WHEN CAN DIFFUSION MODELS SOLVE INVERSE PROBLEMS ACCURATELY?

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models approximate the score of a target distribution to generate new samples. Here we ask how well a diffusion model learns the true data distribution's underlying features by posing this question as an inverse problem. When the target belongs to a parameterized family, can generated samples accurately identify the unknown parameter? We propose sensitivity of the diffusion score matching (DSM) loss as a mechanism for accurate parameter recovery. Specifically, we prove an information-theoretic sufficient condition under which generated samples recover a parameter accurately, even when the target score itself appears insensitive to parameter changes. For example, for Gaussian mixtures in arbitrary dimension, we show that the mixture-weight parameter is recoverable at the same order of accuracy as the DSM loss. This resolves an apparent paradox in the theoretical literature. Although the scores of well-separated Gaussian mixtures can yield poor mixture-weight estimates, we show that scores along the noising process remain sensitive to mixture weights and thereby enable accurate recovery. Experiments on benchmark datasets corroborate this mechanism: noise schedules designed to reduce sensitivity also degrade parameter recovery. Our results characterize when diffusion models can capture rare modes in multimodal mixtures and, more broadly, when diffusion-generated samples can support accurate solutions to scientific inverse problems.

1 INTRODUCTION

Generative models are powerful tools that are increasingly used in inverse problem settings where users seek to accurately estimate important properties of the true target distribution. For instance, parameters such as mixture weights, component variances, and mode amplitudes determine whether a generative model can serve as a reliable surrogate for downstream inference. Generative models that incorrectly reflect the amplitudes or variances of modes (Zhang et al., 2025; Roos et al., 2026; Hakemi et al., 2025) may hinder uncertainty quantification and the recovery of qualitative parameters (Bouchet et al., 2019; Addison et al., 2026; Mudur et al., 2023). Past work notes that diffusion models generate distributions that are close in Wasserstein or total variation distance when score learning is accurate Lee et al. (2022; 2023); De Bortoli (2022). However, these convergence results do not quantify how accurately the parameters of the target distribution are recovered from generated samples. Existing sample-quality metrics such as the Frechet Inception Distance (FID) (Heusel et al., 2017) are similarly structure-agnostic and may fail to detect parameter misrepresentation (Sajjadi et al., 2018). We frame parameter recovery as an explicit evaluation criterion for generative models and ask, *when can diffusion-generated samples support accurate solutions to inverse problems?*

In the theoretical literature, several works have noted that learning the score of a target distribution may be insufficient for accurate sampling (Koehler et al., 2023; Chen et al., 2021; Li et al., 2024), particularly for canonical multimodal distributions such as Gaussian mixtures (Alspach & Sorenson, 1972; Yakowitz & Spragins, 1968). But, in practice, diffusion models, despite the inexact score network trained on finite data, have been observed to cover well-separated modes of a multimodal distribution (Song et al., 2021a; Sajjadi et al., 2018; Ho et al., 2020; Biroli et al., 2024). Our work focuses on resolving this apparent paradox by introducing the idea of *score sensitivity* to parameters and using it to explain the mechanism behind accurate parameter recovery or lack thereof by diffusion models for a general class of distributions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To illustrate our proposed idea of *score sensitivity*, take for instance the weight parameter of a bimodal mixture model, as shown in Figure 1. The top row plots the score functions for two Gaussian mixture distributions that are identical except for the mixture weight γ and shows that they are nearly identical except in a region with very low probability mass. We thus likely lack sufficient training samples for the loss to reliably indicate the value of γ that best represents the training data. Any sampler that relies solely on the learned target score, such as Langevin dynamics, therefore exhibits near-identical behavior across different mixture weights. However, for larger values of t (or noise scale) in the noising process, the corresponding distributions are less well separated, and the score functions are clearly distinguishable over a higher-probability region. Since training diffusion models uses the score across multiple $t \in [0, 1]$, learned models hone in on the correct value of γ.

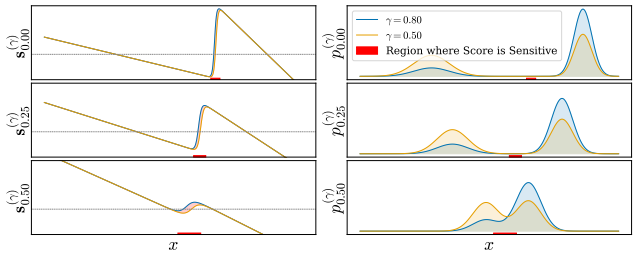


Figure 1: Score sensitivity to mixture weights at different diffusion times. At $t = 0$, the scores for $\gamma = 0.8$ and $\gamma = 0.5$ are nearly identical except in a low-probability region. At $t > 0$, the mixture components overlap more, and the region where the scores differ has a substantially higher probability.

Our contributions are as follows:

- We prove a strict minimax separation in parameter recovery between estimators using only target-score data and those with access to score data at intermediate time t, precisely when the score at t is more sensitive to parameter changes than at time 0 (Theorem 1).
- We propose the diffusion score sensitivity index (DSSI, (9)) as a computable diagnostic for parameter recovery and validate it by implementing sampling schemes with artificially reduced sensitivity that exhibit correspondingly poor recovery.
- For Gaussian mixtures, we prove a dimension-free lower bound on the DSSI showing that mixture-weight estimation error is controlled by the DDPM loss (Theorem 2), and evaluate the DSSI empirically on real data to illustrate the effect of noise schedule choices on both sensitivity and sample fidelity.

2 Preliminaries: Score-based diffusion models

Score-based diffusion models sample from a target distribution on $\mathbb{R}^d$ by tracing out a one-parameter family of distributions $\{p_t\}_{t\in[0,1]}$ indexed by a diffusion time t, with p_0 being the target data distribution and $p_1 \approx \mathcal{N}(\mathbf{0}, \mathbf{I})$ the standard normal distribution. A *forward process* progressively transforms samples from p_0 into samples from p_1 along this family. A *reverse process* inverts the densities along the forward process, starting from approximately $\mathcal{N}(0, \mathbf{I})$, using learned approximations of the scores of the intermediate marginals p_t. Thus, score-based diffusions have 3 components: (a) the forward diffusion process and its corresponding reverse process, (b) a training objective for learning the marginal score functions $\nabla \log p_t$ of the intermediate distributions for $t \in [0, 1]$, and (c) a sampling procedure used to approximately generate samples from the target data distribution at inference time. In this section, we briefly describe each component.

Noising process. Given a data distribution p_0 on $\mathbb{R}^d$, the noising process we consider is the variance-preserving forward SDE (VP-SDE) proposed in Song et al. (2021c). Give a noise schedule $\{\beta_t\}$, this takes the form

$$\mathrm{d}\boldsymbol{x}_t \;=\; -\tfrac{1}{2}\beta_t\,\boldsymbol{x}_t\,\mathrm{d}t \;+\; \sqrt{\beta_t}\,\mathrm{d}\mathbf{w}_t, \qquad \boldsymbol{x}_0 \sim p_0, \quad t \in [0,1], \tag{1}$$

where $\mathbf{w}_t$ is a standard d-dimensional Wiener process. We denote by p_t the marginal distribution of $\boldsymbol{x}_t$. The forward conditional law of $\boldsymbol{x}_t$ given $\boldsymbol{x}_0$ under (1) is Gaussian,

$$p_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0) \;=\; \mathcal{N}(\boldsymbol{x}_t;\; \sqrt{\alpha_t}\,\boldsymbol{x}_0,\; (1-\alpha_t)\,\mathbf{I})\,, \tag{2}$$

with conditional mean $\sqrt{\alpha_t}\boldsymbol{x}_0$ and conditional covariance $(1-\alpha_t)\mathbf{I}$, where $\alpha_t = e^{-\int_0^t \beta_s \mathrm{d}s}$.[1]

Denoising process. Anderson (1982) established that the time-reversal of (1) is itself an SDE:

$$\mathrm{d}\boldsymbol{x}_t \;=\; \left[-\tfrac{1}{2}\beta_t\,\boldsymbol{x}_t \;-\; \beta_t\,\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)\right]\mathrm{d}t \;+\; \sqrt{\beta_t}\,\mathrm{d}\bar{\mathbf{w}}_t, \qquad \boldsymbol{x}_1 \sim p_1, \tag{3}$$

where $\bar{\mathbf{w}}_t$ is the reverse-time Wiener process. Starting from $\boldsymbol{x}_1 \sim p_1$ and integrating the reverse SDE with the exact score function from time $t=1$ to $t=0$ produces samples $\boldsymbol{x}_0 \sim p_0$. In practice, this reverse SDE is simulated starting from $\boldsymbol{x}_1 \sim \mathcal{N}(0, \mathbf{I}) \approx p_1$ using the learned approximation $\boldsymbol{s}_{\boldsymbol{\theta}}$: $\mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ with parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ trained such that $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) \approx \nabla \log p_t(\boldsymbol{x})$ simultaneously for all noise levels to produce samples approximately distributed as the target p_0.

For a given trained score model $\{\boldsymbol{s}_{\boldsymbol{\theta}}(\cdot, t)\}_{t\in[0,1]}$ (or the true score function $\{\nabla \log p_t\}_{t\in[0,1]}$), generation of samples is an inference task: one may simulate the reverse SDE in (3) with any choice of consistent numerical integration schemes (Karras et al., 2022). The overall process has two distinct sources of error: the *approximation error* incurred in learning the score and the *sampling error* incurred by the choice of sampling algorithm (e.g., discretization error, finite timesteps, etc.).

3 Minimax bounds for parameter estimation from diffusion models

Let $\boldsymbol{\Gamma}$ be a finite-dimensional parameter space equipped with a norm, $|\cdot|$. For each $\gamma \in \boldsymbol{\Gamma}$, we write $p_0^{(\gamma)}$, or simply, $p^{(\gamma)}$, to denote the clean distribution at time 0, and $p_t^{(\gamma)}$, to denote the corresponding marginal along the noising process at time t. Similarly, let $\mathbf{s}^{(\gamma)}$ denote the score of the density $p^{(\gamma)}$, i.e., $\mathbf{s}^{(\gamma)} := \nabla \log p^{(\gamma)}$ and $\mathbf{s}_t^{(\gamma)} := \nabla \log p_t^{(\gamma)}$. The main result in this section is to show that when $\mathbf{s}_t^{(\gamma)}$ is sensitive to parameter γ at some t, the information-theoretic barrier preventing the recovery of γ from the target score alone is removed, yielding lower parameter estimation error. Before we introduce the result, we first recall score-matching and diffusion score-matching.

Classical score matching loss/relative Fisher information. The classical score matching (SM) loss measures the relative Fisher information between distributions (Hyvärinen, 2005). For a pair of distributions in our parameterized class,

$$\ell_{\mathrm{SM}}(p^{(\gamma')}, p^{(\gamma)}) := \mathbb{E}_{\boldsymbol{x}\sim p^{(\gamma)}} \|\mathbf{s}^{(\gamma')}(\boldsymbol{x}) - \mathbf{s}^{(\gamma)}(\boldsymbol{x})\|^2. \tag{4}$$

The relative Fisher information between a pair of mixture distributions (when $\gamma \in (0,1)$ is a mixture weight) with identical components but different mixture weights is small when the components are well-separated. In Fig. 1, the scores of Gaussian mixtures may appear nearly identical outside a low-probability region, an effect strengthened with increasing separation between mixture components as shown in Fig. 2d. Any sampling algorithm that relies solely on the score, such as Langevin sampling, exhibits near-identical behavior across distinct mixture distributions; consequently, the generated samples may not accurately reflect the target mixture weights.

Diffusion score matching loss for mixtures. The diffusion score matching (DSM) loss anneals the score matching loss across the noising process (1) (Song & Ermon, 2019).

$$\ell_{\mathrm{DSM}}(p^{(\gamma')}, p^{(\gamma)}) := \mathbb{E}_{\substack{t\sim\mathrm{Unif}(0,1)\\ \boldsymbol{x}\sim p_t^{(\gamma)}}}[\lambda_t \|\mathbf{s}_t^{(\gamma')}(\boldsymbol{x}) - \mathbf{s}_t^{(\gamma)}(\boldsymbol{x})\|^2] = \mathbb{E}_{t\sim\mathrm{Unif}(0,1)}[\lambda_t\, \ell_{\mathrm{SM}}(p_t^{(\gamma')}, p_t^{(\gamma)})]. \tag{5}$$

Here, λ_t is a weighting factor set to $\lambda_t = 1 - \alpha_t$ following Song et al. (2021c). The DSM loss measures the relative Fisher information over the entire sequence of noised densities generated by the process (1). Recall that the marginals p_t's and scores $\boldsymbol{s}_t$'s depend on the noise schedule $\{\alpha_t\}_t$, as shown in (2), and thus, the DSM loss ℓ_{DSM} is dependent on the choice of noise schedule. In practice, we may train a neural network $\boldsymbol{s}_{\boldsymbol{\theta}}$ to approximate the score function. With some abuse of notation, we denote $p^{(\boldsymbol{\theta})}$ the distribution[2] of samples generated using the score functions $\{\boldsymbol{s}_{\boldsymbol{\theta}}(\cdot, t)\}_{t\in(0,1)}$ for a fixed noise schedule as per (3). Since the true score function $\mathbf{s}_t^{(\gamma)}$ and indeed the target distribution $p^{(\gamma)}$ are unknown, calculating the DSM loss is intractable. Instead, we minimize the objective

$$\ell_{\mathrm{NN}}(p^{(\boldsymbol{\theta})}, p^{(\gamma)}) := \mathbb{E}_{\substack{t\sim\mathrm{Unif}(0,1)\\ \boldsymbol{x}_0\sim p^{(\gamma)}, \boldsymbol{x}_t\sim p_t^{(\gamma)}(\cdot|\boldsymbol{x}_0)}}[\lambda_t \|\mathbf{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \nabla \log p_t^{(\gamma)}(\boldsymbol{x}_t|\boldsymbol{x}_0)\|^2]. \tag{6}$$

[1] As a slight abuse of terminology, the sequences $\{\alpha_t\}$ and $\{\beta_t\}$ will interchangably be referred to as the "noise schedule".

[2] The superscript here indicates the trainable parameters $\boldsymbol{\theta}$ of the learned network.

As per Theorem 3, there exists a constant $C \geq 0$, depending only on $p^{(\gamma)}$, such that $\ell_{\mathrm{NN}}(p^{(\boldsymbol{\theta})}, p^{(\gamma)}) = \ell_{\mathrm{DSM}}(p^{(\boldsymbol{\theta})}, p_t^{(\gamma)}) + C$. Thus, $\ell_{\mathrm{DSM}}(p^{(\boldsymbol{\theta})}, p^{(\gamma)}) \leq \ell_{\mathrm{NN}}(p^{(\boldsymbol{\theta})}, p^{(\gamma)})$.

Minimax theory for parameter estimation. From classical minimax theory (e.g., Geer (2000)), we know that the error in any estimator of a parameter is lower-bounded by the informativeness of the data distribution about the parameter. Now, the data distribution loses information about the parameter in the noising process. How, then, would annealing benefit parameter recovery, if the original mixture is insensitive to parameter changes (e.g., as in Figure 1)? The answer lies in the fact that diffusion models learn to approximate the *scores at various noise levels*. Our main result in this section shows that sensitivity at an intermediate noise scale provably implies more accurate paramter estimation than that attainable from target-score data alone.

Definition 1 (Score oracle and score-restricted estimators)**.** *We consider a Gaussian White Noise (GWN) oracle from nonparametric regression (Tsybakov, 2008) of the form:* $\hat{\mathbf{s}}_t^{(\gamma)} = \mathbf{s}_t^{(\gamma)} + \sigma_t\, W$, *where* $\sigma_t > 0$ *is a scalar and* W *is an isonormal field indexed by* $L^2(p_t^{(\gamma)})$. *The statistical experiment is to estimate* γ *from the observation,* $\hat{\mathbf{s}}_t^{(\gamma)}$. *Let* $q_t^{(\gamma)}$ *denote the distribution of the observation, which is a Gaussian measure with mean* $\mathbf{s}_t^{(\gamma)}$. *We define the set,* $\Gamma_{\mathrm{score},t}$, *of all measurable functions of the GWN observations to* $\boldsymbol{\Gamma}$ *to be the class of "score-restricted estimators". Further, we use* $\mathbb{E}_{\gamma^*,t}$ *as a shorthand for expectations with respect to the score oracle model,* $q_t^{(\gamma)}$. *See also Remark 3 for more details on the estimator classes.*

We call $\Gamma_{\mathrm{score},t}$ *score-restricted estimators* since they access the score vector field (corrupted by Gaussian noise) and not the target data samples. Our class of estimators $\Gamma_{\mathrm{score},t}$ models parameter recovery achievable from having oracle access to a corrupt score vector field at time t, reflecting the denoising/generation phase in diffusion models after scores have been trained. Since the observation model, $q_t^{(\gamma)}$, is Gaussian, we can compute, fixing the indexing Hilbert space for W to be $L^2(p_t^{(\gamma)})$, that $\mathrm{KL}(q_t^{(\gamma)} || q_t^{(\gamma')}) = \ell_{\mathrm{SM}}(p_t^{(\gamma')}, p_t^{(\gamma)})/(2\sigma_t^2)$. See Remark 4 for this computation and elaboration on the connection between the KL divergence and ℓ_{DSM}.

Assumption 1 (Stable score-matching at time t)**.** *We assume that at some time* $t > 0$, *the score observation law,* $q_t^{(\cdot)}$, *satisfies* $|\gamma - \gamma'|^2 \leq A_t^2\, \mathrm{KL}(q_t^{(\gamma)} || q_t^{(\gamma')})$ *for some* $A_t > 0$ *and all* $\gamma, \gamma' \in \boldsymbol{\Gamma}$.

Since $\mathrm{KL}(q_t^{(\gamma)} || q_t^{(\gamma')}) = \ell_{\mathrm{SM}}(p_t^{(\gamma')}, p_t^{(\gamma)})/(2\sigma_t^2)$, the above assumption implies that when the score matching loss is bounded above by some C, then the squared difference in the corresponding parameters, $|\gamma - \gamma'|^2 \leq A_t^2\, C/(2\sigma_t^2)$. Intuitively, this means that, for an $O(\sigma_t^2)$ constant A_t, the score observations predict distances on the space of weight parameters. Our next definition allows us to make this setting rigorous by comparing score observations at time t to a different behavior of the scores at time 0 : the target score observations, $\hat{\mathbf{s}}^{(\gamma)}$, are insensitive in law to changes in γ and therefore do not predict distances on parameter space. In the next section, note that we obtain explicit estimates on A_t for Gaussian mixtures with γ being the mixture weight.

