} \left[\|s_G^{(t)}(x) - s_{G^*}^{(t)}(x)\|_{\Sigma_i}^2 \right] \leq 4 \mathbb{E}_{(i,t,x) \sim \bar{\mathbb{P}}_n} \left[\|s_{G^*}^{(t)}(x) - s_n^{(t)}(x)\|_{\Sigma_i}^2 \right],$$

which recovers the claim by rewriting the integrals back in $\mathcal{J}_n$ and $\hat{\mathcal{J}}_*$. $\square$

By the basic inequality eq. (15), we can now relate the empirical risk to the population risk as follows:

$$\mathcal{J}_*(\hat{G}_n) = [\mathcal{J}_*(\hat{G}_n) - \hat{\mathcal{J}}_*(\hat{G}_n)] + \hat{\mathcal{J}}_*(\hat{G}_n). \quad (17)$$

The first term is the uniform deviation between the true and empirical $\mathcal{J}_*$ over the feasible set of G , and the second term is at most four times the score empirical regret $\mathcal{J}_n(G^*)$ conditioned on the observed data $\mathbf{X}_n$. Below we control the two terms in expectation respectively.

B.2.1 Control of $\mathcal{J}_n(G^*)$

We first have:

$$\begin{aligned} \mathcal{J}_n(G^*) &= \int_{t_0}^T \frac{1}{n} \sum_{i=1}^n g_i^2(t) \mathbb{E}_{x \sim K_t(\cdot \mid i)} \left[\|s_{G^*}^{(t)}(x) - s_n^{(t)}(x)\|_{\Sigma_i}^2 \right] dt \\ &\leq \int_{t_0}^T \max_{i \in [n]} \{g_i^2(t) \lambda_{\max}(\Sigma_i)\} \mathbb{E}_{x \sim q_n^{(t)}} \left[\|s_{G^*}^{(t)}(x) - s_n^{(t)}(x)\|_2^2 \right] dt. \end{aligned}$$

To control this term, we integrate and generalize tools from Wibisono et al. [2024], Saha and Guntuboyina [2020] to prove our next lemma.

Lemma 2 (Control of score empirical regret). *For each $t \in [t_0, T]$ where (t_0, T) are specified in Theorem 3,*

$$\mathbb{E}_{\mathbf{X}_n} \left[\mathbb{E}_{x \sim q_n^{(t)}} \|s_{G^*}^{(t)}(x) - s_n^{(t)}(x)\|_2^2 \right] \leq C_{d,M,(\bar{\sigma}, \underline{\sigma})} \frac{1}{n} (\log n)^{d+3}. \quad (18)$$

We now build up the necessary toolbox to prove the above lemma. Following Jiang and Zhang [2009], Saha and Guntuboyina [2020], it is beneficial to consider the following regularized score rule. Given an arbitrary density ρ , define its associated regularized score function with parameter $\epsilon > 0$ as:

$$s^\epsilon := \frac{\nabla \rho}{\rho \vee \epsilon},$$

where $\rho \vee \epsilon$ denotes $\max\{\rho, \epsilon\}$. For a fixed t , when it is clear in context, we now drop the index (t) in the score. Expand the empirical regret via regularization:

$$\mathbb{E}_{q_n} \left[\|s_{G^*} - s_n\|^2 \right] \leq 3\mathbb{E}_{q_n} \left[\|s_{G^*} - s_{G^*}^\epsilon\|^2 + \|s_{G^*}^\epsilon - s_n^\epsilon\|^2 + \|s_n - s_n^\epsilon\|^2 \right].$$

As useful in the following proof, we first define a general class of heteroscedastic mixtures, in which both the base measures G and the Gaussian smoothing kernels can be heteroscedastic across components. We then show that their associated regularized score function is uniformly bounded. The definition will be useful to characterize the behavior of the empirical score s_n , in which each component is a different Dirac measure $\delta_{(\alpha_i \Sigma_i)^{-1/2} X_i}$ smoothed by a personalized transition kernel $\mathcal{N}(0, K_i)$ with $K_i = (1 - \alpha_i^{-1})I_d$.

Definition 2 (General heteroscedastic mixtures). *Let $\{K_i\}_{i=1}^n$ be positive definite covariance matrices such that $\underline{k}I_d \preceq K_i \preceq \bar{k}I_d$. Let $G_{[n]} := \{G_i\}_{i=1}^n$ be a collection of probability measures on $\mathbb{R}^d$. Define the heteroscedastic mixture $p_{G_{[n]},h} = \frac{1}{n} \sum_{i=1}^n G_i * \mathcal{N}(0, hK_i)$ for bandwidth parameter $h > 0$.*

Lemma 3 (Uniform bounding of regularized scores). *Fix a mixture density defined in Definition 2. For every $x \in \mathbb{R}^d$,*

$$\frac{\|\nabla p_{G_{[n]},h}(x)\|_2}{p_{G_{[n]},h}(x) \vee \epsilon} \leq \sqrt{\frac{1}{h\underline{k}} \log \frac{(2\pi h\underline{k})^{-d}}{\epsilon^2}},$$

for $0 < \epsilon \leq (2\pi h\underline{k})^{-d/2} e^{-1/2}$.

Proof. For each component $p_i := G_i * \mathcal{N}(0, hK_i)$ in the mixture, consider the transformation such that if $X \sim p_i$, let $\tilde{X}_i := (hK_i)^{-1/2} X_i$ and let $\tilde{p}_i = \tilde{G}_i * \mathcal{N}(0, I_d)$ be the transformed measure governing $\tilde{X}_i$. Then we have the following equations:

$$\begin{aligned} p_i(x) &= |hK_i|^{-1/2} \tilde{p}_i \left((hK_i)^{-1/2} x \right), \\ \nabla p_i(x) &= |hK_i|^{-1/2} (hK_i)^{-1/2} \nabla \tilde{p}_i \left((hK_i)^{-1/2} x \right). \end{aligned}$$

For $\tilde{p}_i$ being convolution of measure with standard Gaussian kernel, Saha and Guntuboyina [2020, Lemma F.1] directly applies:

$$\begin{aligned} \frac{\|\nabla \tilde{p}_i\|}{\tilde{p}_i} &\leq \sqrt{\log \frac{(2\pi)^{-d}}{\tilde{p}_i^2}} \\ \implies \frac{\|\nabla p_i\|}{p_i} &= \frac{\|(hK_i)^{-1/2} \nabla \tilde{p}_i\|}{\tilde{p}_i} \leq \sqrt{\frac{1}{h\underline{k}} \log \frac{(2\pi)^{-d}}{\tilde{p}_i^2}} \leq \sqrt{\frac{1}{h\underline{k}} \log \frac{(2\pi h\underline{k})^{-d}}{p_i^2}}. \end{aligned}$$

We observe that the function $x\sqrt{\log \beta x^{-2}}$ is concave over $x \in (0, \sqrt{\beta})$, thus by Jensen's inequality:

$$\sum_{i=1}^n \|\nabla p_i\| \leq n p_{G_{[n]},h} \sqrt{\frac{1}{h\underline{k}} \log \frac{(2\pi h\underline{k})^{-d}}{p_{G_{[n]},h}^2}},$$

and

$$\frac{\|\nabla p_{G_{[n]},h}\|}{p_{G_{[n]},h}} \leq \frac{\sum_i \|\nabla p_i\|}{n p_{G_{[n]},h}} \leq \sqrt{\frac{1}{h\underline{k}} \log \frac{(2\pi h\underline{k})^{-d}}{p_{G_{[n]},h}^2}}.$$

The claim then follows from a same argument as in proof of (F.2) in Lemma F.1 of Saha and Guntuboyina [2020] to relate the regularized rule with the unregularized bound derived above under the two cases $p_{G_{[n]},h} \leq \epsilon$ and $p_{G_{[n]},h} > \epsilon$. $\square$

We now turn our attention to tools for bounding the discrepancy between regularized scores:

$$\mathbb{E}_{q_n} [\|s_{G^*}^\epsilon - s_n^\epsilon\|^2],$$

via their corresponding Hellinger distance. We rely on specializing Theorem E.1 of Saha and Guntuboyina [2020] and Lemma 1 of Wibisono et al. [2024] to the following form that is closely related to our setting with heteroscedastic mixture.

Lemma 4 (Bayes discrepancy of regularized scores). *For two arbitrary n -collections of probability measures $G_{[n]}$ and $H_{[n]}$ on $\mathbb{R}^d$. Let the mixtures $p_{G_{[n]},h}$ and $p_{H_{[n]},h}$ be defined as in Definition 2, then*

$$\begin{aligned} & \mathbb{E}_{p_{G_{[n]},h}} \left[\left\| s_{G_{[n]},h}^\epsilon - s_{H_{[n]},h}^\epsilon \right\|_2^2 \right] \\ & \leq \frac{Cd}{h\underline{k}} \max \left\{ \left(\log \frac{(2\pi h\underline{k})^{-d/2}}{\epsilon} \right)^3, |\log \mathfrak{H}(p_{G_{[n]},h}, p_{H_{[n]},h})| \right\} \mathfrak{H}^2(p_{G_{[n]},h}, p_{H_{[n]},h}), \end{aligned}$$

if $h\underline{k} \leq 1$ and $0 < \epsilon \leq (2\pi h\underline{k})^{-d/2} e^{-1/2}$.

