Provably Reliable Classifier Guidance via Cross-Entropy Control

Sharan Sahu[*]
ss4329@cornell.edu
Department of Statistics and Data Science
Cornell University

Arisina Banerjee[*]
ab2957@cornell.edu
Department of Statistics and Data Science
Cornell University

Yuchen Wu
yuchen.wu@cornell.edu
School of Operations Research and Information Engineering
Cornell University

Abstract

Classifier-guided diffusion models generate conditional samples by augmenting the reverse-time score with the gradient of the log-probability predicted by a probabilistic classifier. In practice, this classifier is usually obtained by minimizing an empirical loss function. While existing statistical theory guarantees good generalization performance when the sample size is sufficiently large, it remains unclear whether such training yields an effective guidance mechanism.

We study this question in the context of cross-entropy loss, which is widely used for classifier training. Under mild smoothness assumptions on the classifiers, we show that controlling the cross-entropy at each diffusion model step is sufficient to control the corresponding guidance error. In particular, we show that probabilistic classifiers achieving conditional KL divergence ε^2 induce guidance vectors with mean squared error $\widetilde{O}(d\varepsilon)$, up to constant and logarithmic factors. We demonstrate that the proposed smoothness condition is necessary, by constructing a sequence of non-smooth classifiers that achieve small conditional KL while inducing an exploding guidance error. We also show that the derived upper bound $\widetilde{O}(d\varepsilon)$ is optimal up to poly-logarithmic factors. Our result yields an upper bound on the sampling error of classifier-guided diffusion models and bears resemblance to a reverse log-Sobolev–type inequality. To the best of our knowledge, this is the first result that quantitatively links classifier training to guidance alignment in diffusion models, providing both a theoretical understanding of when such models succeed, along with principled guidelines for selecting classifiers that induce effective guidance. In particular, our findings suggest that, when selecting classifiers for guidance, one should prioritize not only accuracy but also smoothness, highlighting the advantages of smoothness-inducing training methods for diffusion guidance.

1 Introduction

Score-based diffusion models have become a leading approach to generative modeling, achieving high-quality sample generation across a wide range of applications [53, 55, 26, 54, 56, 9]. At a high level, diffusion models generate samples via a sequential denoising process driven by a sequence of time-dependent *score functions*, which in practice are usually learned from data via the score matching approach [30].

[*]Equal contribution

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Among the many extensions of this framework, *classifier guidance* has become a standard technique in modern diffusion models, enabling controllable and high-fidelity conditional sample generation across diverse domains [14, 4, 62, 34]. Specifically, to generate samples conditioned on a label y, the classifier guidance approach trains a probabilistic classifier to approximate the time-dependent label posterior $p_t(y \mid x)$. During sampling, a classifier-guided diffusion model augments the unconditional score $\nabla \log p_t(x)$ with the gradient of the log-probability predicted by the learned classifier, thereby steering the generative process toward the desired condition. In practice, the classifier used for diffusion guidance is typically learned via empirical risk minimization. The most commonly used loss function in this setting is the cross-entropy loss, which is closely related to the Kullback-Leibler (KL) divergence between the true conditional label distributions and the classifier's predictions.

A large body of statistical learning theory guarantees that, with sufficiently large sample sizes, the trained classifier generalizes well, achieving low classification error or, equivalently, a small conditional KL divergence between the true conditional label distribution and the estimated conditional label distribution [25]. However, low classification error alone does not guarantee that a trained classifier yields an effective guidance algorithm, as generalization under the cross-entropy loss does not necessarily imply that the gradient of the log-predicted probability aligns with that of the true conditional probability. In fact, empirical studies suggest that classifier guidance can be brittle: even when the classifier predicts labels correctly, its gradients can point in the wrong direction, leading to oversaturated or semantically inconsistent samples [61, 22, 15, 16]. This observation reveals a fundamental theoretical gap: classifier training objectives are typically formulated in terms of *density divergences* (e.g., conditional KL), whereas the diffusion model sampling dynamics depend on the accuracy of a *vector field* $\nabla_x \log p_t(y \mid x)$. In many applications, however, guarantees are available only in terms of the training loss, hence motivating the following key questions:

- Q1 Can we identify the precise conditions under which a small conditional label KL divergence (equivalently, cross-entropy) guarantees effective classifier guidance?
- Q2 Under the above conditions, how can the sampling error be upper bounded as a function of the conditional KL divergence?

Answering these questions not only provides a theoretical understanding of classifier guidance, but also offers practical guidance for selecting classifiers: one should favor classifiers that satisfy the conditions in Q1, so that strong classification performance translates into effective guidance.