Definition 2 (ϵ-diffusion sensitivity)**.** *We say that a mixture family* $\{p^{(\gamma)}\}$ *exhibits* ϵ*-diffusion sensitivity if for every* $\gamma^* \in \boldsymbol{\Gamma}$, *there exist* ϵ*-packing sets* $S(\gamma^*) \subset \boldsymbol{\Gamma}$ *containing* $N(\gamma^*, \epsilon) \geq 16$ *elements with* $\mathrm{KL}(q^{(\gamma)} || q^{(\gamma')}) = \ell_{\mathrm{SM}}(p^{(\gamma')}, p^{(\gamma)})/(2\sigma_0^2) \leq \epsilon^2/4$, *for all* $\gamma, \gamma' \in S(\gamma^*)$ *satisfying* $|\gamma - \gamma'| \geq 4\sqrt{2}\, A_t \epsilon$. *Note that* A_t *here is as defined in Assumption 1.*

Theorem 1. *Fix* $t > 0$ *satisfying Assumption 1. Let* $\epsilon_t > 0$ *be a covering radius that solves* $\epsilon_t := \sigma_t \sqrt{V(\epsilon_t)}$, *where* $V(\epsilon_t)$ *is the* ϵ_t*-covering entropy*[3] *with respect to the square root KL divergence of the class of Gaussian measures* $\{q_t^{(\gamma)}\}$. *Suppose the bimodal mixture family satisfies Assumption 1 and exhibits an* ϵ*-diffusion sensitivity with* $\epsilon = \epsilon_t$. *Then, we have the following minimax upper bound on the score-restricted estimators at time* t,

$$\min_{\hat{\gamma} \in \Gamma_{\mathrm{score},t}} \max_{\gamma^*} \mathbb{E}_{\gamma^*,t} |\gamma^* - \hat{\gamma}|^2 \leq 2\, A_t^2\, \epsilon_t^2\,. \tag{7}$$

Simultaneously, if the packing set cardinality $N(\gamma^*, \epsilon_t) \geq e^{\sigma_t^2\, V(\epsilon_t)}$, *for all* γ^*, *it holds that the minimax lower bound achievable for target score-restricted estimators is the following:*

$$\min_{\hat{\gamma} \in \Gamma_{\mathrm{score},0}} \max_{\gamma^*} \mathbb{E}_{\gamma^*,0} |\gamma^* - \hat{\gamma}|^2 \geq 4\, A_t^2\, \epsilon_t^2. \tag{8}$$

[3] Refer to Definition 3 for an explicit statement

Proof. The bound in (7) is a direct application of (Yang & Barron, 1999, Theorem 3). The main novelty in the proof is the setup of score oracle-based estimators themselves. Note that the above standard definitions of packing and covering entropies extend verbatim to our infinite-dimensional statistical models, as discussed in Appendix C. Then, using a generalized Fano's method (Yu, 1997, Lemma 3) and an appropriate choice of δ as a function of ϵ in Definition 2 that limits the packing entropy of the cover of $(0, 1)$ with cardinality $e^{V(\epsilon_t)}$, we obtain the desired lower bound in (8). We defer a more detailed proof to Appendix C. □

Theorem 1 elucidates the conditions under which a diffusion model learns the mixture weights accurately when target score estimation followed by Langevin sampling does not. The above analysis applies to any bimodal mixture density and to any scalar parameter in place of the mixture weight. The main mechanism that enables parameter recovery is that the score field at some time t changes more rapidly with parameter changes. In contrast, the target score at time 0 may not vary, creating an information-theoretic barrier that limits the accuracy of any estimator. That is, no estimator can distinguish two close values of γ using a finite number of target score samples.

Yang & Barron (1999) constructs a Bayes density estimator that belongs to the mixture class and achieves the upper bound (7). Thus, we can conclude from Theorem 1 that using score observations at time t leads to provably better parameter recovery.

Resolving data-processing inequality. Saying a family of distributions exhibits an ϵ-sensitivity essentially means that more information about the parameter is created in the noising process, which, at first glance, appears to violate the data processing inequality. The law of noised samples, $p_t^{(\gamma)}$, cannot be more informative about γ at any time t than $p^{(\gamma)}$ (the target mixture), due to the data processing inequality. More precisely, since $p_t^{(\gamma)}$ results from a γ-independent convolution operation on $p_0^{(\gamma)}$, the data processing inequality gives: $\mathrm{KL}(p_t^{(\gamma')}||p_t^{(\gamma)}) \leq \mathrm{KL}(p^{(\gamma')}||p^{(\gamma)})$. But our key idea here is that diffusion models use the score observation fields, $\hat{\mathbf{s}}_t^{(\gamma)}$, to recover γ, even when $\hat{\mathbf{s}}^{(\gamma)}$ can not separate γ values sufficiently. Specifically, under the assumptions of Theorem 1 on the mixture family, the score observations can be more informative about γ at time t than at time 0, i.e., $\mathrm{KL}(q_t^{(\gamma')}||q_t^{(\gamma)}) \geq \mathrm{KL}(q^{(\gamma')}||q^{(\gamma)})$, without violating data processing. Definition 2 defines a structural property of mixture distributions: the family, $\{q_t^{(\gamma)}\}$ at intermediate t can be more separated in γ than at $t = 0$ when mixture components do not overlap. In Lemma 1, we prove why this counterintuitive inequality does not violate data processing and why no estimator in Γ_0 can reconstruct the score observation model that Γ_t has access to.

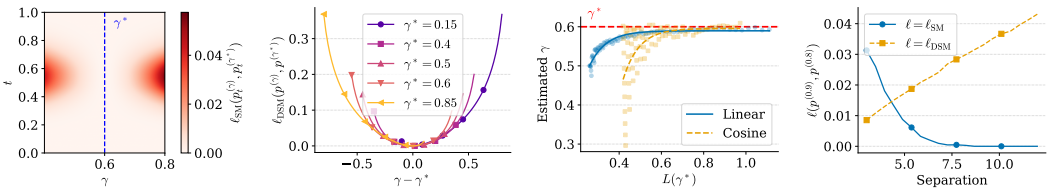
3.1 Effect of modified sensitivity to mixture weight

While Theorem 1 provides a mechanism for parameter recovery, we now discuss experimental evidence for the same mechanism but using a practically computable diagnostic. To quantify this phenomenon of parameter recovery despite insensitivity of the target score, we measure the sensitivity of the DSM loss ℓ_{DSM} (5) to changes in parameters of interest. Specifically, recall the set $\{p^{(\gamma)}\}_{\gamma\in\mathbf{\Gamma}}$ of parameterized distributions. Let $p^{(\gamma^*)}$ be a target distribution with target parameter γ^*. The *diffusion score sensitivity index (DSSI)* of the target parameter γ^* is defined as,

$$L(\gamma^*) = \inf_{\gamma\in\mathbf{\Gamma}} \frac{\ell_{\mathrm{DSM}}(p^{(\gamma)}, p^{(\gamma^*)})}{(\gamma - \gamma^*)^2}. \tag{9}$$

Both ℓ_{DSM} and $L(\gamma^*)$ depend on our choice of noise schedule. If the target data distribution has mixture weight γ^*, and we learn a distribution with mixture parameter γ, then the squared error of mixture weight estimation γ is bounded by the DSM loss divided by the DSSI. Thus, when the DSSI is small, our learned distribution may be highly misrepresentative of γ^* even when the DSM loss is low. In contrast, when $L(\gamma^*)$ is large, we know that a small DSM loss indicates low error in the parameter γ.

DSSI (9), unlike the the information-theoretic quantities used in Definition 2, is practically computable given score models trained on distributions with slightly different parameter values. Our numerical results show that the score sensitivity index quantifies how accurately a diffusion model captures the mixture weight of a Gaussian mixture model. We generate samples using the true analytical score function $\mathbf{s}_t^{(\gamma^*)}(\mathbf{x})$ and a sequence of modified noise schedules with a coarse time



(a) (b) (c) (d)

Figure 2: Score sensitivity for Gaussian mixture models in dimension $d = 10$. (a) $\ell_{\mathrm{SM}}(p_t^{(\gamma)}, p_t^{(\gamma^*)})$ as a function of t and γ. (b) $\ell_{\mathrm{DSM}}(p^{(\gamma)}, p^{(\gamma^*)})$ vs. $\gamma - \gamma^*$ for multiple γ^*, showing non-negligable sensitivity throughout. (c) Mixture weight estimates $\hat{\gamma}$ from generated data for $\gamma^* = 0.6$ under modified linear and cosine noise schedules (see Section E.1). The bias and variance of $\hat{\gamma}$ increase as $L(\gamma^*)$ decreases. (d) ℓ_{SM} and ℓ_{DSM} as a function of separation for $\gamma^* = 0.8$ and $\gamma = 0.9$. ℓ_{DSM} remains nonnegligible while ℓ_{SM} rapidly decays.

discretization ($T = 100$), designed to artificially reduce the DSSI $L(\gamma^*)$ (Section E.1). Recall that $L(\gamma^*)$ and ℓ_{DSM} depend on the choice of noise schedule $\{\alpha_t\}_t$.

We modify both linear and squared-cosine noise schedules and numerically estimate $L(\gamma^*)$ for these schedules. Fig. 2c shows the estimated mixture weight $\hat{\gamma}$ (calculated using Expectation-Maximization) plotted against $L(\gamma^*)$. As the sensitivity index $L(\gamma^*)$ decreases, both the bias and variance in $\hat{\gamma}$ increase. This demonstrates that noise schedules with lower $L(\gamma^*)$ yield mixture-weight estimates that are less robust to discretization errors, even with a perfect score estimator.

On the other hand, the means and covariances of the modes are uniformly well-approximated across the different noise schedules. This shows that errors in estimating the mixture weight can occur without noticeable artifacts in the geometry of individual modes. Modern accelerated diffusion sampling schemes (e.g., Song et al. (2021a)) leverage coarse time discretizations to reduce the computational cost of sampling while maintaining comparable sample fidelity. Our results suggest that such sampling approaches may be vulnerable to misrepresenting parameters of interest in the data distribution while appearing to generate accurate samples.

4 Diffusion Sensitivity to GMM Mixture Weights

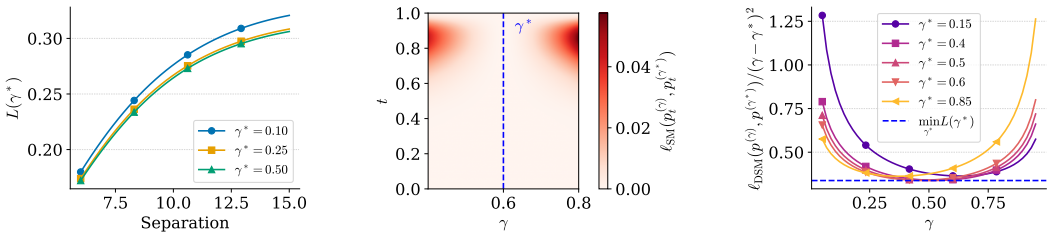
We show that for a bimodal isotropic Gaussian mixture in any dimension, the DSSI $L(\gamma^*)$ is non-negligible for the weight parameter γ^*. As a result, a diffusion model with a well-trained score generates samples with the correct weights even for extreme values of γ. See Remark 7 for why the bounds derived below are well-behaved for γ close to 0 or 1.

Theorem 2 (Diffusion sensitivity in bimodal Gaussian mixtures)**.** *Consider the family* $p^{(\gamma)}(\boldsymbol{x}) = (1-\gamma)\ \mathcal{N}(\boldsymbol{x}; \mu_0, \mathbf{I}) + \gamma\ \mathcal{N}(\boldsymbol{x}; \mu_1, \mathbf{I})$ *and a linear noise schedule* $\beta_t = \beta_1 t$ *for* $\beta_1 > 0$*. Let* $\alpha_t = \exp\left(-\int_0^t \beta_s \mathrm{d}s\right)$*. Then for any* $a \in (0, 1/2)$

$$\min_{\hat{\gamma} \in [a, 1-a]} \frac{\ell_{\mathrm{DSM}}(p^{(\hat{\gamma})}, p^{(\gamma^*)})}{(\hat{\gamma} - \gamma^*)^2} > \frac{(1 - \Phi(\|\mu_1 - \mu_0\|\sqrt{\alpha_1}))\left(\sqrt{k^2+16} - k\right)}{7\beta_1 \gamma^*(1-\gamma^*)} \exp\left(-(k^2+k)/4\right) \tag{10}$$

with $k := \left|\log \frac{1-a}{a}\right| + \left|\log \frac{1-\gamma^*}{\gamma^*}\right|$ *and* $\Phi(\cdot)$ *the CDF of* $\mathcal{N}(0,1)$.

Proof Sketch: The proof proceeds in three steps. We first decompose $\ell_{\mathrm{DSM}}(p^{(\hat{\gamma})}, p^{(\gamma^*)})$ by mixture component (36) to reduce the problem to lower-bounding the expectations over paths from $p^{(0)}$ and $p^{(1)}$ independently. Second, using the Probability Flow ODE (which is equivalent in probability to the forward process) (Song et al., 2021c), we find an analytical form for $\|\mathbf{s}_t^{(\hat{\gamma})} - \mathbf{s}_t^{(\gamma^*)}\|^2$ along paths $\mathbf{z}_t$ starting from $p^{(0)}$ (Proposition 1). Finally, we identify the interval of t values where this expression is bounded below (Proposition 2), integrate over this region of t (Proposition 3), and finally integrate over initial conditions $\mathbf{y}_0$. The bound is dimension-independent because for a given initial condition $\mathbf{y}_0$ starting from, e.g., mode $p^{(0)}$, the bound only depends on $\langle \mathbf{y}_0 - \mu_0, \mu_1 - \mu_0 \rangle / \|\mu_1 - \mu_0\| \sim \mathcal{N}(0,1)$ in any dimension. For full details, see Section D.2. □



(a) (b) (c)

Figure 3: Effect of Gaussian mixture model parameters on score sensitivity in dimension $d = 10$. (a) $L(\gamma^*)$ vs. mode separation for multiple $\gamma^* \leq 0.5$ (by symmetry $L(\gamma^*) = L(1-\gamma^*)$). $L(\gamma^*)$ is an increasing function of separation. (b) $\ell_{\mathrm{SM}}(p_t^{(\gamma)}, p_t^{(\gamma^*)})$ as a function of t and γ for a cosine noise schedule. Compared to a linear noise schedule, the noise schedule only affects the range of t in the forward process at which the score is sensitive to γ. (c) $\ell_{\mathrm{DSM}}(p^{(\gamma)}, p^{(\gamma^*)})/(\gamma - \gamma^*)^2$ vs. γ; $L(\gamma^*) \geq 0.27$ for all γ^*, with the minimum at $\gamma^* = 0.5$.

Remark 1. *If $\ell_{\mathrm{DSM}}(p^{(\gamma)}, p^{(\gamma^*)}) < \delta$ then $(\gamma - \gamma^*)^2 \leq \delta / L(\gamma^*)$ by the definition of $L(\gamma^*)$ in (9). Theorem 2 shows that $L(\gamma^*)$ cannot be too small under mild conditions for two-component Gaussian mixture models. Further, the neural network loss ℓ_{NN} (6) is an upper bound for ℓ_{DSM}, and so if we train a neural network to an error $\ell_{\mathrm{NN}}(p^{(\boldsymbol{\theta})}, p^{(\gamma)}) < \delta$ and our neural network learns a score corresponding to the mixture distributions $\{p_t^{(\gamma')}\}_{t \in (0,1)}$, then $\ell_{\mathrm{DSM}}(p^{(\gamma')}, p^{(\gamma^*)}) < \delta$. Our theory more directly implies that any score model which closely approximates the true score $\mathbf{s}^{(\gamma^*)}$ in L^2 will have small ℓ_{DSM}, and thus its sample distribution will accurately reflect γ^*.*

Remark 2. *If, when generating new samples at inference time, we alter the noise schedule $\{\alpha_t\}$, this may alter the mixture proportions in the generated samples. Specifically, altering the noise schedule changes ℓ_{DSM} and $L(\gamma^*)$, see (5). Hence, similar loss under ℓ_{DSM} for two different noise schedules could potentially correspond to very different guarantees for the accuracy of the recovered mixture weight $\hat{\gamma}$. Discretizing the reverse process SDE accumulates time-discretization errors regardless of score accuracy, meaning noise schedules that lower $L(\gamma^*)$ may worsen mixture-weight errors in the generative distribution, even for well-trained diffusion models. Section 3.1 showcases this effect for GMMs. In Section 5, we conduct further experiments on real-world data.*

Effect of parameters on DSSI. The DSSI $L(\gamma^*)$ depends on the mode separation $\|\mu_1 - \mu_0\|$, true mixture weight γ^*, and choice of noise schedule $\{\alpha_t\}_t$. First, consistent with the fact that $\ell_{\mathrm{DSM}}(p^{(\gamma)}, p^{(\gamma^*)})$ is an increasing function of the separation (Fig. 2d), Fig. 3a shows $L(\gamma^*)$ is an increasing function of the mode separation. Second, although Theorem 2 is only stated for linear noise schedules, different continuous noise schedules are equivalent up to time-reparameterization, guaranteeing $\ell_{\mathrm{SM}}(p_t^{(\hat{\gamma})}, p_t^{(\gamma^*)})$ is nonnegligible for some range of t under any noise schedule (Fig. 3b). Note however that $L(\gamma^*)$ is defined relative to ℓ_{DSM}, which averages $\ell_{\mathrm{SM}}(p_t^{(\hat{\gamma})}, p_t^{(\gamma^*)})$ over $t \in [0,1]$, meaning a noise schedule which reduces this range of t to a small interval $[a, a+\epsilon]$ scales down $L(\gamma^*)$ roughly proportional to ϵ, regardless of the peak value of $\ell_{\mathrm{SM}}(p_t^{(\hat{\gamma})}, p_t^{(\gamma^*)})$. Third, Fig. 3c shows that for mode separation $\|\mu_1 - \mu_0\| = 9$, $L(\gamma^*)$ is bounded below by 0.37 across all γ^*, implying that the squared prediction error in γ is at most ≈ 2.7 times larger than the squared diffusion score error. Notably, this bound is dimension-independent, and hence this holds for bimodal isotropic Gaussian mixtures in any dimension. In Section E.4, we showcase additional experiments on parameters of Gaussian mixtures other than the mixture weight.