Proof of Lemma 4. Examining the proof of Saha and Guntuboyina [2020, Theorem E.1], we have:

$$\mathbb{E}_{p_{G_{[n]},h}} \left[\left\| s_{G_{[n]},h}^\epsilon - s_{H_{[n]},h}^\epsilon \right\|_2^2 \right]^{1/2} \leq T_1 + T_2,$$

with

$$\frac{1}{4}T_1^2 \leq 2\mathfrak{H}^2(p_{G_{[n]},h}, p_{H_{[n]},h})\Upsilon^2, \quad \text{and} \quad \frac{1}{4}T_2^2 \leq \sum_{i=1}^d \Delta_{i,1}^2,$$

where

$$\Upsilon^2 := \sup_{x \in \mathbb{R}^d, \circ \in \{G_{[n]}, H_{[n]}\}} \left\{ \frac{\|\nabla p_{\circ,h}(x)\|}{p_{\circ,h}(x) \vee \epsilon} \right\}^2, \quad \Delta_{i,k}^2 := \int \frac{(\partial^k / \partial x_i^k (p_{G_{[n]},h} - p_{H_{[n]},h}))^2}{p_{G_{[n]},h} \vee \epsilon + p_{H_{[n]},h} \vee \epsilon}.$$

With integration by parts, it is shown there that Δ 's satisfy a recursion that for $k \geq 1$,

$$\Delta_{i,k}^2 \leq 2\Upsilon \Delta_{i,k-1} \Delta_{i,k} + \Delta_{i,k-1} \Delta_{i,k+1}.$$

With generalized Lemma F.2 of Saha and Guntuboyina [2020] and some work around the recursion, we can deduce that for every $a \geq \sqrt{(2k-1)/(h\underline{k})}$:

$$\Delta_{i,k}^2 \leq \frac{2(2\pi h\underline{k})^{-d/2}}{\epsilon} \left[a^{2k} \mathfrak{H}^2(p_{G_{[n]},h}, p_{H_{[n]},h}) + \sqrt{\frac{2}{\pi}} (a^2/h\underline{k})^{(2k-1)/2} e^{-h\underline{k}a^2} \right],$$

and further with $k = k_0 + 1 \geq \frac{h\underline{k}}{2} \Upsilon^2$ and taking $a\sqrt{h\underline{k}} = \max \left\{ \sqrt{2k_0 + 1}, \sqrt{2|\log \mathfrak{H}(p_{n,h}, p_{*,h})|} \right\}$, we have

$$e^{-h\underline{k}a^2} \leq \mathfrak{H}^2(p_{G_{[n]},h}, p_{H_{[n]},h}),$$

and

$$\Delta_{i,1} \leq C \max \{2k_0 \Upsilon, e(1 + \sqrt{h\underline{k}})a\} \mathfrak{H}(p_{n,h}, p_{*,h}).$$

Plugging in the uniform upper bound on Υ by Lemma 3 and some basic algebra around the choice of k_0 and merging of terms yield the desired bound. Because the last algebraic steps exactly follow from Saha and Guntuboyina [2020, Proof of Theorem E.1], we refer our readers to their work for more details. $\square$

The above result implies that we can further reduce the discrepancy between regularized scores to bounding the Hellinger distance between q_{G^*} and its empirical estimate q_n . We recall the following result from Wibisono et al. [2024] that controls the Hellinger convergence rate of smoothed empirical distributions with the vanilla radial Gaussian kernel.

Lemma 5 ([Wibisono et al., 2024, Lemmas 2 & 3]). *Let ρ be an α -subgaussian measure on $\mathbb{R}^d$ whose score $s(x) = \nabla \log \rho(x)$ is L -Lipschitz. Let $\rho_h = \rho * \varphi_h$ where φ_h is the density of $\mathcal{N}(0, hI_d)$. Then for every $x \in \mathbb{R}^d$,*

$$\frac{\rho(x)}{\rho_h(x)} \leq \exp(dLh/2). \quad (19)$$

Let ρ_n be its size- n empirical measure and $\rho_{n,h} = \rho_n * \mathcal{N}(0, hI_d)$. Then provided $h \leq \min\{\frac{1}{4L}, \alpha^2\}$,

$$\mathbb{E}[\mathfrak{H}^2(\rho_{n,h}, \rho_h)] \leq \frac{1}{n} \left(\frac{C\alpha^2 \log n}{h} \right)^{d/2} + \frac{4d}{n},$$

where $C > 0$ is a universal constant.

We extend the above lemma for empirical heteroscedastic kernel densities.

Corollary 1 (Hellinger convergence of heteroscedastically smoothed empirical measures). *Let $\{Y_i\}_{i=1}^n$ be independent random variables, each drawn from an α -subgaussian, L -log-smooth measure p_i on $\mathbb{R}^d$. Consider the heteroscedastic mixture p_n with density: $p_{n,h}(x) := \frac{1}{n} \sum_{i=1}^n \varphi_{h_i I_d}(x - Y_i)$ and denote its population mixture $p_{*,h} = \mathbb{E}[p_{n,h}] = \frac{1}{n} \sum_{i=1}^n p_i * \mathcal{N}(0, hK_i)$. Assume each $\underline{k}I_d \preceq K_i \preceq \bar{k}I_d$, and $h\bar{k} \leq \min\{\alpha^2, 1/2L\}$. Then,*

$$\mathbb{E} [\mathfrak{H}^2(p_{n,h}, p_{*,h})] \leq \frac{C_{\underline{k}, \bar{k}, \alpha}}{n} \left(\frac{\log n}{h} \right)^{d/2} + \frac{4d}{n}, \quad (20)$$

for a constant C that depends on $\alpha, \bar{k}$.

Proof of Corollary 1. By proof of Lemma 2 in Wibisono et al. [2024], let $B \subset \mathbb{R}^d$,

$$\mathbb{E} [\mathfrak{H}^2(p_{n,h}, p_{*,h})] \leq \int_B \frac{\mathbb{E}[(p_{*,h}(x) - p_{n,h}(x))^2]}{p_{*,h}(x)} dx + 2 \int_{B^c} p_{*,h}$$

Fix each x , then for each i -th component of the empirical mixture $\varphi_{hK_i}(x - Y_i)$, by their trick:

$$\mathbb{E} \varphi_{hK_i}(x - Y_i)^2 = |4\pi h K_i|^{-1/2} \mathbb{E} \varphi_{hK_i/2}(x - Y_i).$$

Thus

$$\begin{aligned} \text{Var}[p_{n,h}(x)] &= \frac{1}{n} \sum_{i=1}^n \text{Var}[\varphi_{hK_i}(x - Y_i)] \\ &= \frac{1}{n} \sum_{i=1}^n \left(|4\pi h K_i|^{-1/2} \mathbb{E}[\varphi_{hK_i/2}(x - Y_i)] - \mathbb{E}[\varphi_{hK_i}(x - Y_i)]^2 \right) \leq \frac{(4\pi h \underline{k})^{-d/2}}{n} p_{*,h/2}(x). \end{aligned}$$

Combining with above,

$$\mathbb{E} [\mathfrak{H}^2(p_{n,h}, p_{*,h})] \leq \frac{(4\pi h \underline{k})^{-d/2}}{n} \int_B \frac{p_{*,h/2}(x)}{p_{*,h}(x)} dx + 2 \int_{B^c} p_{*,h}.$$

Notice that $p_{*,h}$ is subgaussian with parameter at most $\sigma = \sqrt{\alpha^2 + h\bar{k}}$. By letting B be the box region with center $\mathbb{E}_{p_*}[\bar{Y}]$ and radius $\sigma\sqrt{\log n}$, standard concentration inequalities for multivariate subgaussian variables yield $\int_{B^c} p_{*,h} \leq \frac{2d}{n}$, and for the first term we can bound:

$$\int_B \frac{p_{*,h/2}(x)}{p_{*,h}(x)} dx \leq \text{Vol}(B) \sup_x \frac{p_{*,h/2}(x)}{p_{*,h}(x)} = (4\sigma^2 \log n)^{d/2} \sup_x \frac{p_{*,h/2}(x)}{p_{*,h}(x)},$$

where the last term can be bounded by applying an immediate extension from (19) and Lemma 6 of Wibisono et al. [2024]. For each component in the mixture, consider the rotation $\tilde{p}_{i,h} = K_i^{-1/2} \# p_i * \mathcal{N}(0, hI_d)$. Then notice $\nabla \log \tilde{p}_{i,h} = K_i^{1/2} \nabla \log p_{i,h}$. For all $h > 0$ and $x \in \mathbb{R}^d$, let $\tilde{x} = K_i^{-1}x$ such that:

$$\begin{aligned} \log \frac{p_{i,h}}{p_{i,2h}}(x) &= \log \frac{\tilde{p}_{i,h}}{\tilde{p}_{i,2h}}(K_i^{-1/2}x) \\ &\leq \mathbb{E}_{Z \sim \mathcal{N}(0, I_d)} [\log \tilde{p}_{i,h}(\tilde{x}) - \log \tilde{p}_{i,h}(\tilde{x} - \sqrt{h}Z)] \\ &= \mathbb{E}_{Z \sim \mathcal{N}(0, I_d)} \int_{-\sqrt{h}}^0 \langle -Z, \nabla \log \tilde{p}_{i,h}(\tilde{x} - uZ) \rangle du \\ &\leq \mathbb{E} \int_{-\sqrt{h}}^0 \|Z\| \cdot \|\nabla \log \tilde{p}_{i,h}(\tilde{x} - uZ) - \nabla \log \tilde{p}_{i,h}(\tilde{x})\| du \leq \bar{k}Ldh, \end{aligned}$$

which implies $p_{*,h}(x)/p_{*,2h}(x) \leq \exp(\bar{k}Ldh)$.