1.1 Our contributions

In this work, we study the theoretical properties of classifier-guided diffusion models by addressing both Q1 and Q2. In particular, we make the following contributions:

Small conditional KL does not necessarily ensure effective guidance. We first show through a counterexample that probabilistic classifiers with small conditional KL may fail to provide effective diffusion guidance. To illustrate this, we construct a sequence of classifiers that attain vanishingly small conditional KL, while the associated guidance errors remain bounded away from zero or even tend to infinity. Our construction of classifiers is based on adding high-frequency perturbations to the true conditional label distributions, and disproves the belief that small classification error alone guarantees effective classifier guidance.

Small conditional KL implies effective guidance under classifier smoothness conditions. On the other hand, we show that probabilistic classifiers that achieve small conditional KL ensure effective guidance, provided they satisfy certain smoothness conditions that are also satisfied by the ground-truth model. Specifically, under mild regularity conditions we show that a conditional KL divergence of $O(\varepsilon^2)$ between the target label conditionals and the classifier-predicted conditionals implies a guidance MSE of $\widetilde{O}(d\varepsilon)$, up to constant and logarithmic factors. We further show that the derived upper bound is *optimal* by constructing an explicit instance that attains this rate.

Implications for sampling with classifier-guided diffusion models. As a direct application of our result, we use it to bound the sampling error in classifier-guided diffusion models. Specifically, we show that for Denoising Diffusion Probabilistic Models (DDPMs) with classifier guidance, if the classifiers at each step satisfy the same smoothness conditions as the corresponding label conditionals, then the KL divergence between the target and output distributions by this diffusion model scales as $\widetilde{O}(d\varepsilon_{\text{average}})$ (ignoring errors from score estimation and discretization), where $\varepsilon_{\text{average}}$ is the average of ε_t, with ε_t^2 denoting the conditional KL achieved by the classifier at time t.

1.2 Related work

Score-based generative models. This line of research traces back to the principle of score matching [29], which seeks to learn a target distribution $p_0(x)$ via estimating its score function $\nabla \log p_0(x)$. Subsequent work has shown that score matching corresponds to minimizing the Fisher divergence over some function classes, and in certain statistical models is closely related to maximum likelihood estimation [43, 57, 35]. Leveraging score matching, score-based generative models, in particular diffusion models, have emerged as a powerful class of generative models, attaining state-of-the-art performance in numerous applications [50, 5, 7]. Representative examples of score-based generative models include Denoising Diffusion Probabilistic Models (DDPM) [26] and Denoising Diffusion Implicit Models (DDIM) [54]. Various theoretical results have been established in attempt to explain the success of score-based generative models, focusing on different perspectives ranging from convergence [37, 11, 3, 24, 41, 66, 42, 28, 40, 39], score learning [64, 48, 68, 17, 36, 63], to statistical applications [18, 46, 67, 45, 10]. We refer interested readers to [9] for an overview of recent theoretical advances in diffusion models. In these examples, sample generation relies on accurately approximating the score functions of progressively noised data distributions, given by $\nabla \log p_t(x)$ at a collection of time points t.

Classifier-guided diffusion models. As discussed in the previous sections, classifier guidance [14] is a technique that steers the sample generation of an unconditional diffusion model toward the conditional distribution associated with a class label y, via augmenting the base model's score with the gradient log-probability of a separately trained, time-dependent probabilistic classifier $\nabla_x \log \hat{p}_t(y \mid x)$. The effectiveness of this framework critically depends on the quality of the guidance vector $\nabla_x \log \hat{p}_t(y \mid x)$, which ideally approximates the ground-truth guidance vector $\nabla_x \log p_t(y \mid x)$. Notably, several works analyze the behavior of guided diffusion models assuming access to the true guidance vectors [65, 6, 12, 38, 31].

Recent work has explored, both empirically and theoretically, the role of classifier quality in the design of effective guidance algorithms. For example, [44] empirically observes that standard off-the-shelf classifiers are often poorly calibrated, and argue that classifier smoothness is crucial for reliable gradient estimation. On the theoretical side, they establish that L^2 convergence of densities implies L^2 convergence of the corresponding guidance vectors under suitable smoothness assumptions. Their analysis requires the densities to be bounded above and below on a compact and convex set Ω, and the estimators to belong to a Sobolev space $H^k(\Omega)$ with smoothness $k > 1$. However, the lower bound assumption could be restrictive, as it precludes regions where the density is vanishingly small. Another work [47] focuses on the binary classification setting, assuming that the conditional probabilities are smooth and follow a specific form determined by a bounded reward function. Their classifier construction relies critically on knowledge of this reward function. More recently, several papers analyze diffusion guidance from different perspectives. For example, [23] recasts diffusion guidance as solving a regularized optimization problem along the diffusion trajectory, and [59] develops a stochastic-analysis framework for understanding and analyzing diffusion guidance. On the application side, [58] shows that classifier guidance remains powerful in rectified-flow models when coupled with a fixed-point formulation and anchored trajectories.