4.1 Related Work

Multimodal capabilities of diffusion models. The ability of diffusion models to sample multimodal targets varies substantially across neural architectures (Hakemi et al., 2025), training recipes (Karras et al., 2022), and sampling algorithms (Roos et al., 2026; Zhang et al., 2025). Existing evaluation metrics (Sajjadi et al., 2018; Kynkäänniemi et al., 2019; Djolonga et al., 2020; Naeem et al., 2020; Alaa et al., 2022) are typically structure-agnostic and may fail to detect mode coverage (Räisä et al., 2025; Stein et al., 2023). The DSSI (9) provides a principled alternative that directly quantifies multimodal capability.

Blindness of score-matching estimators. Prior work (Wenliang & Kanagawa, 2020; Zhang et al., 2022) attributes the failure of naive score-based methods to a mismatch of distributional supports. We instead propose a support-agnostic mechanism, greater score sensitivity at intermediate noise scales, that provably enables parameter recovery for any distribution whose score is differentiable with respect to the parameter, including Gaussian mixtures with overlapping supports. Chewi et al. (2025) study the asymptotic statistical efficiency of DDPM-based estimators, without directly revealing the role of the noising process in enabling parameter recovery from practically trained models. Qin & Risteski (2024) establish sample complexity bounds for annealed score matching, complementing our focus on parameter recovery from generated samples.

Learning GMMs with DSM. DSM is *in principle* expressive enough to learn Gaussian mixtures (Shah et al., 2023), and learning algorithms with polynomial sample complexity exist for GMM density estimation (Gatmiry et al., 2025; Chen et al., 2025). When the learned score is L^2-accurate, the resulting diffusion sampler is provably close to the target in total variation (Chen et al., 2023; Lee et al., 2023). However, this may be insufficient to recover structural parameters that in low-probability regions. Li et al. (2024) suggest that diffusion models may fail to sample multimodal targets due to the lack of expressivity of the Gaussian transition steps in the reverse diffusion process. Our work complements Li et al. (2024) by identifying a different mechanism for multimodal sampling and clarifying the role of noise schedules.

Benefits of Annealing. Classical (unannealed) score matching (Hyvärinen, 2005) is ill-suited to multimodal targets, with asymptotic efficiency governed by the log-Sobolev constant, which grows exponentially in mode separation (Koehler et al., 2023; Chen et al., 2021). Song & Ermon (2019) proposed annealing over noise scale to handle multimodality, and subsequent work observes a phase transition at intermediate noise scales Biroli et al. (2024); Raya & Ambrogioni (2023). Qin & Risteski (2024) establish the first formal statistical benefit of annealing in diffusion via sample complexity bounds. Our work addresses parameter recovery from generated samples.

5 Sensitivity in real datasets

We empirically show that mixture weight information is captured in the score functions of intermediate forward process distributions, even outside the Gaussian mixture setting. Our experiments in this section can be repeated for any other unknown parameter, but we choose the mixture weight of a bimodal mixture due to its ease of visualization and interpretation. We explore how altering the noise schedule affects the DSSI and demonstrate that a small DSSI value quantitatively degrades mixture-weight recovery on non-Gaussian data. This is particularly salient to methods like (Song et al., 2021b; Lu et al., 2022; Karras et al., 2022) where noise schedules are selected to increase the speed of inference-time sampling; such methods have the potential to amplify some modes over others, augmenting the recent observation by Zhang et al. (2025).

We train a simple autoencoder on the ones and eights from MNIST ((LeCun et al., 2010)). Our target distribution $p^{(\gamma^*)}$ is then the mixture distribution of latent representations of eights and ones ($p^{(1)}$ is the distribution of ones). We set $\gamma^* = 0.4$ and use deep ResNets (He et al., 2015) with a linear noise schedule as our score matching models.

Sensitivity at intermediate t. In Fig. 4, the top row displays the score sensitivity over time and γ using trained diffusion models (each with a different noise schedule) to approximate the true score function (see Section E.3 for more details). The first column uses the same linear noise schedule as in Fig. 2a. Note that relative to the Gaussian mixture example, $\ell_{\mathrm{SM}}(p_t^{(\gamma)}, p_t^{(\gamma^*)})$ is non-negligable over a wider range of t values. We numerically estimate that the sensitivity index is $\approx 5\times$ larger than in the Gaussian mixture model example, so our theory guarantees a five-fold smaller error in the predicted mixture weight for the same level of error under ℓ_{DSM} as in the Gaussian setting.

Sensitivity affects parameter recovery. In Fig. 4a, the bottom row shows generated samples from a diffusion model with $T = 100$. We estimate the value of γ^* using generated samples and a trained classifier, shown below the generated samples. The generated samples accurately reflect the true value of the mixture weight, with only a 1% relative error in estimating γ^*.

To demonstrate the effect of the score sensitivity index $L(\gamma^*)$ on the error in estimating γ^*, we perform an ablation study on the trained diffusion model by modifying the noise schedule to reduce $L(\gamma^*)$ as in Section 3.1 (for details, see Section E.1). Fig. 4b and Fig. 4c show the score sensitivity and generated samples for the modified noise schedules, along with the corresponding estimates $\hat{\gamma}$ from the samples. When the score sensitivity index reduces by a factor of 10 (Fig. 4c), we see

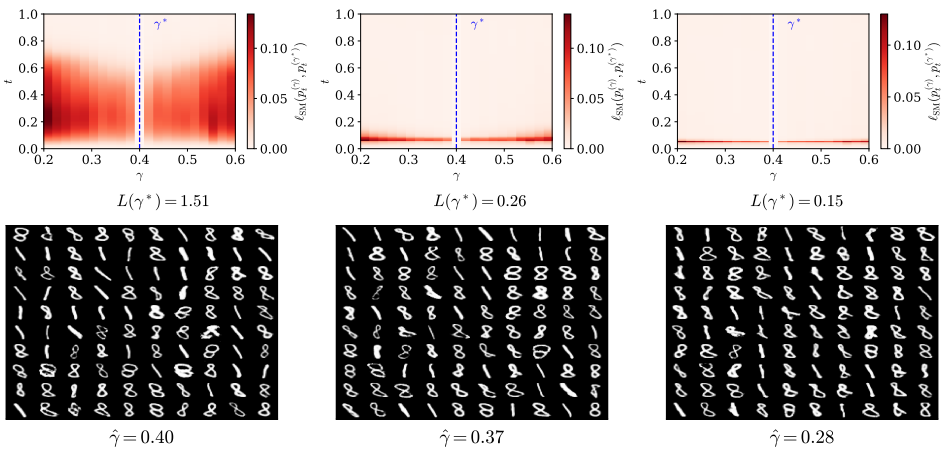


Figure 4: Influence of noise schedule on mixture weight recovery in MNIST. (a) Linear noise schedule; (b-c) two synthetic noise schedules designed to reduce $L(\gamma^*)$ (see Section E.1). Top row: $\ell_{\mathrm{SM}}(p_t^{(\gamma)}, p_t^{(\gamma^*)})$ as a function of t and γ estimated using trained diffusion models, with $\gamma^* = 0.4$. Bottom row: generated samples using each schedule, with estimated $\hat{\gamma}$ below. Larger $L(\gamma^*)$ yields more accurate mixture weight recovery.

that the estimate of γ^* has a relative error of 31%. Surprisingly, the samples from the modified noise schedules have comparable fidelity to those from the linear noise schedule (Fig. 4a). This is in line with the theoretical and empirical observations that say that diffusion models under errors and accelerated sampling schemes can still accurately learn the geometry of the support of the target distribution (Song et al., 2021b; Nichol & Dhariwal, 2021; Stanczuk et al., 2022; Chandramoorthy & de Clercq, 2025). See more in Section E.3.

6 Conclusion

Our main results (Theorem 1 and Theorem 2) address when and how diffusion models can generate samples that support accurate solutions of inverse problems. We answer three questions not covered by prior work:

1. **When does DDPM represent key features of distributions that score matching followed by Langevin sampling cannot?** We put forward the sensitivity of the score to parameter changes as a computable predictor for parameter recovery, rather than studying the statistical efficiency of DDPM-based estimators as prior work does.
2. **Why does intermediate score sensitivity enable more accurate parameter recovery?** Theorem 1 shows that when score sensitivity is larger at time t than at time 0, a practically trained diffusion model provably recovers parameters more accurately than Langevin sampling. Theorem 2 analytically verifies this for bimodal Gaussian mixtures.
3. **What is the effect of the noise schedule?** Sampling schemes that spend more time in sensitive regions achieve more accurate parameter recovery. We verify this numerically by computing the DSSI on benchmark datasets and showing that schedules designed to reduce sensitivity degrade recovery while leaving sample fidelity largely intact.

Limitations. We prove a quantitative relationship between the sensitivity index and diffusion score error only for Gaussian mixtures; Theorem 1 provides an information-theoretic mechanism that cannot be directly verified computationally. Evaluation of the DSSI on benchmark datasets requires surrogate generative models in place of the unknown target score. Finally, while the DSSI and the minimax bounds of Theorem 1 apply to multi-dimensional parameters γ, Theorem 2 is restricted to two-component Gaussian mixtures; generalizing its lower bound to more modes would require handling more complex interaction terms in the total derivative of the score.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for proofreading, and adding clarity to the exposition in some sections. The idea for the main results and the proofs can be entirely attributed to us, but we had helpful conversations with Claude and ChatGPT models during proof developments of Theorems 1 and 2. The final proof and the ideas were written without the use of any LLMs. We take full responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

There are no direct ethical concerns regarding this theoretical paper.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

We have attached detailed appendices with complete proofs and many remarks about the results in the paper. The last appendix section is devoted to ablation studies and further experimental results that verify the claims in the paper.

REFERENCES

Henry Addison, Elizabeth Kendon, Suman Ravuri, Laurence Aitchison, and Peter A. G. Watson. Machine learning emulation of precipitation from km-scale UK regional climate simulations using a diffusion model. *Journal of Advances in Modeling Earth Systems*, 2026. doi: 10.1029/2025MS005140. arXiv:2407.14158.

Ahmed M. Alaa, Boris van Breugel, Evgeny S. Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 290–306, 2022. arXiv:2102.08921.

D. Alspach and H. Sorenson. Nonlinear bayesian estimation using gaussian sum approximations. *IEEE Transactions on Automatic Control*, 17(4):439–448, 1972. doi: 10.1109/TAC.1972.1100034.

Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982. ISSN 0304-4149. doi: https://doi.org/10.1016/0304-4149(82)90051-5. URL https://www.sciencedirect.com/science/article/pii/0304414982900515.

Lukas Biewald. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15:9957, 2024. doi: 10.1038/s41467-024-54281-3.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

Freddy Bouchet, Joran Rolland, and Jeroen Wouters. Rare event sampling methods. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(8):080402, 2019. doi: 10.1063/1.5120509.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Yash Katariya, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

Nisha Chandramoorthy and Adriaan de Clercq. When and how can inexact generative models still sample from the data manifold? In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz,

M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 170326–170357. Curran Associates, Inc., 2025. doi: 10.52202/085713-5668. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f897cdacb44c0dff8845e8e07b448bcd-Paper-Conference.pdf.

Hong-Bin Chen, Sinho Chewi, and Jonathan Niles-Weed. Dimension-free log-sobolev inequalities for mixture distributions, 2021. ISSN 0022-1236. URL https://www.sciencedirect.com/science/article/pii/S0022123621003189.

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

Sitan Chen, Vasilis Kontonis, and Kulin Shah. Learning general gaussian mixtures with efficient score matching. In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 1029–1090. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/chen25e.html.

Sinho Chewi, Alkis Kalavasis, Anay Mehrotra, and Omar Montasser. Ddpm score matching is asymptotically efficient. In *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*, 2025.

Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022. arXiv:2208.05314.

Josip Djolonga, Mario Lucic, Marco Cuturi, Olivier Bachem, Olivier Bousquet, and Sylvain Gelly. Precision-recall curves using information divergence frontiers. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2550–2559, 2020.

Khashayar Gatmiry, Jonathan Kelner, and Holden Lee. Learning mixtures of gaussians using diffusion models. In Nika Haghtalab and Ankur Moitra (eds.), *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 2403–2456. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/gatmiry25b.html.

Sara A Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Shahin Hakemi, Naveed Akhtar, Ghulam Mubashar Hassan, and Ajmal Mian. Deeper diffusion models amplify bias. *arXiv preprint arXiv:2505.17560*, 2025.

Tiankai Hang, Shuyang Gu, Xin Geng, and Baining Guo. Improved noise schedule for diffusion training, 2024. URL https://arxiv.org/abs/2407.03297.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. URL https://api.semanticscholar.org/CorpusID:206594692.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3): 90–95, 2007. doi: 10.1109/MCSE.2007.55.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, December 2005. ISSN 1532-4435.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Patrick Kidger and Cristian Garcia. Equinox: neural networks in JAX via callable PyTrees and filtered transformations. *Differentiable Programming workshop at Neural Information Processing Systems 2021*, 2021.

Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TD7AnQjNzR6.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Krzysztof Łatuszyński, Matthew T Moores, and Timothée Stumpf-Fétizon. Mcmc for multi-modal distributions. *arXiv preprint arXiv:2501.05908*, 2025.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Holden Lee, Andrej Risteski, and Rong Ge. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering langevin monte carlo. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/c6ede20e6f597abf4b3f6bb30cee16c7-Paper.pdf.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. In *Advances in Neural Information Processing Systems*, volume 35, 2022. arXiv:2206.06227.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *Proceedings of the 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pp. 946–985. PMLR, 2023. arXiv:2209.12381.

Yangming Li, Boris van Breugel, and Mihaela van der Schaar. Soft mixture denoising: Beyond the expressive bottleneck of diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=aaBnFAyW9O.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, volume 35, 2022. arXiv:2206.00927.

Nayantara Mudur, Carolina Cuesta-Lazaro, and Douglas P. Finkbeiner. Cosmological field emulation and parameter inference with diffusion models. In *NeurIPS Workshop on Machine Learning and the Physical Sciences*, 2023. arXiv:2312.07534.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185, 2020.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

Yilong Qin and Andrej Risteski. Fit like you sample: sample-efficient generalized score matching from fast mixing diffusions. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 4413–4457. PMLR, 2024.

Ossi Räisä, Boris van Breugel, and Mihaela van der Schaar. Position: All current generative fidelity and diversity metrics are flawed. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=DMRrbb36r5.

Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.19693.

Nathan Roos, Ekaterina Iakovleva, Ani Gjergji, Vito Paolo Pastore, and Enzo Tartaglione. How i met your bias: Investigating bias amplification in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5374–5383, 2026.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

Kulin Shah, Sitan Chen, and Adam Klivans. Learning mixtures of gaussians using the DDPM objective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aig7sgdRfI.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021b. arXiv:2010.02502.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11895–11907, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021c. URL https://openreview.net/forum?id=PxTIG12RRHS.

Jan Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Your diffusion model secretly knows the dimension of the data manifold. *arXiv preprint arXiv:2212.12611*, 2022.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=08zf7kTOoh.

Alexandre B Tsybakov. Nonparametric estimators. In *Introduction to Nonparametric Estimation*, pp. 1–76. Springer, 2008.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. doi: 10.1162/NECO_a_00142.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Li K Wenliang and Heishiro Kanagawa. Blindness of score-based methods to isolated components and mixing proportions. *arXiv preprint arXiv:2008.10087*, 2020.

Sidney J. Yakowitz and John D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968. doi: 10.1214/aoms/1177698520.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

Bin Yu. *Assouad, Fano, and Le Cam*, pp. 423–435. Springer New York, New York, NY, 1997. ISBN 978-1-4612-1880-7. doi: 10.1007/978-1-4612-1880-7_29. URL https://doi.org/10.1007/978-1-4612-1880-7_29.

Mingtian Zhang, Oscar Key, Peter Hayes, David Barber, Brooks Paige, and François-Xavier Briol. Towards healing the blindness of score matching. *arXiv preprint arXiv:2209.07396*, 2022.

Yi Zhang, Zhenyu Liao, Jingfeng Wu, and Difan Zou. On the collapse errors induced by the deterministic sampler for diffusion models, 2025. URL https://arxiv.org/abs/2508.16154.

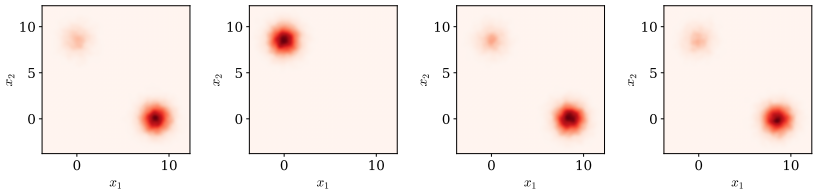
Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar C Tatikonda, Nicha Dvornek, Xenophon Papademetris, and James Duncan. Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. *Advances in Neural Information Processing Systems*, 33, 2020.

A Notation

Symbol	Definition	In words
$\mathcal{P}(p_0, p_1)$	$\{p^{(\gamma)} : p^{(\gamma)} = \gamma\, p_1 + (1-\gamma)\, p_0,\ \gamma \in (0,1)\}$	Family of bimodal mixtures of p_0 and p_1
p_t	$\mathrm{d}\boldsymbol{x}_t = -\frac{\beta_t}{2}\boldsymbol{x}_t \mathrm{d}t + \sqrt{\beta_t}\,\mathrm{d}W_t$ and $\boldsymbol{x}_0 \sim p \Longrightarrow \boldsymbol{x}_t \sim p_t$	Distribution of samples at time t in the forward process starting from p
$\ell_{\mathrm{NN}}(\boldsymbol{s}_{\boldsymbol{\theta}}, p)$	$\mathbb{E}_{[}[\boldsymbol{x}_0 \sim p,\ t,\ \boldsymbol{x}_t \mid \boldsymbol{x}_0]]\lambda_t \left\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \nabla \log p_t^{(\gamma^*)}(\boldsymbol{x}_t \mid \boldsymbol{x}_0)\right\|^2$	Diffusion model training objective
$\ell_{\mathrm{SM}}(p, q)$	$\mathbb{E}_{[}[\boldsymbol{x} \sim q]]\|\nabla \log p(\boldsymbol{x}) - \nabla \log q(\boldsymbol{x})\|^2$	Classical score matching loss
$\ell_{\mathrm{DSM}}(p, q)$	$\mathbb{E}_{[}[t \in (0,1),\ \boldsymbol{x}_t \sim q_t]]\lambda_t \left\|\nabla \log p_t(\boldsymbol{x}_t) - \nabla \log q_t(\boldsymbol{x}_t)\right\|^2$ $= \mathbb{E}_{[}[t \in (0,1)]]\lambda_t \ell_{\mathrm{SM}}(p, q)$	Diffusion score matching loss
$L(\gamma^*)$	$\min_{\gamma'} \frac{\ell_{\mathrm{DSM}}(p^{(\gamma')}, p^{(\gamma^*)})}{(\gamma^* - \gamma')^2}$	Diffusion score sensitivity index
$\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$	$\forall\, \boldsymbol{\theta} \in \boldsymbol{\Theta} = \mathbb{R}^{\#\text{ Parameters}},\ \boldsymbol{s}_{\boldsymbol{\theta}} : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$	Neural network with parameters $\boldsymbol{\theta}$

B Comparison of sampling algorithms for multimodal distributions

Sampling from multimodal distributions using classical Langevin sampling is known to be challenging (Lee et al., 2018; Łatuszyński et al., 2025). For a two-component Gaussian mixture, Fig. 2d shows that as mode separation increases, the sensitivity of the classical score matching loss to the mixture weight decays rapidly to 0 while the DSM loss does not. As a consequence, sampling algorithms based solely on the target score are unsuitable for sampling from multimodal distributions. Fig. 5 illustrates this effect, comparing classical Langevin sampling to diffusion sampling.