Combining the results, we obtain:

$$\mathbb{E} [\mathfrak{H}^2(p_{n,h}, p_{*,h})] \leq \frac{((\frac{4\alpha^2}{n} + h\bar{k}) \log n / 4\pi h \underline{k})^{d/2}}{n} + \frac{4d}{n} \leq \frac{C_d}{n} \left(\frac{4\alpha^2/h + \bar{k}}{\underline{k}} \log n \right)^{d/2} + \frac{4d}{n}.$$

□

Lastly, we study the discrepancy due to regularizing q_n and q_{G^*} , which can be controlled via adapting results from Saha and Guntuboyina [2020]. Specifically, we recall the following result:

Lemma 6 (Saha and Guntuboyina [2020, Lemma 4.3]). *For a probability measure $G \in \mathcal{P}(\mathbb{R}^d)$ and let $\epsilon \in (0, (2\pi)^{-d/2}e^{-1/2}]$. Let $L(\epsilon) := \sqrt{-\log((2\pi)^d\epsilon^2)}$. Then for every compact set $S \subseteq \mathbb{R}^d$,*

$$\Delta(G, \epsilon) = \int \left(1 - \frac{f_{G, I_d}}{f_{G, I_d} \vee \epsilon}\right)^2 \frac{\|\nabla f_{G, I_d}\|_2^2}{f_{G, I_d}} \leq C_d N(4/L(\epsilon), S) L^d(\epsilon) \epsilon + dG(S^c),$$

where $N(\delta, S)$ is the regular Euclidean δ -covering of set S .

Intuitively, the score and its regularized counterpart differ only in the tails.

Corollary 2. *For an arbitrary n -mixture of probability measures $G_{[n]} = \frac{1}{n} \sum_{i=1}^n G_i \in \mathcal{P}(\mathbb{R}^d)$, and let the heteroscedastic mixture $p_{G_{[n]}, h}$ defined as in Definition 2. Let $\epsilon \in (0, (2\pi h\bar{k})^{-d/2}e^{-1/2}n^{-1})$ and define $L(\epsilon) := \sqrt{-\frac{1}{h\bar{k}} \log((2\pi h\bar{k})^d n^2 \epsilon^2)}$. For compact sets $S_1, \dots, S_n \subseteq \mathbb{R}^d$,*

$$\mathbb{E}_{p_{G_{[n]}, h}} [\|s_{G_{[n]}, h} - s_{G_{[n]}, h}^\epsilon\|^2] \leq C_{d, \bar{k}, \underline{k}} \frac{1}{n} \sum_{i=1}^n \left(\frac{1}{h} N(4/L(\epsilon), S_i) L^d(\epsilon) \epsilon + \frac{d}{h} G_i(S_i^c) \right).$$

Proof of Corollary 2. The strategy to prove this slightly generalized version is similar to proof of Lemma 3, the Proof of Lemma F.7 in Saha and Guntuboyina [2020] and proofs in Section D.2 of Soloff et al. [2025]. Specifically, for every fixed x , we can let $\tilde{G}_i$ denote the distribution of $(hK_i)^{-1/2}\theta_i$ where $\theta_i \sim G_i$ and write $x_i = (hK_i)^{-1/2}x$. This allow us to rescale the terms in the score function:

$$\begin{aligned} p_{G_{[n]}, h}(x) &= \frac{1}{n} \sum_{i=1}^n |hK_i|^{-1/2} f_{\tilde{G}_i, I_d}(x_i), \\ \nabla p_{G_{[n]}, h}(x) &= \frac{1}{n} \sum_{i=1}^n |hK_i|^{-1/2} (hK_i)^{-1/2} \nabla f_{\tilde{G}_i, I_d}(x_i). \end{aligned}$$

By convexity,

$$\begin{aligned} \frac{\|\nabla p_{G_{[n]}, h}(x)\|^2}{p_{G_{[n]}, h}(x)} &\leq \frac{1}{n} \sum_{i=1}^n \frac{\| |hK_i|^{-1/2} (hK_i)^{-1/2} \nabla f_{\tilde{G}_i, I_d}(x_i) \|^2}{|hK_i|^{-1/2} f_{\tilde{G}_i, I_d}(x_i)} \\ &\leq (h\bar{k})^{-(d/2+1)} \frac{1}{n} \sum_{i=1}^n \frac{\|\nabla f_{\tilde{G}_i, I_d}(x_i)\|^2}{f_{\tilde{G}_i, I_d}(x_i)}, \end{aligned}$$

and it is trivial to verify that for every $i \in [n]$:

$$\begin{aligned} p_{G_{[n]}, h}(x) &\geq \frac{1}{n} (h\bar{k})^{-d/2} f_{\tilde{G}_i, I_d}(x_i) \\ \implies \left(1 - \frac{p_{G_{[n]}, h}(x)}{p_{G_{[n]}, h}(x) \vee \epsilon}\right)^2 &\leq \left(1 - \frac{f_{\tilde{G}_i, I_d}(x_i)}{f_{\tilde{G}_i, I_d}(x_i) \vee n(h\bar{k})^{d/2}\epsilon}\right)^2. \end{aligned}$$

Thus,

$$\begin{aligned} \mathbb{E}_{p_{G_{[n]}, h}} [\|s_{G_{[n]}, h} - s_{G_{[n]}, h}^\epsilon\|^2] &= \int \left(1 - \frac{p_{G_{[n]}, h}(x)}{p_{G_{[n]}, h}(x) \vee \epsilon}\right)^2 \frac{\|\nabla p_{G_{[n]}, h}(x)\|^2}{p_{G_{[n]}, h}(x)} dx \\ &\leq C_{d, \bar{k}, \underline{k}} \frac{1}{nh} \sum_{i=1}^n \int \left(1 - \frac{f_{\tilde{G}_i, I_d}(x_i)}{f_{\tilde{G}_i, I_d}(x_i) \vee n(h\bar{k})^{d/2}\epsilon}\right)^2 \frac{\|\nabla f_{\tilde{G}_i, I_d}(x_i)\|^2}{f_{\tilde{G}_i, I_d}(x_i)} dx_i \\ &= C_{d, \bar{k}, \underline{k}} \frac{1}{nh} \sum_{i=1}^n \Delta(\tilde{G}_i, n(h\bar{k})^{d/2}\epsilon). \end{aligned}$$

where $C_{d,\underline{k},\bar{k}} = \bar{k}^{d/2}/\underline{k}^{d/2+1}$. Now, Lemma 6 directly applies to each summand. Provided $\epsilon \in (0, (2\pi h\bar{k})^{-d/2}e^{-1/2}\frac{1}{n}]$, and $L(\epsilon) := \sqrt{-\frac{1}{h\bar{k}}\log((2\pi h\bar{k})^d n^2 \epsilon^2)}$, then for compact sets S_i , $i = 1, \dots, n$,

$$\mathbb{E}_{p_{G_{[n],h}}}[\|s_{G_{[n],h}} - s_{G_{[n],h}}^\epsilon\|^2] \leq C_{d,\underline{k},\bar{k}} \frac{1}{nh} \sum_{i=1}^n N(4/L(\epsilon), S_i) L^d(\epsilon) \epsilon + dG_i(S_i^c).$$

where we applied the covering number bound and the change-of-variables identity for pushforward measures under linear maps to re-express the terms defined on S and S^c . $\square$

We are now finally ready to integrate the above pieces together to prove Lemma 2.

Proof of Lemma 2. Provided

$$\begin{aligned} q_n^{(t)} &= \frac{1}{n} \sum_{i=1}^n K_t(\cdot | i) = \frac{1}{n} \sum_{i=1}^n \delta_{(\alpha_i \Sigma_i)^{-1/2} X_i} * \mathcal{N}(0, (1 - \alpha_i^{-1}) I_d) \\ q_{G^*}^{(t)} &= \frac{1}{n} \sum_{i=1}^n f_{G_{i,t}, I_d}^* = \frac{1}{n} \sum_{i=1}^n f_{G_{i,t}, I_d}^* \Big|_{t=1} * \mathcal{N}(0, (1 - \alpha_i^{-1}) I_d) \end{aligned}$$

where each $G_{i,t}^* := (\Sigma_i^{(t)})^{-1/2} \# G^*$ and $\Sigma_i^{(t)} = \alpha_i(t) \Sigma_i = (\int_0^t g_i^2(s) ds) \Sigma_i$. It highlights that q_n and q_G can be viewed as a specialization of the mixtures in Definition 2. We can easily verify that $f_{G_{i,t}, I_d}^* \Big|_{t=1}$ is at most $\sqrt{4M^2/\underline{\sigma} + 1}$ -subgaussian and 1-log-smooth. By construction, $1 - \alpha_i^{-1} \in [\frac{1}{2}, 1)$.