Information-theoretic divergences. From a theoretical perspective, our work is closely related to the connection between the KL divergence and the Fisher divergence. The relationship between these two divergences has long been a central topic in information theory and probability. For example, de Bruijn's identity relates the *time-derivative* of the KL divergence (under a Gaussian convolution) to the Fisher information and has been extended and refined in various settings (e.g., [32]). Beyond the infinitesimal regime, a growing literature studies divergences built directly from Fisher information. In particular, [52] introduces the Jensen-Fisher divergence that parallels the Jensen-Shannon divergence but with Fisher information in place of entropy, while [33] and follow-up works develop generalized and q-Fisher information measures together with their associated Jensen-type divergences. [57] analyze density estimation in (possibly infinite-dimensional) exponential families under Fisher divergence, emphasizing that control of Fisher divergence is strictly stronger than control of KL or total variation in many natural settings. In addition, if a distribution satisfies a logarithmic Sobolev inequality, then the associated KL divergence can be bounded in terms of the relative Fisher information. Our paper addresses a different question: we ask under what conditions a small conditional KL divergence over the label distribution implies a small MSE of the guidance vector, which is related to, but not identical to the Fisher divergence. This question is closely related in spirit to the reverse log-Sobolev inequalities studied in the functional-inequality literature. For instance, [19, 20] show that, for Gaussian reference measures and distributions of low

"gradient complexity", one can derive almost-tight reverse log-Sobolev inequalities that control Fisher information by entropy.

Classifier-free guidance. A popular alternative to classifier-guided approach is classifier-free guidance [27]. This method eliminates the need for a separately trained classifier by jointly training a single diffusion model on both conditional and unconditional objectives, via randomly dropping the conditioning information during training. At sampling time, guidance is achieved by steering the process using a scaled difference between the conditional and unconditional score estimates produced by this single model.

1.3 Organization

The rest of this paper is organized as follows. Section 2 introduces some preliminaries regarding diffusion models and classifier guidance. We state in Section 3 our main theoretical results, and we present in Section 4 numerical experiments that support our theoretical findings.

2 Preliminaries

Notation. For $n \in \mathbb{N}_+$, define $[n] = \{1, 2, \cdots, n\}$. For $r > 0$, we denote the closed Euclidean ball of radius r by $\mathcal{B}_r = \{x \in \mathbb{R}^d : \|x\|_2 \le r\}$, and the unit sphere by $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. We write $C^k(\mathbb{R}^d)$ for the set of k-times continuously differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$. For two positive sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we write $a_n \lesssim b_n$ if there exists a constant $C > 0$ that is independent of n, such that $a_n \le C b_n$. We write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. We write $a_n = O(b_n)$ if $a_n \le C b_n$ for some constant $C > 0$ that is independent of n, we write $a_n = \Omega(b_n)$ if $b_n = O(a_n)$, and write $a_n = \Theta(b_n)$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. We write $a_n = \widetilde{O}(b_n)$ to mean that there exists a constant $C > 0$ that is independent of n, and a poly-logarithmic factor $\mathrm{polylog}(\Pi)$ in the relevant problem parameters Π, such that $|a_n| \le C\, b_n \cdot \mathrm{polylog}(\Pi)$. We use $\nabla_x \cdot$ to represent the divergence operator with respect to the variable x.

Information-theoretic quantities. For two probability distributions P and Q, the KL divergence between P and Q is defined as $D_{\mathrm{KL}}(P \parallel Q) = \int P(\mathrm{d}x) \log \frac{P(\mathrm{d}x)}{Q(\mathrm{d}x)}$. The total variation distance between P and Q is $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|$. The cross-entropy between P and Q is $H(P, Q) = -\mathbb{E}_P[\log Q]$. Note that $H(P, Q) = H(P) + D_{\mathrm{KL}}(P \parallel Q)$, where $H(P)$ denotes the entropy of P. Closely related to cross-entropy is the cross-entropy loss $\mathcal{L}_{CE}$: for a covariate x and a label $y \in \mathcal{Y}$, the cross-entropy loss associated with a probabilistic classifier $\hat{p}$ is $\mathcal{L}_{\mathrm{CE}} = -\sum_{y' \in \mathcal{Y}} \mathbb{1}_{y=y'} \log \hat{p}(y' \mid x)$. When $(x, y) \sim p$, we have $\mathbb{E}[\mathcal{L}_{\mathrm{CE}}] = \mathbb{E}_{x \sim p}\big[H(p(\cdot \mid x), \hat{p}(\cdot \mid x))\big]$. Therefore, minimizing the empirical cross-entropy loss often leads to a small conditional KL between the true and estimated label conditionals.