(a) (b) (c) (d)

Figure 5: Comparison between different sampling approaches on a high-separation Gaussian mixture model. (a) shows a KDE of the true data distribution. (b)-(c) show, in order, KDEs of samples using (b) Langevin sampling (run to time $T = 10^5$), (c) diffusion sampling using the analytic score function, and (d) diffusion sampling using a trained neural network. We can see that, unlike Langevin sampling, diffusion sampling accurately reflects the relative masses of the two modes.

C Minimax upper and lower bounds

Here we provide a more detailed proof of Theorem 1 and provide additional remarks on the Theorem and the underlying assumptions. We begin with a remark on score-restricted estimators (Definition 1) and explain the motivation for their definition.

In minimax theory, and in particular, Fano's method, the accuracy of an estimator for a parameter learned from a finite number of data samples is inherently limited by how well the data samples can distinguish between two parameter values. This distinguishability is often measured in terms of the KL divergence between the data distributions at two nearby parameter values. At first glance, the data samples given to a diffusion model appear to be the target samples, and as noted previously, the data distributions with well-separated modes do not distinguish between nearby values of weight parameters. Further, note $\mathrm{KL}(p_t^{(\gamma)}||p_t^{(\gamma')}) \leq \mathrm{KL}(p^{(\gamma)}||p^{(\gamma')})$. As a result, it might seem that the generated samples from the denoising process do not accurately recover the parameter (particularly, given the role that the KL divergence plays in Fano's inequality (Yu, 1997)), but our empirical results disagree with this expectation. In our empirical results, we see that the noising/denoising process affects parameter recovery compared with score approximation followed by Langevin dynamics. A key idea that helps resolve this mismatch is to recognize that the data a diffusion model uses to accurately estimate parameters is the noisy score (approximated by a neural network), not the noisy sample data. The generated samples, from the reverse dynamics, are a function of the noisy scores (interpolated by neural networks). These neural network interpolations at some time t can have a distribution that distinguishes between parameter values more than at time 0. In that sense, the time t scores contain sufficient information to generate samples corresponding to a nearby parameter value, while the time 0 score (target score) may not.

Our score-restricted estimators define the notion of recovering γ^* from "score data" rather than sample data. Such a notion models parameter estimation from the generated samples in score-based generative models where the score field is first approximated and then used for sampling. To distinguish Langevin sampling from diffusion models, we define two separate classes of estimators: one class has access to only a target score oracle and the other class has access to a score oracle at an intermediate time along the noising process. Hence, the estimator classes $\Gamma_{\mathrm{score},t}$ and $\Gamma_{\mathrm{score},0}$ differ in the data distributions given to them, which are, respectively, $(q_t^{(\gamma)})$ and $(q_0^{(\gamma)})$. In Lemma 1, we show that the oracle model $q_t^{(\gamma)}$ cannot be reconstructed in a γ-independent manner from the oracle $q_0^{(\gamma)}$, and this induces a separation between the two estimator classes.

Remark 3. *While in Theorem 1, we are only comparing $\Gamma_{\mathrm{score},0}$ and a set of estimators that use score data at one other time t, the result continues to hold if instead under a specific redefinition of the two estimator classes. Let t still be a time at which the diffusion sensitivity definition holds for the values of ϵ and δ defined in Theorem 1. The estimator class $\Gamma_{\mathrm{score},0}$ contains any sampling scheme that does not use samples from $q_{t'}^{(\gamma)}$, for any time except $t' = 0$. On the other hand, the estimators in $\Gamma_{\mathrm{score},0}$ must include sampling (denoising) steps that access score data from $q_t^{(\gamma)}$.*

Remark 4. *To compute the KL divergence,* $\mathrm{KL}(q_t^{(\gamma)}||q_t^{(\gamma')})$*, we first recognize that both Gaussian measures are on* $L^2(p_t^{(\gamma)})$*. Hence, by Cameron-Martin formula, the Radon-Nikodym derivative,* $dq_t^{(\gamma)}/dq_t^{(\gamma')}$*. We can take expectations with respect to* $q_t^{(\gamma)}$ *to arrive at the formula* $\mathrm{KL}(q_t^{(\gamma)}||q_t^{(\gamma')}) = \|\mathbf{s}_t^{(\gamma)} - \mathbf{s}_t^{(\gamma')}\|^2_{L^2(p_t^{(\gamma)})}/(2\sigma_t^2)$*. This resembles the KL divergence between finite-dimensional Gaussian distributions with the same covariance. For the KL between the data distributions, applying the de Bruijn identity to (1) gives*

$$\frac{d}{dt}\mathrm{KL}(p_t^{(\gamma)}\|p_t^{(\gamma')}) = -\frac{\beta_t}{2}\,\ell_{\mathrm{SM}}(p_t^{(\gamma')}, p_t^{(\gamma)}).$$

Since $p_1^{(\gamma)} \approx p_1^{(\gamma')} \approx \mathcal{N}(0, I)$*, integrating from* 0 *to* 1 *and using the boundary condition* $\mathrm{KL}(p_1^{(\gamma)}\|p_1^{(\gamma')}) \approx 0$ *gives:*

$$\mathrm{KL}(p^{(\gamma)}\|p^{(\gamma')}) \approx \frac{1}{2}\int_0^1 \beta_t\,\ell_{\mathrm{SM}}(p_t^{(\gamma')}, p_t^{(\gamma)})\,dt.$$

Both the KL and ℓ_{DSM} *are weighted integrals of* $\ell_{\mathrm{SM}}(p_t^{(\gamma')}, p_t^{(\gamma)})$ *over* $t \in [0,1]$*, but with different weights. The KL uses weight* $\beta_t/2$*, while* ℓ_{DSM} *uses* $\lambda_t = 1 - \alpha_t$*. Since* $\alpha_t = e^{-\int_0^t \beta_s ds}$*, these weights are related but not equal in general. The two are proportional if and only if* $\beta_t \propto 1 - \alpha_t$*, which holds approximately for small* t *(where* $1 - \alpha_t \approx \int_0^t \beta_s ds$*).*

Lemma 1. *There is no* γ*-independent post processing of score observations at time 0,* $\hat{\mathbf{s}}_0^{(\gamma)}$*, that can reproduce* $\hat{\mathbf{s}}_t^{(\gamma)}$*.*

Proof. We can prove this by contradiction. If there exists a γ-independent processing of $q^{(\gamma)}$ that produces the distribution $q_t^{(\gamma)}$, the data processing inequality gives $\mathrm{KL}(q_t^{(\gamma)}||q_t^{(\gamma')}) \leq \mathrm{KL}(q^{(\gamma)}||q^{(\gamma')})$. But, we have assumed that at time t, our mixture family is such that $\mathrm{KL}(q^{(\gamma)}||q^{(\gamma')}) \leq \epsilon_t^2/4 \leq (1/128)(\gamma-\gamma')^2/A_t^2 \leq (1/128)\mathrm{KL}(q_t^{(\gamma)}||q_t^{(\gamma')})$. Thus, the oracle observations and the statistical experiment that can be performed with the model $q_t^{(\gamma)}$ cannot be simulated from having access to the oracle, $q^{(\gamma)}$. □

Before we give the proof of Theorem 1, we recapitulate a corollary of Generalized Fano's method from Lemma 3 of (Yu, 1997) and define covering entropy, which is used in Theorem 1.

Lemma 2. *[Lack of identifiability of mixture weights] Let* $G := \{\gamma_1, \cdots, \gamma_N\} \in (0,1)^N$ *be a set of* $N \geq 16$ *values of mixture weights such that* $|\gamma_i - \gamma_j| \geq \varepsilon,\ i \neq j,\ 1 \leq i, j \leq N$*. Then, for the class of score-restricted estimators (Definition 1),* $\Gamma_{\mathrm{score},0}$*, with oracle access to the model,* $q^{(\gamma^*)}$*, if* $\mathrm{KL}(q^{(\gamma_i)}||q^{(\gamma_j)}) \leq \log N/4$*, we have*

$$\inf_{\hat{\gamma}\in\Gamma_{\mathrm{score},0}} \max_{\gamma^*\in G} \mathbb{E}_{q^{(\gamma^*)}}[|\hat{\gamma} - \gamma^*|^2] \geq \varepsilon^2/8. \tag{11}$$

The proof follows from (Yu, 1997, Lemma 3), which is an application of Fano's inequality. To get the error in the squared norm, $\mathbb{E}[|\hat{\gamma} - \gamma^*|^2$, as we have in (11), we need to additionally use Markov's inequality: $\mathbb{E}[|\hat{\gamma} - \gamma|^2 \geq (\varepsilon/2)^2\mathbb{P}(|\hat{\gamma} - \gamma^*| > \varepsilon/2) \geq (\varepsilon/2)^2\,(1/2)$. The second inequality directly follows from the proof of Lemma 3 of (Yu, 1997). The main difference here is that the data used by the estimator corresponds to the score (Definition 1), chosen to exemplify the diffusion model setting.

Definition 3. *[*ϵ*-covering entropy] Let* O_t *be a finite open cover of* $[0,1]$*, with centers* $\gamma_1, \cdots, \gamma_{N_t}$*, such that for any* $\gamma \in (0,1)$*,* $\min_j \mathrm{KL}(q_t^{(\gamma)}||q_t^{(\gamma_j)}) < \epsilon^2$*. Then,* $V_{\epsilon,t} := \log|O_t|$ *is called the* ϵ*-covering entropy.*

The main novelty in the proof of Theorem 1 is the setup of score oracle-based estimators. Note that the above standard definitions of packing and covering entropies extend verbatim to our infinite-dimensional statistical models.

Secondly, while annealing has been known to be statistically beneficial (Koehler et al., 2023; Qin & Risteski, 2024), our result proves better minimax bounds for parameter recovery, rather than density estimation. To complete the upper bound in (7), we apply (Yang & Barron, 1999, Theorem 2) directly. To obtain the lower bound, we use Lemma 2 with a possibly different packing set of size $N(\gamma^*)$ around each γ^*. Using the definition of ϵ-sensitivity and assuming that $\log N(\gamma^*) \geq \max\{\log 16, \sigma_t^2\ V(\epsilon_t)\}$ at all γ^*, we obtain that for any $\gamma, \gamma' \in S(\gamma^*)$, $\mathrm{KL}(q^{(\gamma)}||q^{(\gamma')}) \leq \delta = \epsilon_t^2/(4\sigma_t^2) \leq (\log N(\gamma^*)/V(\epsilon_t))\epsilon_t^2/4 = \log N(\gamma^*)/4$, since $\epsilon_t^2\ \sigma_t^2 = V(\epsilon_t)$. As a result, the conditions of Lemma 2 apply at every γ^*.

D Diffusion process on Gaussian mixture models

D.1 Mixture Family Preliminaries

Let $p^{(0)}$ and $p^{(1)}$ be two fixed probability distributions on $\mathbb{R}^d$. We recall the family of mixed probability densities

$$\mathcal{P}(p^{(0)}, p^{(1)}) \ := \ \left\{ p^{(\gamma)} := (1-\gamma)\, p^{(0)}(\boldsymbol{x}) + \gamma\, p^{(1)}(\boldsymbol{x}) \ \middle|\ \gamma \in [0,1] \right\}.$$

The set of noisy marginals at noise level t produced by the forward process (1) applied to samples $\boldsymbol{x}_0 \sim p^{(\gamma)}$ is denoted $\{p_t^{(\gamma)}\}_{t\in[0,1]}$. The forward process preserves the mixture structure: the noisy marginal $p_t^{(\gamma)}$ is itself a mixture of the noisy component marginals $p_t^{(0)}$ and $p_t^{(1)}$, with the same mixture weight γ. The family of noisy mixtures and the noisy family of mixtures therefore coincide.

Lemma 3 (Mixture closure under forward VP-SDE)**.** *For every $\gamma \in (0,1)$ and every $t \in [0,1]$,*

$$p_t^{(\gamma)}(\boldsymbol{x}) = (1-\gamma)p_t^{(0)}(\boldsymbol{x}) + \gamma p_t^{(1)}(\boldsymbol{x}) \tag{12}$$

Proof. Let $k_t(\mathbf{y} \mid \boldsymbol{x}) := \mathcal{N}(\mathbf{y};\ \sqrt{\alpha_t}\,\boldsymbol{x},\ (1-\alpha_t)\,\mathbf{I})$ denote the forward conditional density (2). Since k_t does not depend on the initial distribution, the marginal at noise level t is obtained by a linear convolution against the initial density:

$$p_t^{(\gamma)}(\boldsymbol{x}) \ = \ \int_{\mathbb{R}^d} k_t(\boldsymbol{x} \mid \boldsymbol{x}_0)\, p^{(\gamma)}(\boldsymbol{x}_0)\, \mathrm{d}\boldsymbol{x}_0.$$

Substituting the definition of the mixture and applying the linearity of integration,

$$\begin{aligned} p_t^{(\gamma)}(\boldsymbol{x}) \ &= \ (1-\gamma) \int_{\mathbb{R}^d} k_t(\boldsymbol{x} \mid \boldsymbol{x}_0)\, p^{(0)}(\boldsymbol{x}_0)\, \mathrm{d}\boldsymbol{x}_0 \ + \ \gamma \int_{\mathbb{R}^d} k_t(\boldsymbol{x} \mid \boldsymbol{x}_0)\, p^{(1)}(\boldsymbol{x}_0)\, \mathrm{d}\boldsymbol{x}_0 \\ &= \ (1-\gamma)\, p_t^{(0)}(\boldsymbol{x}) + \gamma\, p_t^{(1)}(\boldsymbol{x}). \end{aligned}$$

□

A useful consequence of Lemma 3 is that samples from $p_t^{(\gamma)}$ can be drawn by first sampling a component label $c \sim \mathrm{Ber}(\gamma)$ and then drawing $\mathbf{y}_t \sim p_t^{(c)}$, a component-wise decomposition that simplifies subsequent analysis.

D.1.1 The score of the noisy mixture

Differentiating $\log p_t^{(\gamma)}$ and applying Lemma 3 gives the score of the noisy mixture as a pointwise convex combination of the component scores,

$$\nabla_{\boldsymbol{x}} \log p_t^{(\gamma)}(\boldsymbol{x}) \ = \ w_0^t(\boldsymbol{x})\, \nabla_{\boldsymbol{x}} \log p_t^{(0)}(\boldsymbol{x}) \ + \ w_1^t(\boldsymbol{x})\, \nabla_{\boldsymbol{x}} \log p_t^{(1)}(\boldsymbol{x}), \tag{13}$$

where the weights are the EM responsibilities — the posterior probabilities under the generative model that $\boldsymbol{x}$ was produced by component 0 or component 1, respectively (Bishop, 2006, §9.2),

$$w_0^t(\boldsymbol{x}) \ := \ \frac{(1-\gamma)\, p_t^{(0)}(\boldsymbol{x})}{p_t^{(\gamma)}(\boldsymbol{x})}, \qquad w_1^t(\boldsymbol{x}) \ := \ \frac{\gamma\, p_t^{(1)}(\boldsymbol{x})}{p_t^{(\gamma)}(\boldsymbol{x})}, \tag{14}$$

and satisfy $w_0^t(\boldsymbol{x}) + w_1^t(\boldsymbol{x}) = 1$.

D.1.2 SPECIALIZATION: GAUSSIAN COMPONENTS

When the components are Gaussian, $p^{(c)} = \mathcal{N}(\boldsymbol{\mu}_c^*, \boldsymbol{\Sigma}_c^*)$ for $c \in \{0, 1\}$, the noisy marginals remain Gaussian, $p_t^{(c)} = \mathcal{N}(\boldsymbol{\mu}_{c,t}^*, \boldsymbol{\Sigma}_{c,t}^*)$, and the score (13) takes the explicit form

$$\nabla_{\boldsymbol{x}} \log p_t^{(\gamma)}(\boldsymbol{x}) \;=\; \sum_{c\in\{0,1\}} w_c^t(\boldsymbol{x})\,(\boldsymbol{\Sigma}_{c,t}^*)^{-1}(\boldsymbol{\mu}_{c,t}^* - \boldsymbol{x}), \tag{15}$$

a responsibility-weighted sum of precision-scaled displacements $(\boldsymbol{\Sigma}_{c,t}^*)^{-1}(\boldsymbol{\mu}_{c,t}^* - \boldsymbol{x})$. The responsibility weights (14) admit the explicit closed form

$$w_c^t(\boldsymbol{x}) \;=\; \frac{\pi_c \mathcal{N}\big(\boldsymbol{x};\, \boldsymbol{\mu}_{c,t}^*,\, \boldsymbol{\Sigma}_{c,t}^*\big)}{\sum_{c'\in\{0,1\}} \pi_{c'}\, \mathcal{N}\Big(\boldsymbol{x};\, \boldsymbol{\mu}_{c',t}^*,\, \boldsymbol{\Sigma}_{c',t}^*\Big)}, \tag{16}$$

with mixture weights $\pi_0 = 1 - \gamma$ and $\pi_1 = \gamma$. The numerator and denominator together carry information about γ, the Mahalanobis distance from $\boldsymbol{x}$ to each component mean, and the Gaussian normalizing constants $\det(\boldsymbol{\Sigma}_{c,t}^*)^{-1/2}$.