By the earlier decomposition,

$$\mathbb{E}_{\mathbf{X}_n} \mathbb{E}_{q_n} [\|s_{G^*} - s_n\|^2] \leq 3\mathbb{E}_{q_*} [\|s_{G^*} - s_{G^*}^\epsilon\|^2] + 3\mathbb{E}_{\mathbf{X}_n} \mathbb{E}_{q_n|\mathbf{X}_n} [\|s_{G^*}^\epsilon - s_n^\epsilon\|^2 + \|s_n - s_n^\epsilon\|^2].$$

For the first term, the test function is deterministic with respect to the data, thus we can reduce the expectation to its population quantity. On the other hand, for terms involving s_n , it is a random function of the data samples, thus we evaluate first the inner expectation, conditioned on a fixed data.

The first term can be controlled via a direct application of Corollary 2, with $\epsilon = (2\pi)^{-d/2}e^{-1/2}/n$ and the sets $S_i = \{(\Sigma_i^{(t)})^{-1/2}\theta \mid \theta \in [-M, M]^d\}$ so $G_{i,t}^*(S_i) = 1$. Then we have:

$$\begin{aligned} \mathbb{E}_{q_*} [\|s_{G^*} - s_{G^*}^\epsilon\|^2] &\leq C_d [\sup_i N(4, S_i)] \frac{1}{n} \lesssim \left(1 + \frac{M\sqrt{d}}{2(\min_i \alpha_i(t)\underline{\sigma})^{1/2}}\right)^d \frac{1}{n} \\ &\leq C_{d,M,\underline{\sigma}} \frac{1}{n}. \end{aligned}$$

The last term can be controlled similarly. Standard Gaussian concentration yields that: provided $n \geq 2d$, with probability at least $1 - n^{-1}$, $\max \|X_i\|_\infty \leq r_n := M + \bar{\sigma}\sqrt{6\log(n)}$. Now let the compact set in Corollary 2 be $S_i := \{(\Sigma_i^{(t)})^{-1/2}x \mid x \in [-r_n, r_n]^d\}$. Conditioned on the high-probability event, each $\delta_{(\Sigma_i^{(t)})^{-1/2}X_i}$ is compactly supported on S_i and thus satisfies $G_i(S_i^c) = 0$; otherwise $G_i(S_i^c) \leq 1$ is uniformly bounded. Let $\epsilon = (2\pi)^{-d/2}e^{-1/2}/n$ and we obtain:

$$\mathbb{E}_{\mathbf{X}_n} \mathbb{E}_{q_n|\mathbf{X}_n} [\|s_n - s_n^\epsilon\|^2] \lesssim \left(\sup_{i \in [n]} N(4/L(\epsilon), S_i) L^d(\epsilon) + d\right) \frac{1}{n} \lesssim \frac{(\log n)^d}{n}.$$

Lastly for the Bayes discrepancy between regularized scores, by Lemma 4, Corollary 1, and the same strategy to reduce the expected Hellinger distance in Wibisono et al. [2024, Lemma 4],

$$\begin{aligned} \mathbb{E}_{\mathbf{X}_n} \mathbb{E}_{q_n|\mathbf{X}_n} [\|s_{G^*}^\epsilon - s_n^\epsilon\|^2] &\leq C_{d,\underline{\sigma}} \mathbb{E}_{\mathbf{X}_n} [\max \{(\log n)^3, |\log \mathfrak{H}(q_n, q_{G^*})|\} \mathfrak{H}^2(q_n, q_{G^*})] \\ &\leq C_{d,M,(\bar{\sigma},\underline{\sigma})} \frac{(\log n)^{d/2+3}}{n}. \end{aligned}$$

Combining the terms, we obtain the desired result that for all $t \in [t_0, T]$,

$$\mathbb{E}_{\mathbf{X}_n} \left[\mathbb{E}_{x \sim q_n^{(t)}} \left\| s_{G^*}^{(t)}(x) - s_n^{(t)}(x) \right\|_2^2 \right] \leq C_{d,M,(\bar{\sigma}, \underline{\sigma})} \frac{1}{n} (\log n)^{d+3}.$$

□

B.2.2 Control of $\mathcal{J}_*(\widehat{G}_n) - \widehat{\mathcal{J}}_*(\widehat{G}_n)$

We control the term $\mathcal{J}_*(\widehat{G}_n) - \widehat{\mathcal{J}}_*(\widehat{G}_n)$ in (17) via empirical process theory together with the regularization trick introduced above. In particular, we crucially rely on local Rademacher complexity analysis following Bartlett et al. [2005], Lei et al. [2016]. Localization also appears essential to obtain sharp learning rate in empirical Bayes guarantees [Saha and Guntuboyina, 2020, Soloff et al., 2025, Stein, 1981]. Under the uniform bounding assumption, it is without loss of generality to let $g_i = g$ for all $i \in [n]$ to ease the notations.

Denote the pointwise loss ψ with:

$$\psi_{i,t}(x; G) := \left\| s_G^{(t)}(x) - s_{G^*}^{(t)}(x) \right\|_{\Sigma_i}^2,$$

and let the regularized point-wise discrepancy ψ^ϵ , where we write the shorthand notation $s_{G,t}^\epsilon := \nabla q_G^{(t)} / (q_G^{(t)} \vee \epsilon)$:

$$\psi_{i,t}^\epsilon(x; G) := \|s_{G,t}^\epsilon(x) - s_{G^*,t}^\epsilon(x)\|_{\Sigma_i}^2.$$

We also let:

$$h_{i,t}(x; G) = \mathbb{E}_{Z \sim \mathcal{N}(0, (1-\alpha_i(t)^{-1})I_d)} \psi_{i,t}(x + Z; G), \quad h_{i,t}^\epsilon(x; G) = \mathbb{E}_{Z \sim \mathcal{N}(0, (1-\alpha_i(t)^{-1})I_d)} \psi_{i,t}^\epsilon(x + Z; G).$$

such that for all fixed G , denote $\widetilde{X}_i := (\Sigma_i^{(t)})^{-1/2} X_i$

$$\mathbb{E}[h_{i,t}(\widetilde{X}_i; G)] = \mathbb{E}_{X \sim f_{G^*,t}, I_d} [\psi_{i,t}(X; G)],$$

$$\widehat{\mathcal{J}}_*(G) = \int g(t)^2 \frac{1}{n} \sum_i h_{i,t}(\widetilde{X}_i; G) dt, \quad \mathcal{J}_*(G) = \mathbb{E} \widehat{\mathcal{J}}_*(G).$$

Consistent with Bartlett et al. [2005], we write for a function $f : \mathcal{X} \rightarrow \mathbb{R}$,

$$P_n f = \frac{1}{n} \sum_{i=1}^n f(X_i), \quad P f = \mathbb{E} f(X), \quad R_n f = \frac{1}{n} \sum_{i=1}^n \sigma_i f(X_i).$$

Here $\sigma_1, \dots, \sigma_n$ are n independent Rademacher variables. For a function class $\mathcal{F}$, its Rademacher complexity is $\mathbb{E} R_n \mathcal{F} := \mathbb{E} [\sup_{f \in \mathcal{F}} R_n f]$.

To facilitate metric entropy toolkit derived in Saha and Guntuboyina [2020], Soloff et al. [2025], we consider the following pseudometric that restricts the localized score discrepancy to a compact set S :

$$m_\epsilon^S(G, H) := \max_{i \in [n]} \sup_{x \in S} \|s_{G,t}^\epsilon(x) - s_{H,t}^\epsilon(x)\|_{\Sigma_i}.$$

It is natural to consider the following compact set: $A := \{x \in \mathbb{R}^d : q_G^{(t)}(x) \geq \epsilon, \forall G \in \mathcal{P}([-M, M]^d)\}$, in which we guarantee $s_G = s_G^\epsilon$ for all measures G . Denote

$$(\underline{\sigma}_t, \bar{\sigma}_t) := \left(\min_{i \in [n]} \sigma_{\min}(\Sigma_i^{(t)}), \max_{i \in [n]} \sigma_{\max}(\Sigma_i^{(t)}) \right).$$

It is straightforward to show that a sufficient radial condition that implies $x \in A$ is:

$$\|x\|_2 \leq r_\epsilon \leq \left[\sqrt{2 \log((2\pi))^{-d/2} \epsilon^{-1}} - M \sqrt{d/\underline{\sigma}_t} \right]_+.$$

Let $B = \{x \in \mathbb{R}^d : \|x\|_2 \leq r_\epsilon\}$ then $B \subseteq A$. By standard Gaussian concentration, provided

$$\epsilon \leq \epsilon_n := (2\pi)^{-d/2} \exp \left(-\frac{1}{2} (2M \sqrt{d/\underline{\sigma}_t} + \sqrt{d} + 2\sqrt{\log n})^2 \right),$$

then

$$\mathbb{P}(X \in A) \geq \mathbb{P}(X \in B) \geq 1 - n^{-2}, \forall X \sim f_{G_{i,t}, I_d}, \forall i \in [n]. \quad (21)$$

With the above construction, we define the discrepancy function $\tilde{h}_{G,t} : [n] \times \mathbb{R}^d \rightarrow \mathbb{R}$ that restricts the regularized discrepancy to the high-probability set A :

$$\tilde{h}_{G,t}(i, x) = \mathbb{E}_{Z \sim \mathcal{N}(0, (1 - \alpha_i(t)^{-1})I_d)} [\psi_{i,t}^\epsilon(x + Z; G) \mathbf{1}\{x + Z \in A\}].$$

Then for arbitrary fixed G , we separate:

$$\begin{aligned} \mathcal{J}_*(G) - \widehat{\mathcal{J}}_*(G) &= \int_{t_0}^T g(t)^2 [P \tilde{h}_{G,t} - P_n \tilde{h}_{G,t}] dt \\ &\quad + \int_{t_0}^T g(t)^2 \frac{1}{n} \sum_{i=1}^n \left\{ \mathbb{E}[\psi_{i,t} \mathbf{1}\{A^c\}] - \mathbb{E}_{Z_i^{(t)}} \psi_{i,t}(\tilde{X}_i + Z_i^{(t)}) \mathbf{1}\{\tilde{X}_i + Z_i^{(t)} \in A^c\} \right\} dt \\ &\leq \int_{t_0}^T g(t)^2 [P \tilde{h}_{G,t} - P_n \tilde{h}_{G,t}] dt + \int_{t_0}^T g(t)^2 \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\psi_{i,t} \mathbf{1}\{A^c\}] dt, \end{aligned}$$

where P and P_n are defined with respect to the product measures $\frac{1}{n} \sum_{i=1}^n \delta_i \otimes f_{G_{i,t}, I_d} \Big|_{t=1}$ and $\frac{1}{n} \sum_{i=1}^n \delta_{(i, \tilde{X}_i)}$ defined over the augmented space $[n] \times \mathbb{R}^d$. We drop the last term since ψ 's are nonnegative.

Control of $[P - P_n] \tilde{h}$. We apply the following results that establishes generalization analysis with local Rademacher complexities:

Lemma 7 (Theorem 3.3 of Bartlett et al. [2005]). *Let $\mathcal{F}$ be a class of functions with range in $[a, b]$ and assume that there are some functional $T : \mathcal{F} \rightarrow \mathbb{R}_+$ and some constant B such that for every $f \in \mathcal{F}$, $\text{Var}[f] \leq T[f] \leq B \cdot Pf$. Let φ be a sub-root function that satisfies the following properties:*

- $\varphi : [0, \infty) \rightarrow [0, \infty)$ is non-decreasing and $r \rightarrow \varphi(r)/\sqrt{r}$ is non-increasing for $r > 0$.
- φ is continuous on $[0, \infty)$ and the equation $\varphi(r) = r$ has a unique positive solution r^* . For all $r > 0$, $r \geq r^* \iff r \geq \varphi(r)$.

If for any $r \geq r^*$, φ satisfies,

$$\varphi(r) \geq B \mathbb{E} R_n \{f \in \mathcal{F} : T(f) \leq r\},$$

then for any $K > 1$, and any $t > 0$, with probability at least $1 - \exp(-t)$,

$$Pf \leq \frac{K}{K-1} P_n f + c_1 \frac{K}{B} r^* + (c_2(b-a) + c_3 BK) \frac{t}{n}, \quad \forall f \in \mathcal{F}$$

where c_0, c_1, c_2 are numerical constants.

It is convenient to then apply Lei et al. [2016]'s results that presents the local Rademacher complexity bounds in terms of metric entropy.

Lemma 8 (Corollary 1 of Lei et al. [2016]). *Let $\mathcal{F}$ be a function class with $\sup_{f \in \mathcal{F}} \|f\|_\infty \leq b$. If $\mathcal{F}$ can be finitely covered with respect to the metric $\|\cdot\|_{L_2(P_n)} := (\int |f|^2 d\mu)^{1/2}$; that is there exists positive numbers γ, δ, p such that $\log N(u, \mathcal{F}, \|\cdot\|_2) := \sup_n \sup_{P_n} \log N(u, \mathcal{F}, \|\cdot\|_{L_2(P_n)}) \leq \delta \log^p(\gamma/u)$ for any $u \in (0, \gamma]$, then for any $r \in (0, \gamma^2]$,*

$$\mathbb{E} R_n \{f \in \mathcal{F} : Pf^2 \leq r\} \leq c_{b,p,\gamma} \left[\sqrt{\frac{\delta r \log^p(2\gamma n^{1/2})}{n}} + \frac{\delta \log^p(2\gamma n^{1/2})}{n} \right]. \quad (22)$$

In particular, the right hand side of (22) defines a valid sub-root function whose fixed point satisfies $r^* \leq c\delta \log(n)^p/n$ (see Section B.3. of Lei et al. [2016]). Thus, it suffices for us to show the function class of interests satisfies the required covering number bounds.

For every fixed t , the (relaxed) function class of interests is $\mathcal{H}_t = \{\tilde{h}_{G,t} : G \in \mathcal{P}(\mathbb{R}^d)\}$. We have the following:

- For all $h \in \mathcal{H}_t$, there exists $b_t(\epsilon)$ such that $0 \leq h \leq \psi^\epsilon \leq 4 \sup_G \|s_G^\epsilon\|^2 \leq b_t^2(\epsilon)$. By Lemma 3, let

$$b_t(\epsilon) = 2\sqrt{\log[(2\pi)^{-d}\epsilon^{-2}]}$$

- The variance condition is satisfied by letting $T(h) := Ph^2$. Then $\text{Var}[h] \leq T[h]$ by definition and $T(h) \leq b_t^2(\epsilon)Ph$ by the uniform bounding above.
- The covering number of $\mathcal{H}_t$ satisfies $N(u, \mathcal{H}_t, \|\cdot\|_2) \leq N(\frac{u}{2b_t(\epsilon)}, \mathcal{P}(\mathbb{R}^d), m_\epsilon^A)$. For $G, H \in \mathcal{P}(\mathbb{R}^d)$, and $(i, x) \in P_n$,

$$|\tilde{h}_{G,t} - \tilde{h}_{H,t}|(i, x) \leq 2b_t(\epsilon)\mathbb{E}_Z[|\psi_{i,t}^\epsilon(x+Z; G)^{1/2} - \psi_{i,t}^\epsilon(x+Z; H)^{1/2}| \cdot \mathbf{1}\{x+Z \in A\}].$$

By the reverse triangle inequality that $||a - c| - |b - c|| \leq |a - b|$ and conditioned on $y = x + Z \in A$:

$$\begin{aligned} & |\psi_{i,t}^\epsilon(y; G)^{1/2} - \psi_{i,t}^\epsilon(y; H)^{1/2}| \\ &= ||s_{G,t}^\epsilon(y) - s_{G^*,t}^\epsilon(y)|_{\Sigma_i} - |s_{H,t}^\epsilon(y) - s_{G^*,t}^\epsilon(y)|_{\Sigma_i}| \\ &\leq |s_{G,t}^\epsilon(y) - s_{H,t}^\epsilon(y)|_{\Sigma_i} \leq m_\epsilon^A(G, H). \end{aligned}$$

Thus for every n and P_n ,

$$\begin{aligned} & \|\tilde{h}_{G,t} - \tilde{h}_{H,t}\|_{L_2(P_n)} \leq 2b_t(\epsilon)m_\epsilon^A(G, H) \\ \implies & \log N(u, \mathcal{H}_t, \|\cdot\|_2) \leq \log N(\frac{u}{2b_t(\epsilon)}, \mathcal{P}(\mathbb{R}^d), m_\epsilon^A). \end{aligned}$$

By the same strategy in Soloff et al. [2025], we have:

$$\begin{aligned} m_\epsilon^A(G, H) &\leq \bar{\sigma}_t^{1/2}(2\epsilon)^{-1}b_t(\epsilon) \sup_{x \in A} |q_G^{(t)}(x) - q_H^{(t)}(x)| \\ &\quad + \bar{\sigma}_t^{1/2}\epsilon^{-1} \sup_{x \in A} \|\nabla q_G^{(t)}(x) - \nabla q_H^{(t)}(x)\|_2 \\ &\leq \bar{\sigma}_t^{1/2}(2\epsilon)^{-1}b_t(\epsilon) \max_{i \in [n]} \sup_{x \in A} |f_{G_{i,t}, I_d}(x) - f_{H_{i,t}, I_d}(x)| \\ &\quad + \bar{\sigma}_t^{1/2}\epsilon^{-1} \max_{i \in [n]} \sup_{x \in A} \|\nabla f_{G_{i,t}, I_d}(x) - \nabla f_{H_{i,t}, I_d}(x)\|_2. \end{aligned}$$