D.1.3 DDPM-STYLE FORWARD PROCESS

Lemma 4. *For $p^{(0)} := \mathcal{N}(\boldsymbol{x}; \mu_0, \mathbf{I})$ and $p^{(1)} := \mathcal{N}(\boldsymbol{x}; \mu_1, \mathbf{I})$ denote two individual component distributions. For a variance-preserving forward process with noise schedule $(\bar{\alpha}_t)_{t\in[0,T]}$, the parameters of the mixture modes at time t are*

$$\boldsymbol{\mu}_{c,t}^* \;=\; \sqrt{\bar{\alpha}_t}\,\boldsymbol{\mu}_c^*, \qquad \boldsymbol{\Sigma}_{c,t}^* \;=\; \bar{\alpha}_t\,\boldsymbol{\Sigma}_c^* + (1 - \bar{\alpha}_t)\,\mathbf{I}, \tag{17}$$

Hence, the noisy marginal distributions are,

$$p_t^{(0)}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \sqrt{\alpha_t}\mu_0, \mathbf{I}), \qquad p_t^{(1)}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \sqrt{\alpha_t}\mu_1, \mathbf{I})$$

Proof. Let $\boldsymbol{x}_0 \sim \mathcal{N}(\mu, \boldsymbol{\Sigma})$. We have that

$$\begin{aligned} &\boldsymbol{x}_t|\boldsymbol{x}_0 \sim \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{x}_0, (1-\alpha_t)\mathbf{I}) \\ \implies &\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\mathbf{z} \end{aligned}$$

where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ is independent of $\boldsymbol{x}_0$. Therefore,

$$\boldsymbol{x}_t \sim \mathcal{N}\left(\sqrt{\alpha_t}\mu, \alpha_t\boldsymbol{\Sigma} + (1-\alpha_t)\mathbf{I}\right)$$

□

As per (17)m α_t decreases from 1 to 0, the component means contract toward the origin and the covariances inflate toward $\mathbf{I}$. Substituting (17) into (16) reveals two limits of interest. At small t ($\alpha_t \to 1$), the perturbed components remain well separated, and the responsibilities concentrate sharply on the closest mode, recovering the M-step regime of the EM algorithm. At large t ($\alpha_t \to 0$), the perturbed components all approach $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and the responsibilities flatten toward the prior weights $(1-\gamma, \gamma)$, becoming nearly independent of $\boldsymbol{x}$. Sensitivity of the score to the mixture parameter γ is therefore concentrated at the noise scales for which the responsibility weights are neither saturated nor uninformative — a window whose location and width are determined by the geometry of the components.

We now demonstrate an equivalence between the neural network loss (6) and the DSM loss (5).

Theorem 3. *Let $p^{(\gamma)}$ be a data distribution on $\mathbb{R}^d$ with $\mathbb{E}_{p^{(\gamma)}}\|\boldsymbol{x}_0\|^2 < \infty$, and let $\{p_t^{(\gamma)}\}_{t\in[0,1]}$ be its marginals under the VP-SDE (1)–(2). Let $\boldsymbol{s}_{\boldsymbol{\theta}} : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ be a score model with $\mathbb{E}_{p_t^{(\gamma)}}\|\boldsymbol{s}_{\boldsymbol{\theta}}(\cdot, t)\|^2 < \infty$ for almost every $t \in [0,1]$. Then*

$$\ell_{\mathrm{NN}}(p^{(\boldsymbol{\theta})}, p^{(\gamma)}) \;=\; \ell_{\mathrm{DSM}}(p^{(\boldsymbol{\theta})}, p^{(\gamma)}) \;+\; C(p^{(\gamma)}), \tag{18}$$

where the constant $C(p^{(\gamma)})$ does not depend on $\boldsymbol{\theta}$ and admits the closed form

$$C(p^{(\gamma)}) \;=\; \int_0^1 \lambda_t \cdot \frac{\alpha_t}{(1-\alpha_t)^2}\,\mathrm{MMSE}_t(p^{(\gamma)})\,dt, \tag{19}$$

$$\mathrm{MMSE}_t(p^{(\gamma)}) \;:=\; \mathbb{E}_{\substack{\boldsymbol{x}_0 \sim p^{(\gamma)} \\ \boldsymbol{x}_t \sim p_t^{(\gamma)}(\cdot\mid \boldsymbol{x}_0)}} \Big[\, \big\| \boldsymbol{x}_0 - \widehat{\boldsymbol{x}}_0(\boldsymbol{x}_t) \big\|^2 \,\Big], \tag{20}$$

$$\widehat{\boldsymbol{x}}_0(\boldsymbol{x}_t) \;:=\; \mathbb{E}_{\boldsymbol{x}_0' \sim p^{(\gamma)}(\cdot\mid \boldsymbol{x}_t)}[\boldsymbol{x}_0']. \tag{21}$$

Moreover, $C(p^{(\gamma)}) \geq 0$, with equality if and only if $p^{(\gamma)}$ is a Dirac mass. In particular,

$$\ell_{\mathrm{NN}}(p^{(\boldsymbol{\theta})}, p^{(\gamma)}) \;\geq\; \ell_{\mathrm{DSM}}(p^{(\boldsymbol{\theta})}, p^{(\gamma)}) \tag{22}$$

for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Proof. Our arguments follow the reasoning in Vincent (2011) with additional logic for guaranteeing the required relation between the neural network loss and the DSM loss. Throughout, all expectations are taken under $p^{(\gamma)}(\boldsymbol{x}_0)$, $p_t^{(\gamma)}(\boldsymbol{x}_t \mid \boldsymbol{x}_0)$ for the joint, and under $p_t^{(\gamma)}(\boldsymbol{x}_t)$ for the marginal. We suppress the dependence on t in $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ when convenient. Fix $t \in [0,1]$. Expanding the squared norm in the inner expectation of ℓ_{DSM},

$$\mathbb{E}_{\boldsymbol{x}_t}\big\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t) - \mathbf{s}_t^{(\gamma)}(\boldsymbol{x}_t)\big\|^2 = \mathbb{E}_{\boldsymbol{x}_t}\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\|^2 \;-\; 2\,b_t(\boldsymbol{\theta}) \;+\; \mathbb{E}_{\boldsymbol{x}_t}\big\|\mathbf{s}_t^{(\gamma)}(\boldsymbol{x}_t)\big\|^2, \tag{23}$$

with $b_t(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{x}_t}\big\langle \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t),\, \mathbf{s}_t^{(\gamma)}(\boldsymbol{x}_t)\big\rangle$. Similarly,

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{x}_t}\big\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t) - \nabla_{\boldsymbol{x}_t}\log p_t^{(\gamma)}(\boldsymbol{x}_t\mid\boldsymbol{x}_0)\big\|^2 \\ &\qquad = \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{x}_t}\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\|^2 \;-\; 2\,a_t(\boldsymbol{\theta}) \;+\; \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{x}_t}\big\|\nabla_{\boldsymbol{x}_t}\log p_t^{(\gamma)}(\boldsymbol{x}_t\mid\boldsymbol{x}_0)\big\|^2,\end{aligned} \tag{24}$$

with $a_t(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{x}_t}\big\langle \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t),\, \nabla_{\boldsymbol{x}_t}\log p_t^{(\gamma)}(\boldsymbol{x}_t\mid\boldsymbol{x}_0)\big\rangle$. The first term in each expansion coincides because $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)$ depends only on $\boldsymbol{x}_t$. We now claim that for all θ, $b_t(\boldsymbol{\theta}) = a_t(\boldsymbol{\theta})$. Using $\mathbf{s}_t^{(\gamma)} = \nabla p_t^{(\gamma)}/p_t^{(\gamma)}$ and the marginalization $p_t^{(\gamma)}(\boldsymbol{x}_t) = \int p_t^{(\gamma)}(\boldsymbol{x}_t\mid\boldsymbol{x}_0)\,p^{(\gamma)}(\boldsymbol{x}_0)\,d\boldsymbol{x}_0$,

$$\begin{aligned}b_t(\boldsymbol{\theta}) &= \int p_t^{(\gamma)}(\boldsymbol{x}_t)\,\big\langle \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t),\, \mathbf{s}_t^{(\gamma)}(\boldsymbol{x}_t)\big\rangle\,d\boldsymbol{x}_t \\ &= \int \big\langle \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t),\, \nabla_{\boldsymbol{x}_t} p_t^{(\gamma)}(\boldsymbol{x}_t)\big\rangle\,d\boldsymbol{x}_t \\ &= \int \Big\langle \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t),\, \nabla_{\boldsymbol{x}_t}\int p_t^{(\gamma)}(\boldsymbol{x}_t\mid\boldsymbol{x}_0)\,p^{(\gamma)}(\boldsymbol{x}_0)\,d\boldsymbol{x}_0\Big\rangle\,d\boldsymbol{x}_t \\ &= \iint \big\langle \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t),\, \nabla_{\boldsymbol{x}_t} p_t^{(\gamma)}(\boldsymbol{x}_t\mid\boldsymbol{x}_0)\big\rangle\,p^{(\gamma)}(\boldsymbol{x}_0)\,d\boldsymbol{x}_0\,d\boldsymbol{x}_t \\ &= \iint p_t^{(\gamma)}(\boldsymbol{x}_t\mid\boldsymbol{x}_0)\,p^{(\gamma)}(\boldsymbol{x}_0)\,\big\langle \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t),\, \nabla_{\boldsymbol{x}_t}\log p_t^{(\gamma)}(\boldsymbol{x}_t\mid\boldsymbol{x}_0)\big\rangle\,d\boldsymbol{x}_0\,d\boldsymbol{x}_t \\ &= a_t(\boldsymbol{\theta}),\end{aligned}$$

where Leibniz's rule applies by smoothness and Gaussian decay of the forward kernel, and Fubini's theorem applies due to the regularity assumption on $\boldsymbol{s}_{\boldsymbol{\theta}}$. Combining the above,

$$\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{x}_t}\big\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t) - \nabla_{\boldsymbol{x}_t}\log p_t^{(\gamma)}(\boldsymbol{x}_t\mid\boldsymbol{x}_0)\big\|^2 - \mathbb{E}_{\boldsymbol{x}_t}\big\|\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t) - \mathbf{s}_t^{(\gamma)}(\boldsymbol{x}_t)\big\|^2 = \mathrm{gap}(t), \tag{25}$$

where

$$\mathrm{gap}(t) := \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{x}_t}\big\|\nabla_{\boldsymbol{x}_t}\log p_t^{(\gamma)}(\boldsymbol{x}_t\mid\boldsymbol{x}_0)\big\|^2 - \mathbb{E}_{\boldsymbol{x}_t}\big\|\mathbf{s}_t^{(\gamma)}(\boldsymbol{x}_t)\big\|^2, \tag{26}$$

which is independent of $\boldsymbol{\theta}$. The Gaussian forward kernel (2) gives

$$\nabla_{\boldsymbol{x}_t}\log p_t^{(\gamma)}(\boldsymbol{x}_t\mid\boldsymbol{x}_0) = -\frac{\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_0}{1-\alpha_t} = -\frac{\mathbf{z}}{\sqrt{1-\alpha_t}}, \quad \mathbf{z}\sim\mathcal{N}(\mathbf{0},\mathbf{I}), \tag{27}$$

hence the first term of (26) equals

$$\mathbb{E}\big\|\nabla_{\boldsymbol{x}_t}\log p_t^{(\gamma)}(\boldsymbol{x}_t\mid\boldsymbol{x}_0)\big\|^2 = \frac{d}{1-\alpha_t}. \tag{28}$$

For the marginal score, take expectations of (27) conditional on $\boldsymbol{x}_t$ we obtain Tweedie's formula:

$$\mathbf{s}_t^{(\gamma)}(\boldsymbol{x}_t) = \mathbb{E}\left[\nabla_{\boldsymbol{x}_t} \log p_t^{(\gamma)}(\boldsymbol{x}_t \mid \boldsymbol{x}_0) \,\middle|\, \boldsymbol{x}_t\right] = \frac{\sqrt{\alpha_t}\,\mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t] - \boldsymbol{x}_t}{1-\alpha_t}. \tag{29}$$

Setting $\hat{\boldsymbol{\eta}}(\boldsymbol{x}_t) := \sqrt{\alpha_t}\,\mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t] = \mathbb{E}[\sqrt{\alpha_t}\boldsymbol{x}_0 \mid \boldsymbol{x}_t]$, we see that,

$$\mathbb{E}\left\|\mathbf{s}_t^{(\gamma)}(\boldsymbol{x}_t)\right\|^2 = \frac{\mathbb{E}\|\boldsymbol{x}_t - \hat{\boldsymbol{\eta}}(\boldsymbol{x}_t)\|^2}{(1-\alpha_t)^2}. \tag{30}$$

Decomposing $\boldsymbol{x}_t - \hat{\boldsymbol{\eta}} = (\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_0) + (\sqrt{\alpha_t}\boldsymbol{x}_0 - \hat{\boldsymbol{\eta}}) = \sqrt{1-\alpha_t}\,\mathbf{z} + \sqrt{\alpha_t}(\boldsymbol{x}_0 - \mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t])$ and expanding the expectation we get,

$$\mathbb{E}\|\boldsymbol{x}_t - \hat{\boldsymbol{\eta}}\|^2 = (1-\alpha_t)\,d + 2\sqrt{\alpha_t(1-\alpha_t)}\,\mathbb{E}\big\langle \mathbf{z},\, \boldsymbol{x}_0 - \mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]\big\rangle + \alpha_t\,\mathrm{MMSE}_t(p^{(\gamma)}). \tag{31}$$

For the cross term, use $\mathbf{z} = (\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_0)/\sqrt{1-\alpha_t}$ and orthogonality of conditional expectation, $\mathbb{E}\left[(\boldsymbol{x}_0 - \mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t])\,\widehat{\boldsymbol{x}}_0(\boldsymbol{x}_t)\right] = 0$ for any g:

$$\begin{aligned}
\mathbb{E}\big\langle \mathbf{z},\, \boldsymbol{x}_0 - \mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]\big\rangle &= \tfrac{1}{\sqrt{1-\alpha_t}}\,\mathbb{E}\big\langle \boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_0,\, \boldsymbol{x}_0 - \mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]\big\rangle \\
&= \tfrac{1}{\sqrt{1-\alpha_t}}\,\widehat{\boldsymbol{x}}_0\big(0 - \sqrt{\alpha_t}\,\mathrm{MMSE}_t(p^{(\gamma)})\big) \\
&= -\frac{\sqrt{\alpha_t}\,\mathrm{MMSE}_t(p^{(\gamma)})}{\sqrt{1-\alpha_t}}.
\end{aligned}$$

Substituting into (31),

$$\mathbb{E}\|\boldsymbol{x}_t - \hat{\boldsymbol{\eta}}\|^2 = (1-\alpha_t)\,d - \alpha_t\,\mathrm{MMSE}_t(p^{(\gamma)}), \tag{32}$$

and therefore via (30),

$$\mathbb{E}\left\|\mathbf{s}_t^{(\gamma)}(\boldsymbol{x}_t)\right\|^2 = \frac{d}{1-\alpha_t} - \frac{\alpha_t\,\mathrm{MMSE}_t(p^{(\gamma)})}{(1-\alpha_t)^2}. \tag{33}$$

Combining (28) and (33) in (26),

$$\mathrm{gap}(t) = \frac{\alpha_t\,\mathrm{MMSE}_t(p^{(\gamma)})}{(1-\alpha_t)^2} \;\geq\; 0. \tag{34}$$

Hence, multiplying (25) by λ_t and integrating over $t \sim \mathrm{Unif}(0,1)$, the difference in the neural network loss and the DSM loss can be written as,

$$\ell_{\mathrm{NN}}(p^{(\boldsymbol{\theta})}, p^{(\gamma)}) - \ell_{\mathrm{DSM}}(p^{(\boldsymbol{\theta})}, p^{(\gamma)}) = \int_0^1 \lambda_t\,\mathrm{gap}(t)\,dt = \int_0^1 \lambda_t\,\frac{\alpha_t\,\mathrm{MMSE}_t(p^{(\gamma)})}{(1-\alpha_t)^2}\,dt =: C(p^{(\gamma)}), \tag{35}$$

proving (18)–(19). For the equality case, note that $C(p^{(\gamma)}) = 0$ iff $\mathrm{MMSE}_t(p^{(\gamma)}) = 0$ for almost every $t \in (0,1)$. The latter requires $\boldsymbol{x}_0 = \mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]$ almost surely, i.e., $\boldsymbol{x}_0$ is $\sigma(\boldsymbol{x}_t)$-measurable. Since $\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\mathbf{z}$ with $\mathbf{z}$ independent of $\boldsymbol{x}_0$ and $1-\alpha_t > 0$, this forces $\boldsymbol{x}_0$ to be deterministic. □

D.2 Proof of Theorem 2

Remark 5. *In this theorem, we use $\lambda_t = 1$ rather than $\lambda_t = 1-\alpha_t$. The value of t where $\ell_{\mathrm{SM}}(p_t^{(\gamma)}, p_t^{(\gamma^*)})$ is large correspond to when α_t is close to 0 for when the modes are reasonably well-separated, so that $1-\alpha_t$ is close to 1. The same analysis can be performed in the presence of any weighting function λ_t if one is willing to accept a more complicated expression for the constant C; we elaborate on this in Proposition 3. For all of our numerical results, we use the convention that $\lambda_t = 1-\alpha_t$ to show that our results still hold for the weighting factor used in practice.*

Throughout the proof, let $\Delta := \|\mu_1 - \mu_0\|$. From Lemma 3 and Lemma 4, we can generate samples from $p_t^{(\gamma)}$ by sampling $c \sim \mathrm{Ber}(\gamma)$ and sampling $\mathbf{y}_t \sim p_t^{(c)}$. Furthermore, the next lemma will show that we can generate these samples using the Probability Flow ODE ((Song et al., 2021c)) instead of the stochastic forward process.