Lemma 9 (Lemmas 5 and 6 of Soloff et al. [2025]). *Define the semi-norm over $S \subset \mathbb{R}^d$:*

$$\begin{aligned} \|f_{G,\circ} - f_{H,\circ}\|_{\infty,S} &:= \max_{i \in [n]} \sup_{x \in S} |f_{G_{i,t}, I_d}(x) - f_{H_{i,t}, I_d}(x)|, \\ \|f_{G,\circ} - f_{H,\circ}\|_{\nabla,S} &:= \max_{i \in [n]} \sup_{x \in S} \|\nabla f_{G_{i,t}, I_d}(x) - \nabla f_{H_{i,t}, I_d}(x)\|. \end{aligned}$$

and let $S^a := \{x \in \mathbb{R}^d : \inf_{y \in S} \|x - y\| \leq a\}$. For $\mathbb{F} := \{(f_{G_i, K_i})_{i=1}^n : \{G_i\}_{i=1}^n \in \mathcal{P}(\mathbb{R}^d), \underline{k}I_d \preceq \{K_i\}_{i=1}^n \preceq \bar{k}I_d\}$, there exists positive constants C_d and $c_{d, \bar{k}, \underline{k}}$ such that for all compact $S \subset \mathbb{R}^d$, $M > 0$, and sufficiently small $\eta > 0$ such that

$$\begin{aligned} \log N(\eta, \mathbb{F}, \|\cdot\|_{\infty,S}) &\leq C_d N(a, S^a) \left(\log \frac{c_{d, \bar{k}, \underline{k}}}{\eta} \right)^{d+1}, \\ \log N(\eta, \mathbb{F}, \|\cdot\|_{\nabla,S}) &\leq C_d N(a, S^a) \left(\log \frac{c_{d, \bar{k}, \underline{k}}}{\eta} \right)^{d+1}. \end{aligned}$$

where $a = \sqrt{-2\bar{k} \log \frac{\sqrt{\bar{k}\wedge 1}}{5} (2\pi \underline{k})^{d/2} \eta}$.

Apply Lemma 9 to our metric, by letting

$$u^* = \bar{\sigma}_t^{1/2}(b_t(\epsilon)/2 + 1)\frac{\eta}{\epsilon},$$

provided u^* sufficiently small, we have

$$\begin{aligned}\log N(u^*, \mathcal{P}(\mathbb{R}^d), m_\epsilon^A) &\leq \log N(\eta/2, \mathbb{F}, \|\cdot\|_{\infty, A}) + \log N(\eta/2, \mathbb{F}, \|\cdot\|_{\nabla, A}) \\ &\leq C_d N(a, A^a) (\log \frac{C_d}{\eta})^{d+1} \\ &\leq C_d a^{-d} \text{Vol}(A^{3a/2}) (\log \frac{C_d(b_t(\epsilon)/2 + 1)}{\epsilon u^*})^{d+1}.\end{aligned}$$

where the last step is due to Lemma F.6 of Saha and Guntuboyina [2020]. Since a necessary radial condition for $x \in A$ is given by

$$d(x, [-\frac{M}{\sqrt{\sigma_t}}, \frac{M}{\sqrt{\sigma_t}}]^d)^2 = \inf_{y \in [-\frac{M}{\sqrt{\sigma_t}}, \frac{M}{\sqrt{\sigma_t}}]^d} \|x - y\|_2^2 \leq 2 \log((2\pi))^{-d/2} \epsilon^{-1} =: \bar{r}_t^2(\epsilon),$$

we have that:

$$\begin{aligned}a^{-d} \text{Vol}(A^{3a/2}) &\leq a^{-d} \cdot C_d \left(\bar{r}_t(\epsilon) + M \sqrt{d/\sigma_t} + \frac{3}{2}a \right)^d \\ &\lesssim_{d, M, \sigma_t} (\bar{r}_t(\epsilon) \vee 1)^d \log \left(\frac{5(2\pi)^{-d/2}}{\eta} \right)^{d/2}.\end{aligned}$$

This implies that:

$$\log N \left(\frac{u^*}{2b_t(\epsilon)}, \mathcal{P}(\mathbb{R}^d), m_\epsilon^A \right) \lesssim_{d, M, \sigma_t} (\bar{r}_t(\epsilon) \vee 1)^d \left(\log \frac{C'(b_t(\epsilon)/2 + 1)b_t(\epsilon)}{\epsilon u^*} \right)^{3d/2+1}.$$

By choosing $\epsilon \equiv \epsilon_n = o(n^{-2})$, we have $b_t(\epsilon_n) \asymp \sqrt{\log n}$ and $\bar{r}_t^2(\epsilon_n) \asymp \log n$, provided $n \geq N_{d, M, \sigma_t, \sigma_t}$ sufficiently large such that terms involving n dominate constant terms. Thus, one can apply Lemma 7 (with $K = 2$) and Lemma 8 (with $\gamma \asymp (b_t(\epsilon)/2 + 1)b_t(\epsilon)/\epsilon$, $\delta \asymp (\bar{r}_t(\epsilon) \vee 1)^d$, $p = 3d/2 + 1$) to show that for arbitrary $k \geq 1$ and $G \in \mathcal{P}([-M, M]^d)$, with probability at least $1 - n^{-k}$,

$$\begin{aligned}P\tilde{h}_{G,t} - P_n\tilde{h}_{G,t} &\leq P_n\tilde{h}_{G,t} + c_1 \frac{2(\bar{r}_t(\epsilon) \vee 1)^d}{b_t(\epsilon)^2} \cdot \frac{\log(n)^{3d/2+1}}{n} + c_2 b_t(\epsilon) \frac{k \log(n)}{n} \\ \implies \int_{t_0}^T g(t)^2 [P\tilde{h}_{\hat{G}_n,t} - P_n\tilde{h}_{\hat{G}_n,t}] dt &\leq \hat{\mathcal{J}}_*(\hat{G}_n) + O\left(\frac{(\log n)^{2d}}{n}\right),\end{aligned}$$

since $\tilde{h}_{G,t}(i, x) \leq h_{i,t}(x; G)$ and the time integral of $P_n h_{i,t}(x; G)$ recovers $\hat{\mathcal{J}}_*(G)$ by definition.

Control of tail. We remain to bound for every t and i :

$$\begin{aligned}\mathbb{E}_{x \sim G_{i,t}^* \star \mathcal{N}(0, I_d)}[\psi_{i,t}(x; G) \mathbf{1}\{x \in A^c\}] \\ = \mathbb{E}[(\psi_{i,t}(x; G) - \psi_{i,t}^\epsilon(x; G)) \mathbf{1}\{x \in A^c\}] + \mathbb{E}[\psi_{i,t}^\epsilon(x; G) \mathbf{1}\{x \in A^c\}].\end{aligned}$$

The second term is trivially bounded via uniform bounding $b_t(\epsilon)$ of every regularized score functional ψ^ϵ and eq.(21) that $\mathbb{P}(A^c) \leq n^{-2}$ provided $\epsilon \equiv \epsilon_n$ chosen above. Thus,

$$\mathbb{E}[\psi_{i,t}^\epsilon(x; G) \mathbf{1}\{x \in A^c\}] \leq b_t^2(\epsilon) n^{-2}. \quad (\text{i})$$

By standard algebraic operations via triangle inequality, let $\delta_1 = \max_{c \in \{a, b\}} \|c - c^\epsilon\|$ and $\delta_2 = \max_{c \in \{a, b\}} \|c^\epsilon\|$,

$$\begin{aligned}\|a - b\| &\leq \|a^\epsilon - b^\epsilon\| + \|a - a^\epsilon\| + \|b - b^\epsilon\| = \|a^\epsilon - b^\epsilon\| + 2\delta_1 \\ \implies \| \|a - b\|^2 - \|a^\epsilon - b^\epsilon\|^2 \| &\leq \| \|a - b\| - \|a^\epsilon - b^\epsilon\| \| \cdot (\|a - b\| + \|a^\epsilon - b^\epsilon\|) \\ &\leq 2\delta_1 \cdot (2\|a^\epsilon - b^\epsilon\| + 2\delta_1) \leq 4(\delta_1^2 + 2\delta_1\delta_2).\end{aligned}$$

Letting $a = s_G^{(t)}$, $b = s_{G^*}^{(t)}$, $a^\epsilon = s_{G,t}^\epsilon$, $b^\epsilon = s_{G^*,t}^\epsilon$, and applying Cauchy-Schwarz inequality, the first term in above becomes:

$$\begin{aligned}\mathbb{E}[(\psi_{i,t} - \psi_{i,t}^\epsilon) \mathbf{1}\{A^c\}] &= \mathbb{E}[(\|s_G - s_{G^*}\|_{\Sigma_i^{(t)}}^2 - \|s_{G,t}^\epsilon - s_{G^*,t}^\epsilon\|_{\Sigma_i^{(t)}}^2) \mathbf{1}\{A^c\}] \\ &\leq \max_{H, H' \in \mathcal{P}([-M, M]^d)} 4\bar{\sigma}_t \mathbb{E}[(\|s_H - s_H^\epsilon\|^2 + 2\|s_H - s_H^\epsilon\| \cdot \|s_{H'}^\epsilon\|) \mathbf{1}\{A^c\}] \\ &\leq \max_{H \in \mathcal{P}([-M, M]^d)} 4\bar{\sigma}_t \mathbb{E}[(\|s_H - s_H^\epsilon\|^2) + b_t(\epsilon)(\mathbb{E}[\|s_H - s_H^\epsilon\|^2])^{1/2} \mathbb{P}(A^c)^{1/2}] \\ &\leq 4\bar{\sigma}_t \left(\Delta + \frac{b_t(\epsilon)}{n} \Delta^{1/2} \right).\end{aligned} \quad (\text{ii})$$

where we again used the uniform bounding. The discrepancy $\Delta := \mathbb{E}[\|s_H - s_H^\epsilon\|^2]$ can be bounded by Corollary 2 with a change of measure. By the Cauchy-Schwarz inequality,

$$\Delta \leq \sqrt{1 + \chi^2(q_H \| q_{G^*})} (\mathbb{E}_{q_H} \|s_H - s_H^\epsilon\|^4)^{1/2},$$

where the fourth moment of the score discrepancy can be analogously controlled as the second moment with only the change of the leading constant and the dependency from $L^d(\epsilon)$ to $L^{d+2}(\epsilon)$. The chi-squared distance $\chi^2(p \| q) = \int \frac{p^2}{q} - 1$ is jointly convex and we have $\chi^2(f_{G_{i,t}, I_d} \| f_{H_{i,t}, I_d}) \leq C_{d, M, \underline{\sigma}, \bar{\sigma}}$ is upper bounded by a constant. Consequently, given G^*, H compactly supported, with the choice of $\epsilon \lesssim n^{-2}$, we have that:

$$\Delta \lesssim \frac{(\log n)^{(d+2)/4}}{n}.$$

Recall $b_t(\epsilon) \asymp \sqrt{\log n}$, so combining (i) + (ii) yields:

$$\int_{t_0}^T g(t)^2 \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\psi_{i,t} \mathbf{1}\{A^c\}] dt \lesssim \frac{(\log n)^{(d+2)/4}}{n}.$$

Combining the high-probability part and the tail, we can conclude this section by asserting:

$$\mathcal{J}_*(\hat{G}_n) - \hat{\mathcal{J}}_*(\hat{G}_n) \lesssim_{d, M, (\bar{\sigma}, \underline{\sigma}), (\underline{g}, \bar{g})} \hat{\mathcal{J}}_*(\hat{G}_n) + \frac{\text{polylog}(n)}{n}.$$

By further upper bounding $\hat{\mathcal{J}}_*(\hat{G}_n)$ by $\mathcal{J}_n(G^*)$ from the basic inequality and applying Lemma 2, we recover the first claim in Theorem 3 on score estimation guarantee that with probability at least $1 - n^{-k}$ for arbitrary $k \geq 1$,

$$\mathbb{E} \mathcal{J}_*(\hat{G}_n) \lesssim 8 \mathbb{E} \mathcal{J}_n(G^*) + \frac{(\log n)^{2d+1/2}}{n} + \frac{\log(n)^{O(1)}}{n} \lesssim \frac{1}{n} (\log n)^{O(d)}.$$

B.3 Control of Hellinger divergence

Our next result controls the KL divergence between $q_G^{(t_0)}$ and $q_{G^*}^{(t_0)}$ by the Fisher risk controlled above. Without loss of generality, assume $g_i(t) = g(t)$ so $\alpha_i(t) = \alpha(t)$. Recall that we define:

$$q_G^{(t)} = \frac{1}{n} \sum_{i=1}^n (\Sigma_i^{(t)})^{-1/2} \#G * \mathcal{N}(0, I_d).$$

Let

$$\begin{aligned} \tilde{q}_G^{(t)} &= \frac{1}{n} \sum_{i=1}^n (\Sigma_i)^{-1/2} \#G * \mathcal{N}(0, \alpha(t) I_d) \\ \implies \nabla \log q_G^{(t)}(x) &= \frac{1}{\sqrt{\alpha(t)}} \nabla \log \tilde{q}_G^{(t)}(x / \sqrt{\alpha(t)}). \end{aligned}$$

Lemma 10. *For every fixed $G \in \mathcal{P}(\mathbb{R}^d)$,*

$$\text{KL}(q_{G^*}^{(t_0)} \| q_G^{(t_0)}) \lesssim \tilde{\mathfrak{F}}_{[t_0, T]}(q_{G^*}^{(t_0)} \| q_G^{(t_0)}).$$

Proof. Denote the path space by $\Omega = \mathcal{C}([t_0, T]; \mathbb{R}^d)$. Under the latent measure G and the same conditions on Σ_i and g_i as in Assumption 1, define the path law Q_G to be the unique solution for the following reverse-time SDE on $[t_0, T]$:

$$dX^{(t)} = -g(t)^2 \nabla \log \tilde{q}_G^{(t)}(X^{(t)}) dt + g(t) dB_t, \quad X^{(t_0)} \sim \tilde{q}_G^{(t_0)}. \quad (23)$$

Similarly to Song et al. [2021b, Proof of Theorem 1], consider the measurable Markov kernel $K : \Omega \times \mathcal{B}(\mathbb{R}^d) \rightarrow [0, 1]$ with $K(\omega, x) = \delta_{\omega(t_0)=x}$, then we have:

$$\tilde{q}_G^{(t_0)} = Q_G K := \int K(\omega, \cdot) Q_G(d\omega)$$

that shows that the marginal mixture law $\tilde{q}_G$ at t_0 is the push-forward of the aggregated path measure Q_G under the kernel K . By the data processing inequality,

$$\text{KL}(\tilde{q}_{G^*}^{(t_0)} || \tilde{q}_G^{(t_0)}) = \text{KL}(Q_{G^*} K || Q_G K) \leq \text{KL}(Q_{G^*} || Q_G).$$

In addition, the KL divergence is invariant with the linear rescaling:

$$\text{KL}(q_{G^*}^{(t_0)} || q_G^{(t_0)}) = \text{KL}(\tilde{q}_{G^*}^{(t_0)} || \tilde{q}_G^{(t_0)}).$$

Since the SDEs associated with Q_G and Q_{G^*} differ only in the drift coefficient, we can apply the Girsanov theorem [Oksendal, 2013] to yield:

$$\begin{aligned} \text{KL}(\tilde{Q}_{G^*} || \tilde{Q}_G) &= \frac{1}{2} \mathbb{E}_{Q_{G^*}} \left[\int_{t_0}^T g(t)^2 \left\| \nabla \log \tilde{q}_{G^*}^{(t)} - \nabla \log \tilde{q}_G^{(t)} \right\|_2^2 dt \right] \\ &= \frac{1}{2} \int_{t_0}^T \frac{g(t)^2}{\alpha(t)} \mathbb{E}_{x \sim q_{G^*}^{(t)}} \left[\left\| \nabla \log q_{G^*}^{(t)} - \nabla \log q_G^{(t)} \right\|_2^2 \right] dt \\ &\lesssim \frac{1}{2n} \sum_{i=1}^n \int_{t_0}^T \mathbb{E}_{x \sim f_{G^*, t, I_d}} \left[\left\| \nabla \log q_{G^*}^{(t)} - \nabla \log q_G^{(t)} \right\|_2^2 \right] dt \\ &\lesssim \frac{1}{2} \bar{\mathfrak{F}}_{[t_0, T]}(q_{G^*}^{(t)} || q_G^{(t)}). \end{aligned}$$

□

Plugging in the first claim that bounds $\bar{\mathfrak{F}}_{[t_0, T]}(q_{G^*}^{(t)} || q_G^{(t)})$ in Lemma 10 and use the basic divergence inequality that $\mathfrak{H}^2 \leq KL$ yields the desired second claim.

Lastly, the third claim on the Wasserstein divergence:

$$\mathbb{E} W_2^2 \left(\frac{1}{n} \sum_{i=1}^n (\Sigma_i)^{-1/2} \# G^*, \frac{1}{n} \sum_{i=1}^n (\Sigma_i)^{-1/2} \# \hat{G}_n \right) \lesssim \frac{1}{\log n}$$

follows from a direct application of Theorem 10 of Soloff et al. [2025] with the upper bound of $\mathfrak{H}^2(\tilde{q}_G^{(t_0)}, \tilde{q}_{\hat{G}_n}^{(t_0)})$ from above.

B.4 Reduction with Homoscedastic Noise

Proof of Theorem 1. When $\Sigma_i = \Sigma$ are all equal, We have:

$$\begin{aligned} q_G^t &= f_{(\Sigma^{(t)})^{-1/2} \# G, I_d} \implies f_{G, \Sigma^{(t)}}(x) = |\Sigma^{(t)}|^{-1/2} q_G^t \left((\Sigma^{(t)})^{-1/2} x \right), \\ \nabla \log f_{G, \Sigma^{(t)}}(x) &= (\Sigma^{(t)})^{-1/2} \nabla \log q_G^t \left((\Sigma^{(t)})^{-1/2} x \right). \end{aligned}$$

We have the following equivalence:

$$\begin{aligned} \mathfrak{F}(f_{G, \Sigma^{(t)}} || f_{H, \Sigma^{(t)}}) &\lesssim \lambda_{\min}(\Sigma^{(t)})^{-1} \mathfrak{F}(q_G^t || q_H^t), \\ \text{KL}(f_{G, \Sigma^{(t)}} || f_{G, \Sigma^{(t)}}) &= \text{KL}(q_G^t || q_H^t), \\ W_2^2(G, H) &\leq \sqrt{\lambda_{\max}(\Sigma)} W_2^2(\Sigma^{-1/2} \# G, \Sigma^{-1/2} \# H). \end{aligned}$$

This shows that Theorem 1 is a direct corollary of Theorem 3 with only adjustments to the leading constants. □

C Implementation Details and Additional Results

Experiments compute resources. We fit all our models and performed all experiments on an internal HPC cluster. Jobs were scheduled on GPU-enabled nodes, each using one NVIDIA GeForce GTX 2080 Ti GPU and its accompanying Intel Xeon Skylake 6130 @ 2.1 GHz CPUs with 96 GB of RAM per node. Individual training runs require a maximal of 2 GPU hours and 4 CPU hours, while

the denoising runs require a maximal of 1 GPU hour and 2 CPU hours. The entire experiment takes approximately 120 hours. These estimates do not include preliminary or failed experiments.