First, note that

$$\ell_{\mathrm{DSM}}(p^{(\hat{\gamma})}, p^{(\gamma^*)}) = \mathbb{E}_[t], \mathbf{x}_t \sim p_t^{(\gamma^*)}] \|\mathbf{s}_t^{(\gamma^*)}(\mathbf{x}_t) - \mathbf{s}_t^{(\hat{\gamma})}\|(\mathbf{x}_t) \tag{36}$$

$$= (1-\gamma^*)\mathbb{E}_[t], \mathbf{x}_t \sim p_t^{(\gamma^*)}|c=0] \|\mathbf{s}_t^{(\gamma^*)}(\mathbf{x}_t) - \mathbf{s}_t^{(\hat{\gamma})}\|(\mathbf{x}_t) \tag{37}$$

$$+ \gamma^* \mathbb{E}_[t], \mathbf{x}_t \sim p_t^{(\gamma^*)}|c=1] \|\mathbf{s}_t^{(\gamma^*)}(\mathbf{x}_t) - \mathbf{s}_t^{(\hat{\gamma})}\|(\mathbf{x}_t) \tag{38}$$

$$= (1-\gamma^*)\mathbb{E}_[t], \mathbf{x}_t \sim p_t^{(0)}] \|\mathbf{s}_t^{(\gamma^*)}(\mathbf{x}_t) - \mathbf{s}_t^{(\hat{\gamma})}\|(\mathbf{x}_t) \tag{39}$$

$$+ \gamma^* \mathbb{E}_[t], \mathbf{x}_t \sim p_t^{(1)}] \|\mathbf{s}_t^{(\gamma^*)}(\mathbf{x}_t) - \mathbf{s}_t^{(\hat{\gamma})}\|(\mathbf{x}_t) \tag{40}$$

In particular, we can lower-bound ℓ_{DSM} by independently lower-bounding each of these expectations.

Lemma 5. *Consider any distribution π and let π_t be the noisy marginal at time t of the forward process. Then, if $\mathbf{y}_t$ satisfies the Probability Flow ODE*

$$\mathrm{d}\mathbf{y}_t = -\frac{1}{2}\beta_t \left(\mathbf{y}_t + \nabla \log p_{\mathbf{y}_t}(\mathbf{y}_t)\right) \mathrm{d}t \tag{41}$$

and $\mathbf{y}_0 \sim \pi$, then $\mathbf{y}_t \sim \pi_t$.

Proof. Let $\boldsymbol{x}_t$ satisfy the Variance Preserving SDE and $\mathbf{y}_t$ be as in the lemma statement. Then, we have the Fokker-Planck equations

$$\begin{aligned}\frac{\partial p_{\boldsymbol{x}_t}}{\partial t} &= -\nabla \cdot \left(-\frac{1}{2}\beta_t \boldsymbol{x} p_{\boldsymbol{x}_t}\right) + \frac{1}{2}\beta_t \Delta p_{\boldsymbol{x}_t} \\ &= \frac{1}{2}\beta_t \nabla \cdot (\boldsymbol{x} p_{\boldsymbol{x}_t}) + \frac{1}{2}\beta_t \Delta p_{\boldsymbol{x}_t}\end{aligned}$$

$$\begin{aligned}\frac{\partial p_{\mathbf{y}_t}}{\partial t} &= \frac{1}{2}\beta_t \nabla \cdot (\mathbf{y} p_{\mathbf{y}_t}) + \frac{1}{2}\beta_t \nabla \cdot (p_{\mathbf{y}_t} \nabla \log p_{\mathbf{y}_t}) \\ &= \frac{1}{2}\beta_t \nabla \cdot (\mathbf{y} p_{\mathbf{y}_t}) + \frac{1}{2}\beta_t \Delta p_{\mathbf{y}_t}\end{aligned}$$

□

The Probability Flow ODE ((41)) substantially simplifies the analysis because of the following fact:

Lemma 6. *For $\mathbf{y}_0$, and $\mathbf{y}_t$ following the Probability Flow ODE associated with c and starting from $\mathbf{y}_0 \sim p^{(c)}$, we have that*

$$\mathbf{y}_t - \sqrt{\alpha_t}\mu_c = \mathbf{y}_0 - \mu_c$$

Proof. Note that $\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{\alpha_t} = \frac{1}{2}\beta_t\sqrt{\alpha_t}$. Therefore,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{y}_t - \sqrt{\alpha_t}\mu_c) &= -\frac{1}{2}\beta_t \left(\mathbf{y}_t + \nabla \log p_t^{(c)}(\mathbf{y}_t)\right) - \frac{1}{2}\beta_t\sqrt{\alpha_t}\mu_c \\ &= -\frac{1}{2}\beta_t \left(\mathbf{y}_t + \sqrt{\alpha_t}\mu_c - \mathbf{y}_t - \sqrt{\alpha_t}\mu_c\right) \\ &= 0\end{aligned}$$

□

We now focus on the case where $c = 0$ (the case of $c = 1$ will follow via symmetry). Let $\mathbf{z}_t = \mathbf{y}_t|c = 0$. Our first goal is to lower bound $\|\mathbf{s}_t^{(\gamma^*)}(\mathbf{z}_t) - \mathbf{s}_t^{(\hat{\gamma})}(\mathbf{z}_t)\|^2$ pointwise in t and $\mathbf{z}_t$. We first show an analytic expression for this score difference, then identify a range of t's along each path where this expression is sufficiently large. Finally, by integrating over initial conditions $\mathbf{z}_0$, we get a lower bound for $\mathbb{E}_[t], \mathbf{z}_t \sim p_t^{(0)}] \|\mathbf{s}_t^{(\gamma^*)}(\mathbf{z}_t) - \mathbf{s}_t^{(\hat{\gamma})}(\mathbf{z}_t)\|^2$.

Proposition 1. *Let $\hat{\gamma}, \gamma^* \in [0,1]$. Then,*

$$\|\mathbf{s}_t^{(\hat{\gamma})}(\mathbf{z}_t) - \mathbf{s}_t^{(\gamma^*)}(\mathbf{z}_t)\|^2 = (\hat{\gamma}-\gamma^*)^2 \frac{\|\mu_0-\mu_1\|^2 \alpha_t e^{2f_0(\mathbf{z}_0,t)}}{(1-\hat{\gamma}+\hat{\gamma}e^{f_0(\mathbf{z}_0,t)})^2(1-\gamma^*+\gamma^* e^{f_0(\mathbf{z}_0,t)})^2}$$

where

$$f_0(\mathbf{z}_0,t) := -\frac{1}{2}\|\mu_1-\mu_0\|^2\alpha_t + \langle \mathbf{z}_0-\mu_0, \mu_1-\mu_0\rangle \sqrt{\alpha_t}$$

Proof. We

$$p_t^{(0)}(\boldsymbol{x}) = \frac{1}{Z_t}\exp\left(-\frac{1}{2}\|\boldsymbol{x}-\sqrt{\alpha_t}\mu_0\|^2\right)$$

$$p_t^{(1)}(\boldsymbol{x}) = \frac{1}{Z_t}\exp\left(-\frac{1}{2}\|\boldsymbol{x}-\sqrt{\alpha_t}\mu_1\|^2\right)$$

Expanding the log likelihoods, we have

$$\log\frac{p_1(\mathbf{z}_t)}{p_0(\mathbf{z}_t)} = -\frac{1}{2}\|\mu_0-\mu_1\|^2\alpha_t + \langle \mathbf{z}_0-\mu_0, \mu_1-\mu_0\rangle\sqrt{\alpha_t}$$

Let $f(\mathbf{z}_0,t) := \log\frac{p_1(\mathbf{z}_t)}{p_0(\mathbf{z}_t)}$. Then,

$$\begin{aligned}
\mathbf{s}_t^{(\gamma)}(\mathbf{z}_t) &= \frac{(1-\gamma)\,p_t^{(0)}(\mathbf{z}_t)\mathbf{s}_t^{(0)}(\mathbf{z}_t) + \gamma\, p_t^{(1)}(\mathbf{z}_t)\mathbf{s}_t^{(1)}(\mathbf{z}_t)}{p_t^{(\gamma)}(\mathbf{z}_t)} \\
&= \mathbf{s}_t^{(0)}(\mathbf{z}_t) + \frac{\gamma\, p_t^{(1)}(\mathbf{z}_t)\left(\mathbf{s}_t^{(1)}(\mathbf{z}_t)-\mathbf{s}_t^{(0)}(\mathbf{z}_t)\right)}{p_t^{(\gamma)}(\mathbf{z}_t)} \\
&= \mathbf{s}_t^{(0)}(\mathbf{z}_t) + \frac{\gamma\, p_t^{(1)}(\mathbf{z}_t)}{p_t^{(\gamma)}(\mathbf{z}_t)}\sqrt{\alpha_t}\,(\mu_1-\mu_0) \\
&= \mathbf{s}_t^{(0)}(\mathbf{z}_t) + \frac{\gamma\, e^{f(\mathbf{z}_0,t)}}{1-\gamma+\gamma\, e^{f(\mathbf{z}_0,t)}}\sqrt{\alpha_t}\,(\mu_1-\mu_0)
\end{aligned}$$

where we used $\nabla \log p_t^{(c)}(\boldsymbol{x}) = \sqrt{\alpha_t}\mu_c - \boldsymbol{x}$. Therefore,

$$\mathbf{s}_t^{(\hat{\gamma})}(\mathbf{z}_t) - \mathbf{s}_t^{(\gamma^*)}(\mathbf{z}_t) = \frac{(\hat{\gamma}-\gamma^*)\, e^{f(\mathbf{z}_0,t)}}{\left(1-\hat{\gamma}+\hat{\gamma}\, e^{f(\mathbf{z}_0,t)}\right)\left(1-\gamma^*+\gamma^*\, e^{f(\mathbf{z}_0,t)}\right)}\sqrt{\alpha_t}\,(\mu_1-\mu_0)$$

This gives

$$\|\mathbf{s}_t^{(\hat{\gamma})}(\mathbf{z}_t) - \mathbf{s}_t^{(\gamma^*)}(\mathbf{z}_t)\|^2 = (\hat{\gamma}-\gamma^*)^2 \frac{\|\mu_1-\mu_0\|^2\alpha_t\, e^{2f(\mathbf{z}_0,t)}}{\left(1-\hat{\gamma}+\hat{\gamma}\, e^{f(\mathbf{z}_0,t)}\right)^2\left(1-\gamma^*+\gamma^*\, e^{f(\mathbf{z}_0,t)}\right)^2}$$

□

Lemma 7. *For all $x \in \mathbb{R}$ and $\gamma \in [0,1]$, we have*

$$\frac{e^x}{(1-\gamma+\gamma e^x)^2} \geq \frac{1}{4\gamma(1-\gamma)} e^{-\frac{1}{4}\left(x-\log\frac{1-\gamma}{\gamma}\right)^2}$$

with equality only when $x = \log\frac{1-\gamma}{\gamma}$. In particular,

$$\frac{e^{2f_0(\mathbf{z}_0,t)}}{\left(1-\hat{\gamma}+\hat{\gamma}e^{f_0(\mathbf{z}_0,t)}\right)^2\left(1-\gamma^*+\gamma^* e^{f_0(\mathbf{z}_0,t)}\right)^2} \geq \frac{e^{-\frac{1}{4}\left(f_0(\mathbf{z}_0,t)-\log\frac{1-\hat{\gamma}}{\hat{\gamma}}\right)^2-\frac{1}{4}\left(f_0(\mathbf{z}_0,t)-\log\frac{1-\gamma^*}{\gamma^*}\right)^2}}{16\hat{\gamma}(1-\hat{\gamma})\gamma^*(1-\gamma^*)}$$

Proof. Let $g(x) = \frac{e^x}{(1-\gamma+\gamma e^x)^2}$ and $h(x) = \frac{1}{4\gamma(1-\gamma)} e^{-\frac{1}{4}\left(x - \log\frac{1-\gamma}{\gamma}\right)^2}$. Then,

$$\frac{\mathrm{d}}{\mathrm{d}x}\log g(x) = \frac{1-\gamma-\gamma e^x}{1-\gamma+\gamma e^x}$$

For all $\gamma \in [0,1]$ $1-\gamma+\gamma e^x > 0$. Therefore, $x^* := \log\frac{1-\gamma}{\gamma}$ is the only critical point of $g(x)$, and $g(x^*) = \frac{1}{4\gamma(1-\gamma)}$. Similarly $h(x^*) = \frac{1}{4\gamma(1-\gamma)}$ and $\frac{\mathrm{d}}{\mathrm{d}x}\log h(x^*) = 0$. We have

$$\frac{\mathrm{d}}{\mathrm{d}x}[\log h(x) - \log g(x)] = -\frac{1}{2}(x-x^*) - \frac{1-e^{x-x^*}}{1+e^{x-x^*}}$$
$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}[\log h(x) - \log g(x)] = -\frac{(1-e^{x-x^*})^2}{2(1+e^{x-x^*})^2}$$

So, $\frac{\mathrm{d}^2}{\mathrm{d}x^2}[\log h(x) - \log g(x)] \le 0$ with equality only when $x = x^*$. Therefore, $\log h - \log g$ is strictly concave away from x^* and maximized at x^*, where $h(x^*) = g(x^*)$. Therefore, $h < g$. □

Lemma 8. *Let $\mathbf{z}_t$ and $\mathbf{z}_0$ be as above. Let $M(\mathbf{z}_0) := \langle \mathbf{z}_0 - \mu_0, \mu_1 - \mu_0\rangle$, and define the log-odds ratios*

$$\hat{r} := \log\frac{1-\hat{\gamma}}{\hat{\gamma}}$$
$$r^* := \log\frac{1-\gamma^*}{\gamma^*}$$

Let $\epsilon > 0$ be such that

$$\epsilon \le \exp\left(-\frac{1}{8}\left[(\hat{r}-r^*)^2 + \left(\hat{r}+r^* - \frac{M(\mathbf{z}_0)^2}{\Delta^2}\right)_+^2\right]\right)$$

where $(\cdot)_+ := \max\{\cdot, 0\}$.

Let

$$a = \max\left\{\frac{M(\mathbf{z}_0)}{\Delta^2} + \frac{1}{\Delta}\sqrt{\left(\frac{M(\mathbf{z}_0)^2}{\Delta^2} - (\hat{r}+r^*) - \sqrt{8\log\frac{1}{\epsilon} - (\hat{r}-r^*)^2}\right)_+}, \sqrt{\alpha_1}\right\}$$
$$b = \min\left\{\frac{M(\mathbf{z}_0)}{\Delta^2} + \frac{1}{\Delta}\sqrt{\frac{M(\mathbf{z}_0)^2}{\Delta^2} - (\hat{r}+r^*) + \sqrt{8\log\frac{1}{\epsilon} - (\hat{r}-r^*)^2}}, 1\right\}$$

Then for all t such that $\sqrt{\alpha_t} \in [a,b]$,

$$e^{-\frac{1}{4}(f_0(\mathbf{z}_0,t)-\hat{r})^s 2 - \frac{1}{4}(f_0(\mathbf{z}_0,t)-r^*)^2} \ge \epsilon$$

Proof. First, the inequality is equivalent to

$$(f_0(\mathbf{z}_0,t)-r^*)^2 + (f_0(z_0,t)-r^*)^2 \le 4\log\frac{1}{\epsilon}$$

Plugging in the definition of f, the inequality holds whenever

$$\left[\frac{\Delta^2}{2}\left(\sqrt{\alpha_t} - \frac{M(\mathbf{z}_0)}{\Delta^2}\right)^2 - \frac{M(\mathbf{z}_0)^2}{2\Delta^2} + \frac{\hat{r}+r^*}{2}\right]^2 \le 2\log\frac{1}{\epsilon} - \frac{(\hat{r}-r^*)^2}{4}$$

We wish to identify the values of $\sqrt{\alpha_t}$ which make this inequality hold. This function is continuous, so its sign can only change at the roots of the corresponding equality constraint:

$$\left[\frac{\Delta^2}{2}\left(\sqrt{\alpha_t} - \frac{M(\mathbf{z}_0)}{\Delta^2}\right)^2 - \frac{M(\mathbf{z}_0)^2}{2\Delta^2} + \frac{\hat{r}+r^*}{2}\right]^2 = 2\log\frac{1}{\epsilon} - \frac{(\hat{r}-r^*)^2}{4}$$

Solving this quartic gives the roots

$$\sqrt{\alpha_t} = \frac{M(\mathbf{z}_0)}{\Delta^2} \pm \sqrt{\frac{M(\mathbf{z}_0)^2}{\Delta^4} - \frac{\hat{r}+r^*}{\Delta^2} \pm \frac{1}{\Delta^2}\sqrt{8\log\frac{1}{\epsilon} - (\hat{r}-r^*)^2}}$$

The condition on ϵ is such that the square roots are all well-defined. Let b' be the largest of these roots (both $\pm$'s are set to $+$). As $\sqrt{\alpha_t} \to \infty$, we would have $f_0(\mathbf{z}_0, t) \to -\infty$ (because $-\Delta^2/2$ is strictly negative), and hence

$$e^{-\frac{1}{4}(f_0(\mathbf{z}_0,t)-\hat{r})^2 - \frac{1}{4}(f_0(\mathbf{z}_0,t)-r^*)^2} \to 0$$

Therefore, for $\sqrt{\alpha_t} > b'$ the inequality does not hold. Note that the four roots above are distinct (with two being complex for ϵ sufficiently small), meaning that each root has multiplicity 1, and hence the polynomial changes signs at b'. Therefore, for $\sqrt{\alpha_t}$ between b' and the next largest real root, the inequality holds. The next largest real root is

$$a' = \begin{cases} \frac{M(\mathbf{z}_0)}{\Delta^2} + \sqrt{\frac{M(\mathbf{z}_0)^2}{\Delta^4} - \frac{\hat{r}+r^*}{\Delta^2} - \frac{1}{\Delta^2}\sqrt{8\log\frac{1}{\epsilon} - (\hat{r}-r^*)^2}} & \text{if } \frac{M(\mathbf{z}_0)^2}{\Delta^2} - (\hat{r}+r^*) - \sqrt{8\log\frac{1}{\epsilon} - (\hat{r}-r^*)^2} \geq 0 \\ \frac{M(\mathbf{z}_0)}{\Delta^2} - \sqrt{\frac{M(\mathbf{z}_0)^2}{\Delta^4} - \frac{\hat{r}+r^*}{\Delta^2} + \frac{1}{\Delta^2}\sqrt{8\log\frac{1}{\epsilon} - (\hat{r}-r^*)^2}} & \text{otherwise} \end{cases}$$

In the second case we have $a' \leq \frac{M(\mathbf{z}_0)}{\Delta^2}$, so

$$a' \leq \frac{M(\mathbf{z}_0)}{\Delta^2} + \sqrt{\left(\frac{M(\mathbf{z}_0)^2}{\Delta^4} - \frac{\hat{r}+r^*}{\Delta^2} - \frac{1}{\Delta^2}\sqrt{8\log\frac{1}{\epsilon} - (\hat{r}-r^*)^2}\right)_+} =: a''$$