Hyperparameters and architectures. For our implementations, we consider two variants SDGM-C and SDGM-D, for continuous and discrete latent model, respectively. In the continuous case, the flow uses four repeats of the following block: ActNorm $\rightarrow$ random permutation $\rightarrow$ MaskedPiecewiseRationalQuadraticAutoregressiveTransform, where each autoregressive transform conditions on all previous dimensions via a 2-layer MADE network (128 hidden features), uses eight linear-tail bins (tail bound 10), enforces minimum bin width/height and derivative of 5×10^{-3} , leaky ReLU activations, residual blocks, and no dropout or batch norm. The base distribution is a standard normal distribution. In the discrete case, we use a trainable mixture of 2,000 components (initialized from random normals or random draws from the observations) and corresponding randomly initialized weights.

For the baselines, used the official implementation of Dual-ALM provided by the authors of Zhang et al. [2022] at <https://github.com/YangjingZhang/Dual-ALM-for-NPMLE>. For NPMLE-N, we use all available training data samples as the fixed supports. Thus, the latent mixture is given by: $\sum_{j=1}^n w_j \delta_{X_j}$ with w_j 's trainable. For PEM, we fix a latent mixture of 2,000 components, i.e., $\hat{G} = \sum_{j=1}^{2000} w_j \delta_{\mu_j}$ with both $\{(w_j, \mu_j)\}_{j=1}^{2000}$ optimizable. We randomly sample points from the dataset as the initial supports and uniformly initialize the weights.

SDE setting and training objective. Throughout our experiments, we consider the following family of variance-exploding SDEs:

$$dY_i^{(t)} = g(t)\Sigma_i^{1/2}dB_t, \quad \text{with } Y_i^{(0)} \sim G^*.$$

In particular, we use a shared diffusion coefficient $g(t) = \sqrt{\frac{2}{e^{2t}-1}}e^t$. Under this SDE, the marginal distribution at any time $t \geq 0$ is $Y_i^{(t)} \sim G^* * \mathcal{N}(0, \Sigma_i^{(t)})$, where the effective time-dependent covariance is $\Sigma_i^{(t)} = \frac{e^{2t}-1}{e^2-1}\Sigma_i$. Our observed data X_i correspond to $Y_i^{(1)}$.

The training objective for our SDGM models is based on DSM to match the conditional score of $f_{\tilde{G}, \Sigma_i^{(t)}}$ with $p_{1t}(\cdot|X_i)$. For each data point (X_i, Σ_i) in a training batch of size N_B :

1. A time $t_i > 1$ is sampled (specifically, t_i is drawn uniformly from $[1 + \epsilon, T_{\max}]$, where we use $\epsilon = 10^{-3}$ and $T_{\max} = 1.5$ in our experiments).
2. A perturbed data point, X'_i , is generated by corrupting X_i with additional noise: $X'_i = X_i + Z_i$, where $Z_i \sim \mathcal{N}(0, \Sigma_i^{(t_i)} - \Sigma_i^{(1)})$. This X'_i is a sample from the SDE path at time t_i , conditioned on X_i being the state at $t = 1$.

To compute the score of our model's current estimate of the data distribution at time t_i , we first represent our learned latent prior, $\tilde{G}$.

- For SDGM-C, $\tilde{G}$ is represented by a neural spline flow. We obtain a discrete approximation by drawing $m = 2,000$ samples $\{\mu_j\}_{j=1}^m$ from this flow and assigning uniform weights $w_j = 1/m$.
- For SDGM-D, $\tilde{G} = \sum_{j=1}^m w_j \delta_{\mu_j}$, where the support points $\{\mu_j\}_{j=1}^m$ and their corresponding logits (which determine weights w_j via softmax) are directly trainable parameters, with $m = 2,000$.

The score of the marginal distribution conditional on i -th SDE at time t_i implied by $\tilde{G}$ and noise covariance $\Sigma_i^{(t_i)}$ is denoted $\nabla \log f_{\tilde{G}, \Sigma_i^{(t_i)}}(X'_i)$. This score is computed as $\sum_{j=1}^m p(\mu_j|X'_i, \Sigma_i^{(t_i)})(\Sigma_i^{(t_i)})^{-1}(\mu_j - X'_i)$, where $f_{\tilde{G}, \Sigma_i^{(t_i)}}(x) = \sum_{j=1}^m w_j \mathcal{N}(x; \mu_j, \Sigma_i^{(t_i)})$ and $p(\mu_j|X'_i, \Sigma_i^{(t_i)})$ are the posterior probabilities of the mixture components given the perturbed data X'_i . The objective function minimizes the expected squared difference between this model score and the true score of the perturbation kernel, $\nabla_{X'_i} \log p_{1t_i}(X'_i|X_i) = -(\Sigma_i^{(t_i)} - \Sigma_i^{(1)})^{-1}Z_i$. For a batch,

the loss is:

$$\mathcal{L}(\tilde{G}) = \frac{1}{N_B} \sum_{i=1}^{N_B} g^2(t_i) \left\| \nabla \log f_{\tilde{G}, \Sigma_i^{(t_i)}}(X'_i) - \nabla \log p_{1t_i}(X'_i | X_i) \right\|_{\Sigma_i}^2$$

All our experiments are performed with the importance sampling strategy as in Song et al. [2021b] that further reduce the variance in estimating the time integral by directly sample t_i according to $p(t) \propto g(t)^2$. The optimization is performed using the Adam optimizer. For the experiments presented, we used a learning rate of 5×10^{-5} (unless specified otherwise for particular experiments), $\beta_1 = 0.9$, Adam $\epsilon = 10^{-8}$, no weight decay (weight decay = 0.0), and gradient norm clipping at 1.0.

Denoising. For denoising, we run the “dopri5” ODE solver implemented in `torchdiffeq` [Chen, 2018] with absolute and relative tolerance fixed at 10^{-4} from $t = 0.999$ to $t = 0.001$. We use the same procedure as in the training process to obtain a discrete approximation of the learned latent prior for estimating the score along the reverse-time path.

More numerical results. We performed an additional synthetic experiment in which the nontrivial structure in the first two dimensions is discretely supported:

Example 3 (discrete): The first two coordinates of each θ_i^* are drawn uniformly from the discrete ensembles of three points $\{(0, 0), (0, 6), (6, 0)\}$.

The noises are then drawn with the same procedure as described earlier to produce the observations.

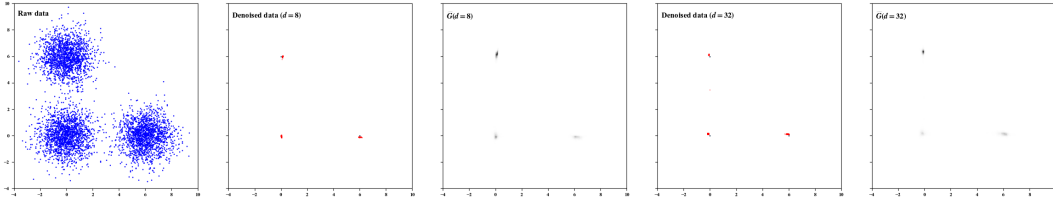


Figure 5: Projected plots of the discrete experiment onto the first two dimensions in dimensions $d \in (8, 16)$ and training sample size $n = 2 \times 10^4$. SDGM-C stays effective in estimating the discrete locations as d grows. Recent NPMLE solver in Zhang et al. [2022] demonstrates significantly degraded performance in such dimensions.

We also performed additional circle experiments with d further increases.

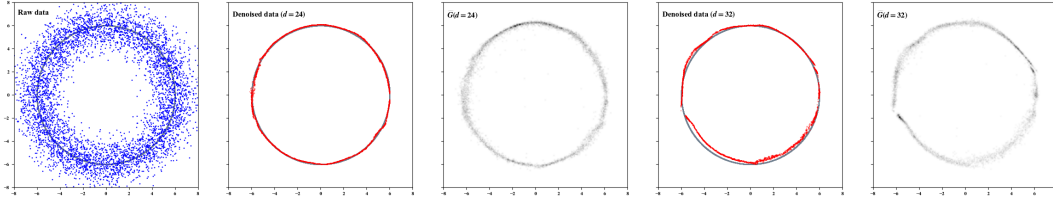


Figure 6: Projected plots of the circle experiment onto the first two dimensions in higher dimensions $d \in (24, 32)$ and training sample size $n = 10^7$. SDGM-C stays relatively stable in recovering the underlying circular structure as d grows to a dimension for which numerical NPMLE solvers struggle.