Therefore, the inequality holds on $[a', b']$ and $[a'', b'] \subseteq [a', b']$, so the inequality holds on $[a'', b']$. Note that for $t \in [0,1]$, $\alpha_t \in [\alpha_1, 1]$ where $\alpha_1 > 0$ is small, so we let $b := \min\{b', 1\}$ and $a := \max\{a, \sqrt{\alpha_1}\}$. □

Proposition 2. *For all ϵ satisfying the condition in Lemma 8, for all t s.t. $\sqrt{\alpha_t} \in [a, b]$ (with a and b as defined in the previous lemma) we have*

$$\|\mathbf{s}_t^{(\hat{\gamma})}(\mathbf{z}_t) - \mathbf{s}_t^{(\gamma^*)}(\mathbf{z}_t)\|^2 \geq \frac{(\hat{\gamma}-\gamma^*)^2\|\mu_1-\mu_0\|^2\alpha_t}{16\hat{\gamma}(1-\hat{\gamma})\gamma^*(1-\gamma^*)}\epsilon$$

Proof. Combining the above two lemmas,

$$\begin{aligned}\|\mathbf{s}_t^{(\hat{\gamma})}(\mathbf{z}_t) - \mathbf{s}_t^{(\gamma^*)}(\mathbf{z}_t)\|^2 &\geq \frac{(\hat{\gamma}-\gamma^*)^2\|\mu_1-\mu_0\|^2\alpha_t}{16\hat{\gamma}(1-\hat{\gamma})\gamma^*(1-\gamma^*)} e^{-\frac{1}{4}(f_0(\mathbf{z}_0,t)-\hat{r})^2 - \frac{1}{4}(f_0(\mathbf{z}_0,t)-r^*)^2} \\ &\geq \frac{(\hat{\gamma}-\gamma^*)^2\|\mu_1-\mu_0\|^2\alpha_t}{16\hat{\gamma}(1-\hat{\gamma})\gamma^*(1-\gamma^*)}\epsilon\end{aligned}$$

□

We now turn our attention to lower bounding

$$\int_0^1 \|\mathbf{s}_t^{(\hat{\gamma})}(\mathbf{z}_t) - \mathbf{s}_t^{(\gamma^*)}(\mathbf{z}_t)\|^2 \mathrm{d}t$$

along a given path starting from $\mathbf{z_0}$. From the lemmas above, we have

$$\int_0^1 \|\nabla \log p_t^{(\hat{\gamma})}(\mathbf{z}_t) - \nabla \log p_t^{(\gamma^*)}(\mathbf{z}_t)\|^2 \mathrm{d}t > \frac{(\hat{\gamma}-\gamma^*)^2\|\mu_1-\mu_0\|^2}{16\hat{\gamma}(1-\hat{\gamma})\gamma^*(1-\gamma^*)}\epsilon \int_{t_{b^2}}^{t_{a^2}} \alpha_t \mathrm{d}t$$

where t_{a^2} and t_{b^2} are such that $\alpha_{t_{a^2}} = a^2$ and $\alpha_{t_{b^2}} = b^2$. Note α_t is a decreasing function of t, hence why the bounds of the integral are reversed. If we let $\alpha^{-1}(\tau)$ be the inverse function of α_t, such that $\alpha_{\alpha^{-1}(\tau)} = \tau$, then

$$\int_{t_{b^2}}^{t_{a^2}} \alpha_t \mathrm{d}t = 2\int_a^b \frac{\tau}{\beta_{\alpha^{-1}(\tau^2)}} \mathrm{d}\tau$$

In particular, if we consider the linear noise schedule

$$\beta_t := \beta_1 t$$

then

$$\int_{t_{b^2}}^{t_{a^2}} \alpha_t \mathrm{d}t = \frac{1}{\sqrt{\beta_1}} \int_a^b \frac{\tau}{\sqrt{-\ln \tau}} \mathrm{d}\tau$$

Proposition 3. *Let a, b, Δ, z_0, $\hat{r}$, r^*, and M, be as above, and assume ϵ satisfies the bound from Lemma 8. Define*

$$R_0 := \sqrt{8 \ln \frac{1}{\epsilon} - (\hat{r} - r^*)^2} - (\hat{r} + r^*)$$

Then, when $\frac{M(\mathbf{z}_0)}{\Delta^2} \geq \sqrt{\alpha_1}$ and

$$\epsilon \geq \exp\left(-\frac{1}{8}\left[(\Delta^2 - 2M(\mathbf{z}_0) + (\hat{r} + r^*))^2 + (\hat{r} - r^*)^2\right]\right)$$

then

$$\int_a^b \frac{\tau}{\sqrt{-\ln \tau}} \mathrm{d}\tau \geq \frac{\frac{M(\mathbf{z}_0)}{\Delta}\sqrt{\frac{M(\mathbf{z}_0)^2}{\Delta^2} + R_0} + \frac{M(\mathbf{z}_0)^2}{2\Delta^2} + \frac{1}{2}R_0}{\Delta^2 \sqrt{\ln(\Delta) - \ln\left(\frac{M(\mathbf{z}_0)}{\Delta} + \frac{1}{2}\sqrt{\frac{M(\mathbf{z}_0)^2}{\Delta^2} + R_0}\right)}}$$

Remark 6. *This proposition requires $\lambda_t = 1$, as the analysis is far more challenging when the integrand g is nonconvex. If one wishes to analyze alternative weighting factors or, indeed, an alternative noise schedule, all the analysis up to this point applies. The difference is that one now needs to lower bound the more challenging expression $\int_{t_{b^2}}^{t_{a^2}} \alpha_t \lambda_t \mathrm{d}t$. For the linear noise schedule and $\lambda_t = 1 - \alpha_t$, this take the form $\frac{1}{\sqrt{\beta_1}} \int_a^b \frac{\tau(1-\tau^2)}{\sqrt{-\ln \tau}} \mathrm{d}\tau$.*

Proof. Let

$$g(t) := \frac{\tau}{\sqrt{-\ln \tau}}$$

Note that

$$g''(t) = \frac{1}{2t(-\ln t)^{3/2}} + \frac{3}{4t(-\ln t)^{5/2}} > 0$$

So, $g(t)$ is convex, meaning

$$\int_a^b g(\tau)\mathrm{d}\tau \geq (b-a) g\left(\frac{a+b}{2}\right)$$

The condition that $\frac{M(\mathbf{z}_0)}{\Delta^2} \geq \sqrt{\alpha_1}$ guarantees $a \geq \sqrt{\alpha_1}$. Similarly, the condition on ϵ guarantees $b \leq 1$, so

$$\begin{aligned} b - a &= \frac{1}{\Delta}\sqrt{\frac{M(\mathbf{z}_0)^2}{\Delta^2} - (\hat{r} + r^*) + \sqrt{8 \log \frac{1}{\epsilon} - (\hat{r} - r^*)^2}} \\ &= \frac{1}{\Delta}\sqrt{\frac{M(\mathbf{z}_0)^2}{\Delta^2} + R_0} \end{aligned}$$

We have

$$\frac{a+b}{2} = \frac{M(\mathbf{z}_0)}{\Delta^2} + \frac{1}{2\Delta}\sqrt{\frac{M(\mathbf{z}_0)^2}{\Delta^2} + R_0}$$

and so

$$\implies g\left(\frac{a+b}{2}\right) = \frac{\frac{M(\mathbf{z}_0)}{\Delta} + \frac{1}{2}\sqrt{\frac{M(\mathbf{z}_0)^2}{\Delta^2} + R_0}}{\Delta\sqrt{\ln(\Delta) - \ln\left(\frac{M(\mathbf{z}_0)}{\Delta} + \frac{1}{2}\sqrt{\frac{M(\mathbf{z}_0)^2}{\Delta^2} + R_0}\right)}}$$

Combining these,

$$\int_a^b g(\tau)\mathrm{d}\tau \geq \frac{\frac{M(\mathbf{z}_0)}{\Delta}\sqrt{\frac{M(\mathbf{z}_0)^2}{\Delta^2} + R_0} + \frac{M(\mathbf{z}_0)^2}{2\Delta^2} + \frac{1}{2}R_0}{\Delta^2\sqrt{\ln(\Delta) - \ln\left(\frac{M(\mathbf{z}_0)}{\Delta} + \frac{1}{2}\sqrt{\frac{M(\mathbf{z}_0)^2}{\Delta^2} + R_0}\right)}}$$

□

Combining this lower bound with what we showed before, we have

$$\int_0^1 \|\nabla \log p_t^{(\hat{\gamma})}(\mathbf{z}_t) - \nabla \log p_t^{(\gamma^*)}(\mathbf{z}_t)\|^2 \mathrm{d}t$$
$$\geq \frac{(\hat{\gamma}-\gamma^*)^2\epsilon}{16\sqrt{\beta_1}\,\hat{\gamma}(1-\hat{\gamma})\gamma^*(1-\gamma^*)} \frac{\frac{M(\mathbf{z}_0)}{\Delta}\sqrt{\frac{M(\mathbf{z}_0)^2}{\Delta^2} + R_0} + \frac{M(\mathbf{z}_0)^2}{2\Delta^2} + \frac{1}{2}R_0}{\sqrt{\ln(\Delta) - \ln\left(\frac{M(\mathbf{z}_0)}{\Delta} + \frac{1}{2}\sqrt{\frac{M(\mathbf{z}_0)^2}{\Delta^2} + R_0}\right)}}$$

Recall that $\mathbf{z}_0 \sim \mathcal{N}(\mu_0, \mathbf{I})$, so

$$M(\mathbf{z}_0) = \langle \mathbf{z}_0 - \mu_0, \mu_1 - \mu_0 \rangle \sim \mathcal{N}(0, \|\mu_1 - \mu_0\|^2)$$
$$\implies \frac{M(\mathbf{z}_0)}{\Delta} \sim \mathcal{N}(0,1)$$

Proposition 4. *Let Φ be the CDF of $\mathcal{N}(0,1)$. Let $\tau \geq \Delta\sqrt{\alpha_1}$, and let $\epsilon > 0$ satisfy $\epsilon \in (\epsilon^0_{\min}, \epsilon^0_{\max})$ with*

$$\epsilon^0_{\max} = \exp\left(-\frac{1}{8}\left[(\hat{r}-r^*)^2 + \left(\hat{r}+r^*-\tau^2\right)_+^2\right]\right)$$
$$\epsilon^0_{\min} = \exp\left(-\frac{1}{8}\left[(\Delta(\Delta-2\tau) + (\hat{r}+r^*))^2 + (\hat{r}-r^*)^2\right]\right)$$

Then, with probability $1-\Phi(\tau)$ over samples of $\mathbf{z}_0$,

$$\int_0^1 \|\nabla \log p_t^{(\hat{\gamma})}(\mathbf{z}_t) - \nabla \log p_t^{(\gamma^*)}(\mathbf{z}_t)\|^2 \mathrm{d}t$$
$$\geq \frac{(\hat{\gamma}-\gamma^*)^2\epsilon}{16\sqrt{\beta_1}\,\hat{\gamma}(1-\hat{\gamma})\gamma^*(1-\gamma^*)} \frac{\tau\sqrt{\tau^2+R_0} + \frac{1}{2}\tau^2 + \frac{1}{2}R_0}{\sqrt{\ln(\Delta) - \ln\left(\tau + \frac{1}{2}\sqrt{\tau^2+R_0}\right)}}$$

Proof. If we let $K := \frac{M(\mathbf{z}_0)}{\Delta}$ our bound becomes

$$\int_0^1 \|\nabla \log p_t^{(\hat{\gamma})}(\mathbf{z}_t) - \nabla \log p_t^{(\gamma^*)}(\mathbf{z}_t)\|^2 \mathrm{d}t$$
$$\geq \frac{(\hat{\gamma}-\gamma^*)^2\epsilon}{16\hat{\gamma}(1-\hat{\gamma})\gamma^*(1-\gamma^*)} \frac{K\sqrt{K^2+R_0} + \frac{1}{2}K^2 + \frac{1}{2}R_0}{\sqrt{\ln(\Delta) - \ln\left(K + \frac{1}{2}\sqrt{K^2+R_0}\right)}}$$

Assume that $K \geq \tau$, which occurs with probability $1-\Phi(\tau)$. Our bound is an increasing function in K and is thus minimized at $K = \tau$. Similarly, our upper bound from before on ϵ is a decreasing function of K, and our lower bound from before is an increasing function of ϵ, so our bounds on ϵ are tightest when we plug in τ, giving $\epsilon^0_{\max}$ and $\epsilon^0_{\min}$ from the statement of the proposition. □

We now state the corresponding proposition for the case where $c = 1$. Let $\mathbf{w}_t := \mathbf{y}_t | c = 1$.

Proposition 5. *Let $\tau \geq \Delta\sqrt{\alpha_1}$, and let $\epsilon > 0$ satisfy $\epsilon \in (\epsilon^1_{\min}, \epsilon^1_{\max})$ with*

$$\epsilon^1_{\max} = \exp\left(-\frac{1}{8}\left[(\hat{r} - r^*)^2 + \left(-(\hat{r} + r^*) - \tau^2\right)_+^2\right]\right)$$
$$\epsilon^1_{\min} = \exp\left(-\frac{1}{8}\left[(\Delta(\Delta - 2\tau) + (\hat{r} + r^*))^2 - (\hat{r} - r^*)^2\right]\right)$$

Then, with probability $1 - \Phi(\tau)$ over samples of $\mathbf{w}_0$,

$$\int_0^1 \|\nabla \log p_t^{(\hat{\gamma})}(\mathbf{w}_t) - \nabla \log p_t^{(\gamma^*)}(\mathbf{w}_t)\|^2 \mathrm{d}t$$
$$\geq \frac{(\hat{\gamma} - \gamma^*)^2 \epsilon}{16\sqrt{\beta_1}\,\hat{\gamma}(1-\hat{\gamma})\gamma^*(1-\gamma^*)} \frac{\tau\sqrt{\tau^2 + R_1} + \frac{1}{2}\tau^2 + \frac{1}{2}R_1}{\sqrt{\ln(\Delta) - \ln\left(\tau + \frac{1}{2}\sqrt{\tau^2 + R_1}\right)}}$$

where

$$R_1 := \sqrt{8\ln\frac{1}{\epsilon} - (\hat{r} - r^*)^2} + (\hat{r} + r^*)$$

Proof. This follows from repeating all of our analysis above for the $c = 0$ case with γ^* and $\hat{\gamma}$ replaced with $1 - \gamma^*$ and $1 - \hat{\gamma}$, respectively. □

Combining Proposition 4 and Proposition 5, we have by our decomposition of ℓ_{DSM} from the start of this section that

$$\ell_{\mathrm{DSM}}(p^{(\hat{\gamma})}, p^{(\gamma^*)}) \geq C(\hat{\gamma} - \gamma^*)^2$$

where

$$C = \frac{\epsilon(1 - \Phi(\tau))}{16\sqrt{\beta_1}\,\hat{\gamma}(1-\hat{\gamma})\gamma^*(1-\gamma^*)}\left[(1-\gamma^*)\frac{\tau\sqrt{\tau^2 + R_0} + \frac{1}{2}\tau^2 + \frac{1}{2}R_0}{\sqrt{\ln(\Delta) - \ln\left(\tau + \frac{1}{2}\sqrt{\tau^2 + R_0}\right)}} + \gamma^*\frac{\tau\sqrt{\tau^2 + R_1} + \frac{1}{2}\tau^2 + \frac{1}{2}R_1}{\sqrt{\ln(\Delta) - \ln\left(\tau + \frac{1}{2}\sqrt{\tau^2 + R_1}\right)}}\right]$$

for all $\tau \geq \Delta\sqrt{\alpha_1}$ and $\epsilon \in [\epsilon_{\min}, \epsilon_{\max}]$, where

$$\epsilon_{\max} = \min\{\epsilon^0_{\max}, \epsilon^1_{\max}\} = \exp\left(-\frac{1}{8}\left[(\hat{r} - r^*)^2 + \max\{\left(\hat{r} + r^* - \tau^2\right)_+^2, \left(-(\hat{r} + r^*) - \tau^2\right)_+^2\}\right]\right)$$
$$\epsilon_{\min} = \max\{\epsilon^0_{\min}, \epsilon^1_{\min}\} = \exp\left(-\frac{1}{8}\left[\min\{(\Delta(\Delta - 2\tau) + (\hat{r} + r^*))^2, (\Delta(\Delta - 2\tau) - (\hat{r} + r^*))^2\} + (\hat{r} - r^*)^2\right]\right)$$

To simplify this expression, note that if

$$g(x) := \frac{x(\tau + \frac{1}{2}x)}{\sqrt{\ln(\Delta) - \ln(\tau + \frac{1}{2}x)}}$$

then direct calculation shows that if $h(x) := \ln(\Delta) - \ln(\tau + \frac{1}{2}x)$,

$$g'(x) = \frac{\frac{x}{4} + (\tau + x)h(x)}{h(x)^{3/2}} > 0$$

Therefore, g is increasing, meaning

$$C = \frac{\epsilon(1 - \Phi(\tau))}{16\sqrt{\beta_1}\,\hat{\gamma}(1-\hat{\gamma})\gamma^*(1-\gamma^*)}\left[(1-\gamma^*)g\left(\sqrt{\tau^2 + R_0}\right) + \gamma^* g\left(\sqrt{\tau^2 + R_1}\right)\right]$$
$$\geq \frac{\epsilon(1 - \Phi(\tau))}{16\sqrt{\beta_1}\,\hat{\gamma}(1-\hat{\gamma})\gamma^*(1-\gamma^*)} g\left(\min\{\sqrt{\tau^2 + R_0}, \sqrt{\tau^2 + R_1}\}\right)$$
$$= \frac{\epsilon(1 - \Phi(\tau))}{16\sqrt{\beta_1}\,\hat{\gamma}(1-\hat{\gamma})\gamma^*(1-\gamma^*)} g\left(\sqrt{\tau^2 + \sqrt{8\ln\frac{1}{\epsilon} - (\hat{r} - r^*)^2} - |\hat{r} - r^*|}\right)$$

Let $R_{\min} := \sqrt{8 \ln \frac{1}{\epsilon} - (\hat{r} - r^*)^2} - |\hat{r} - r^*|$. To obtain the bound from the body of the paper, we now specify values for τ and ϵ which simplify the expression. First, we will set τ to be its minimum value: $\tau = \Delta\sqrt{\alpha_1}$. We choose this because numerically, it seems to be the value of τ that maximizes the expression. This gives us

$$C \geq \frac{\epsilon(1 - \Phi(\Delta\sqrt{\alpha_1}))}{16\sqrt{\beta_1}\,\hat{\gamma}(1-\hat{\gamma})\gamma^*(1-\gamma^*)} \frac{\sqrt{\Delta^2\alpha_1 + R_{\min}}\left(\Delta\sqrt{\alpha_1} + \frac{1}{2}\sqrt{\Delta^2\alpha_1 + R_{\min}}\right)}{\sqrt{\log(\Delta) - \log\left(\sqrt{\alpha_1}\right) - \log\left(\Delta + \frac{1}{2}\sqrt{\Delta^2 + \frac{R_0}{\alpha_1}}\right)}}$$

Noting that $\log(\Delta) < \log\left(\Delta + \frac{1}{2}\sqrt{\Delta^2 + \frac{R_0}{\alpha_1}}\right)$, the denominator of the second term is less than $\sqrt{-\log(\sqrt{\alpha_1})} = \frac{1}{4}\beta_1$ (recall $\alpha_t = \exp\left(-\int_0^t \beta_s \mathrm{d}s\right) = \exp\left(-\frac{1}{2}\beta_1 t^2\right)$). Additionally, note that the numerator of the second term is an increasing function of Δ, meaning we can lower-bound it by setting $\Delta = 0$. Combining these gives us

$$C \geq \frac{1 - \Phi(\Delta\sqrt{\alpha_1})}{8\beta_1\,\hat{\gamma}(1-\hat{\gamma})\gamma^*(1-\gamma^*)} \epsilon\, R_{\min}$$

Finally, direct calculation shows that $\epsilon\, R_{\min}$ is maximized when $\epsilon = \exp\left(-\frac{1}{8}\left((\hat{r} - r^*)^2 + \delta^2\right)\right)$ where $\delta = \frac{|\hat{r}-r^*| + \sqrt{(\hat{r}-r^*)^2+16}}{2}$, giving

$$\epsilon\, R_{\min} = \frac{\sqrt{(\hat{r} - r^*)^2 + 16} - |\hat{r} - r^*|}{2} \exp\left(-\frac{1}{8}\left((\hat{r} - r^*)^2 + \delta^2\right)\right)$$

Note $\delta \leq |\hat{r} - r^*| + 2$ (because $\sqrt{a^2 + b^2} \leq |a| + |b|$), and $\delta^2 \leq |\hat{r} - r^*|^2 + 2|\hat{r} - r^*| + 4$. Applying this, and the fact that $\hat{\gamma}(1-\hat{\gamma}) \leq \frac{1}{4}$

$$\begin{aligned} C &\geq \frac{\left(1 - \Phi(\Delta\sqrt{\alpha_1})\right)\left(\sqrt{(\hat{r} - r^*)^2 + 16} - |\hat{r} - r^*|\right)}{4\beta_1\,\gamma^*(1-\gamma^*)} \exp\left(-\frac{1}{4}\left((\hat{r} - r^*)^2 + |\hat{r} - r^*|\right) - \frac{1}{2}\right) \\ &= \frac{\left(1 - \Phi(\Delta\sqrt{\alpha_1})\right)\left(\sqrt{(\hat{r} - r^*)^2 + 16} - |\hat{r} - r^*|\right)}{4\sqrt{e}\beta_1\,\gamma^*(1-\gamma^*)} \exp\left(-\frac{1}{4}\left((\hat{r} - r^*)^2 + |\hat{r} - r^*|\right)\right) \end{aligned}$$

Finally, for $\hat{\gamma} \in [a, 1-a]$, $|\hat{r} - r^*| \leq \log\frac{1-a}{a} + |\log\frac{1-\gamma^*}{\gamma^*}|$. Our expression is a decreasing function of $|\hat{r} - r^*|$, and thus for all $\hat{\gamma} \in [a, 1-a]$, if we let $k := \log\frac{1-a}{a} + |\log\frac{1-\gamma^*}{\gamma^*}|$

$$C \geq \frac{\left(1 - \Phi(\Delta\sqrt{\alpha_1})\right)\left(\sqrt{k^2 + 16} - k\right)}{7\beta_1\,\gamma^*(1-\gamma^*)} \exp\left(-\left(k^2 + k\right)/4\right)$$

where we additionally used $4\sqrt{e} < 7$. Bringing this all together,

$$\min_{\hat{\gamma} \in [a, 1-a]} \frac{\ell_{\mathrm{DSM}}(p^{(\hat{\gamma})}, p^{(\gamma^*)})}{(\hat{\gamma} - \gamma^*)^2} > \frac{\left(1 - \Phi(\|\mu_1 - \mu_0\|\sqrt{\alpha_1})\right)\left(\sqrt{k^2 + 16} - k\right)}{7\beta_1\gamma^*(1-\gamma^*)} \exp\left(-(k^2 + k)/4\right) \tag{42}$$

□

Remark 7. *The bound in the theorem vanishes as $\gamma \to 0$, $\gamma \to 1$, or $a \to 0$, via the term*

$$\frac{\sqrt{k^2 + 16} - k}{\gamma(1-\gamma)} \exp\left(-\frac{k^2 + k}{4}\right)$$

with

$$k = \log\frac{1-a}{a} + \left|\log\frac{1-\gamma}{\gamma}\right|.$$

The decay is mild, however: for small a, γ we have

$$k \approx \log\frac{1}{a\gamma},$$

so the term behaves as

$$\exp\left(-\frac{1}{4}\left(\log\frac{1}{a\gamma}\right)^2\right).$$

If each mode carries at least 10% *of the mass (*$\gamma \in [0.1, 0.9]$, $a = 0.1$*), the term is uniformly at least* 0.04*. Some degradation is expected, as data from a rare mode is itself vanishingly unlikely.*

The bound in Theorem 4.1 likewise vanishes as

$$|\mu_0 - \mu_1| \to \infty,$$

through the factor

$$1 - \Phi(|\mu_1 - \mu_0|\sqrt{\alpha_1})$$

(Φ *the standard normal CDF). The decay is superexponential, but* α_1 *is tiny (*e^{-10} *for a linear schedule with* $\beta_1 = 20$*) so the factor exceeds* 0.1587 *for all*

$$|\mu_1 - \mu_0| < e^5 \approx 150,$$

and the bound is non-vacuous in any realistic regime.

E Additional numerical experiments

E.1 Noise schedules and time reparameterization

In this work, we reference two noise schedules used in practice. The linear noise schedule is defined by $\beta_t := \beta_1 t$ for a constant $\beta_1 > 0$, such that $\alpha_t = \exp\left(-\frac{\beta_1}{2}t^2\right)$. The squared-cosine noise schedule (Hang et al., 2024)

$\bar{\alpha}_t = \cos\left(\frac{t+s}{1+s}\cdot\frac{\pi}{2}\right)^2$ and $\alpha_t = \frac{\bar{\alpha}_t}{\bar{\alpha}_0}$ for $s = 0.008$.

We want to analyze the effect of artificially reducing the sensitivity index on the estimation of γ^*. To do this, we can reparameterize the time in the forward process as $t = t_{a,b,v}(\tau)$, where

$$t_{a,b,v}(\tau) = \begin{cases} \tau & 0 \le \tau < a \\ a + v(\tau - a) & a \le \tau < a + \frac{b-a}{v} \\ b + \frac{1-b}{1-\left(a+\frac{b-a}{v}\right)}\left(t - a - \frac{b-a}{v}\right) & a + \frac{b-a}{v} \le t \le 1 \end{cases}$$

Under this parameterization, if we let $\mathbf{y}_\tau := \boldsymbol{x}_{t(\tau)}$ where $\boldsymbol{x}_t$ follows the Variance Preserving SDE with noise schedule $\{\alpha_t\}$, then $\mathbf{y}_t$ satisfies

$$\mathrm{d}\mathbf{y}_\tau = -\frac{\beta_{t(\tau)}}{2}\mathbf{y}_\tau\frac{\mathrm{d}t}{\mathrm{d}\tau}\mathrm{d}\tau + \sqrt{\beta_{t(\tau)}}\cdot\sqrt{\frac{\mathrm{d}t}{\mathrm{d}\tau}}\mathrm{d}W_t \tag{43}$$

$$= -\frac{\tilde{\beta}_\tau}{2}\mathbf{y}_\tau\mathrm{d}\tau + \sqrt{\tilde{\beta}_\tau}\mathrm{d}W_t \tag{44}$$

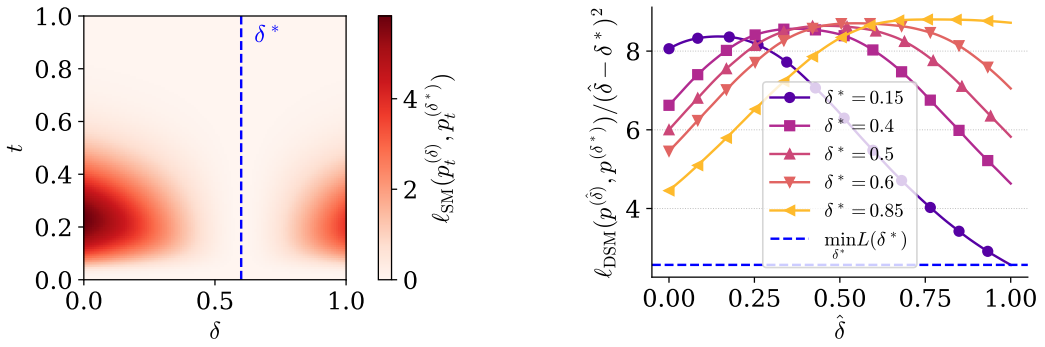
where $\tilde{\beta}_\tau := \beta_{t(\tau)}\cdot\frac{\mathrm{d}t}{\mathrm{d}\tau}$. In particular, reparameterizing time is equivalent to choosing a different noise schedule $\{\tilde{\alpha}_\tau\}$. For $v \ge 1$, $t_{a,b,v}$ maps the interval $[a, b]$ to $\left[a, a + \frac{b-a}{v}\right]$, scaling the length of the interval by $1/v$. To reduce the value of $L(\gamma^*)$, we let $[a, b]$ contain the region of time where $\frac{\ell_{\mathrm{SM}}(p_t^{(\gamma)}, p_t^{(\gamma^*)})}{(\gamma-\gamma^*)^2} > \epsilon$ for fixed $\epsilon > 0$ and increase the value of v.

E.2 Neural network experiment details

For the score models $\boldsymbol{s}_{\boldsymbol{\theta}}$ in our experiments, we use deep ResNets (He et al., 2015) with width equal to the ambient dimension and depth 8. Unless otherwise specified, we use a linear noise schedule with $\beta_1 = 20$ and a time discretization of $T = 4000$ steps on the interval $[10^{-5}, 1]$ for training and $[10^{-3}, 1]$ for sampling, following (Song et al., 2021c).

We train our models using the loss described in this paper (and (Song et al., 2021c)) with AdaBelief (Zhuang et al., 2020) with a linear warmup and cosine decay learning rate. We use JAX (Bradbury et al., 2018) and Equinox (Kidger & Garcia, 2021) for our neural network code, SciPy (Virtanen et al., 2020) for additional machine learning algorithms, Matplotlib (Hunter, 2007) for figure generation, and Weights and Biases (Biewald, 2020) for experiment tracking. [4]

[4] We will make our code available upon publication.



(a) (b)

Figure 6: Sensitivity of ℓ_{DSM} to means of mixture components. (a) $\ell_{\mathrm{SM}}(p_t^{(\delta)}, p_t^{(\delta^*)})$ as a function of t and δ. (b) $\frac{\ell_{\mathrm{DSM}}(p^{(\delta)}, p^{(\delta^*)})}{(\delta - \delta^*)^2}$ as a function of δ.

E.3 MNIST ABLATION STUDY DETAILS

The architecture of our autoencoder is simple: two convolutional layers followed by a dense layer for the encoder, and the reverse for the decoder. For our MNIST experiments, we use a latent dimension of 12. We train the autoencoder on all examples of 1s and 8s in MNIST. We additionally train an MNIST classifier on only 1s and 8s using the architecture in the Equinox documentation (Kidger & Garcia, 2021).

Unlike in the case of Gaussian mixture models, we do not know the analytical score function for the latents of our autoencoder, and thus estimation of $\ell_{\mathrm{DSM}}(p_t^{(\gamma)}, p_t^{(\gamma^*)})$ and $L(\gamma^*)$ analytically is intractable. Instead, for $\gamma^* \in \{0.2, 0.22, ..., 0.8\}$, we train diffusion models $\boldsymbol{s}_{\boldsymbol{\theta}_\gamma}$ on data from $p^{(\gamma)}$ as approximations for the analytical scores using a linear noise schedule. We keep the number of training examples constant across all experiments (5800, the smaller of the counts of 1s and 8s in MNIST) and train the models identically. For each $t \in [0, 1]$, we then have the estimate

$$\ell_{\mathrm{DSM}}(p_t^{(\gamma)}, p_t^{(\gamma^*)}) \approx \mathbb{E}_{[\boldsymbol{x}]_t \sim p_t^{(\gamma^*)}} \lambda_t \| \boldsymbol{s}_{\boldsymbol{\theta}_{\gamma^*}}(\boldsymbol{x}_t, t) - \boldsymbol{s}_{\boldsymbol{\theta}_{\gamma'}}(\boldsymbol{x}_t, t) \|^2$$

We generate all samples used in Fig. 4 using $\boldsymbol{s}_{\boldsymbol{\theta}_{0.6}}$, varying the noise schedule used for sampling.

E.4 DSSI FOR OTHER PARAMETERS

Here, we apply our DSSI framework to parameters other than the mixture weight in a Gaussian model in dimension $d = 10$. In what follows, we will let Σ_0 and Σ_1 be fixed random positive definite matrices, such that

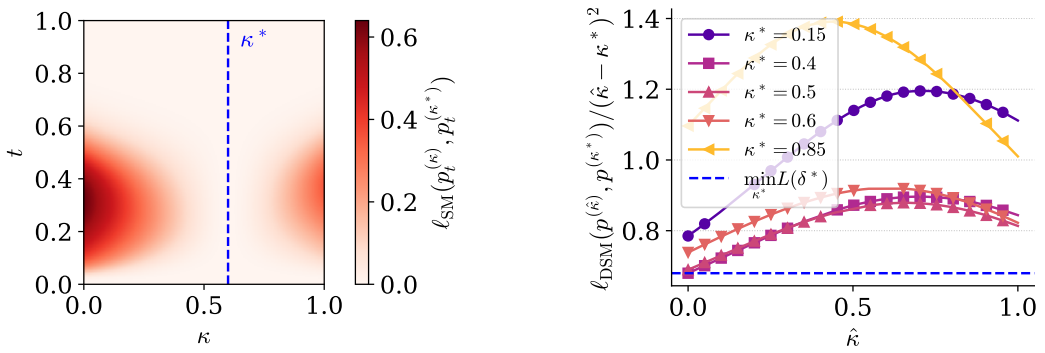
$$\Sigma_i = A^\top A + 10^{-8} I, \quad A_{ij} \sim \mathcal{N}(0, 1)$$

for $i \in \{0, 1\}$ and use a linear noise schedule.

For our analysis of the sensitivity of ℓ_{DSM} to the modes of the mixture components, we consider the one-dimensional family of distributions

$$p^{(\delta)} := 0.6\, \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_0) + 0.4\, \mathcal{N}\left(\frac{12\,\delta}{\sqrt{d}}\mathbf{1}, \boldsymbol{\Sigma}_1\right)$$

in dimension $d = 10$ where we range δ between 0 and 1. Fig. 6a shows $\ell_{\mathrm{SM}}(p_t^{(\delta)}, p_t^{(\delta^*)})$ as a function of δ and t, and Fig. 6b shows $\frac{\ell_{\mathrm{DSM}}(p_t^{(\delta)}, p_t^{(\delta^*)})}{(\delta - \delta^*)^2}$ as a function of δ for multiple values of δ^*. We see that, as with the mixture weight, $\ell_{\mathrm{SM}}(p_t^{(\delta)}, p_t^{(\delta^*)})$ is nonnegligible only for intermediate times in the forward process. However, note that the minimum value of $L(\delta^*)$ we see here is greater than 2 (compared to 0.37 in Fig. 2b), meaning that our theory implies a $5.4\times$ stronger control on the error in estimating μ using a diffusion model trained to the same level of error under ℓ_{DSM}.



(a) (b)

Figure 7: Sensitivity of ℓ_{DSM} to covariance matrix of mixture components. (a) $\ell_{\mathrm{SM}}(p_t^{(\kappa)}, p_t^{(\kappa^*)})$ as a function of t and κ. (b) $\frac{\ell_{\mathrm{DSM}}(p^{(\kappa)}, p^{(\kappa^*)})}{(\kappa-\kappa^*)^2}$ as a function of κ.

We now apply our DSSI framework to covariance matrix recovery. Consider the one-dimensional family of distributions

$$p^{(\kappa)} := 0.6\,\mathcal{N}(\mathbf{0}, \mathbf{I}) + 0.4\,\mathcal{N}\left(\frac{9.0}{\sqrt{d}}\mathbf{1}, \mathbf{\Sigma}_\kappa\right)$$

in dimension $d = 10$ where we range δ between 0 and 1. Here, we define

$$\mathbf{\Sigma}_\kappa := \mathbf{\Sigma}_0^{1/2}\left(\mathbf{\Sigma}_0^{-1/2}\mathbf{\Sigma}_1\mathbf{\Sigma}_0^{-1/2}\right)^\kappa \mathbf{\Sigma}_0^{1/2}$$

to be the geodesic interpolating between $\mathbf{\Sigma}_0$ and $\mathbf{\Sigma}_1$. Fig. 7a shows $\ell_{\mathrm{SM}}(p_t^{(\kappa)}, p_t^{(\kappa^*)})$ as a function of κ and t, and Fig. 7b shows $\frac{\ell_{\mathrm{DSM}}(p^{(\kappa)}, p^{(\kappa^*)})}{(\kappa-\kappa^*)^2}$ as a function of κ.

Again, $\ell_{\mathrm{SM}}(p_t^{(\delta)}, p_t^{(\delta^*)})$ is nonnegligable for intermediate times in the forward process. We additionally see that $L(\kappa^*)$ is greater than 0.6 uniformly across κ.