2.1 Diffusion models through a continuous-time perspective

Diffusion models are often studied via their continuous-time limits, and are formulated as the numerical approximations to these limits. Conceptually, diffusion models consist of a forward and a reverse process. A common choice for the forward process is based on the Ornstein-Uhlenbeck (OU) process. Specifically, let the data distribution P_{data} be a probability measure over $\mathbb{R}^d$. The OU process $(\overrightarrow{X}_t)_{t \in [0,T]}$ is defined as the solution to the following SDE:

$$d\overrightarrow{X}_t = -\overrightarrow{X}_t\, dt + \sqrt{2}\, dB_t, \qquad \overrightarrow{X}_0 \sim P_{\mathrm{data}}, \tag{1}$$

where $(B_t)_{t \in [0,T]}$ is a standard Brownian motion in $\mathbb{R}^d$. Process (1) has an explicit solution: $\overrightarrow{X}_t = \lambda_t \overrightarrow{X}_0 + \sigma_t Z$, where $Z \sim N(0, I_d)$, $\overrightarrow{X}_0 \perp\!\!\!\perp Z$, $\lambda_t = e^{-t}$, and $\sigma_t^2 = 1 - e^{-2t}$. We use p_t to denote the marginal density of $\overrightarrow{X}_t$, and we define $m_t(x_t) = \mathbb{E}[\overrightarrow{X}_0 \mid \overrightarrow{X}_t = x_t] \in \mathbb{R}^d$ and $\Sigma_t(x_t) = \mathrm{Cov}[\overrightarrow{X}_0 \mid \overrightarrow{X}_t = x_t] \in \mathbb{R}^{d \times d}$ as the posterior mean and covariance of $\overrightarrow{X}_0$ given $\overrightarrow{X}_t = x_t$. By Tweedie's formula [51], we have $\nabla_x \log p_t(x) = -\sigma_t^{-2} x + \lambda_t \sigma_t^{-2} m_t(x)$. Likewise we define the time-reversal $(\overleftarrow{X}_t)_{t \in [0,T]}$ of process (1) by setting $\overleftarrow{X}_t = \overrightarrow{X}_{T-t}$. Under mild assumptions on the target distribution, the process $(\overleftarrow{X}_t)_{t \in [0,T]}$ satisfies the following SDE:

$$\mathrm{d}\overleftarrow{X}_t = \big\{\overleftarrow{X}_t + 2\nabla_x \log p_{T-t}(\overleftarrow{X}_t)\big\}\, \mathrm{d}t + \sqrt{2}\, \mathrm{d}B'_t, \qquad \overleftarrow{X}_0 \sim p_T, \tag{2}$$

where $(B'_t)_{t\in[0,T]}$ is another Brownian motion in $\mathbb{R}^d$ that is independent of $\overleftarrow{X}_0$ [1, 8]. Sampling $\overleftarrow{X}_T$ according to process (2) yields samples from P_{data}. In practice, the score $\nabla_x \log p_t(x)$ is not accessible and is approximated by a neural network $s_\theta(x,t)$ trained to minimize the following score-matching objective $\mathcal{L}_{\text{score}}(s_\theta) = \int_\delta^T \mathbb{E}_{X_t\sim p_t}[\|s_\theta(X_t,t) - \nabla_x \log p_t(X_t)\|_2^2]\,\mathrm{d}t$, where $\delta > 0$ denotes an early stopping cutoff. In practice, an empirical version of this objective is minimized using samples from the target distribution, for instance via score matching. To simulate the reverse dynamics (2), a standard approach is to discretize time on a grid $0 = t_0 < t_1 < \cdots < t_N < T$, with step sizes $\tau_k = t_{k+1} - t_k$. For sufficiently large T, we have $p_T \approx \pi_d$, where π_d denotes the distribution of a d-dimensional standard Gaussian random vector. Consequently, it is natural to initialize the reverse process from π_d. In practice, rather than tracking the reverse process all the way to time T, one often employs *early stopping* at time $T - \delta$ for small $\delta > 0$, in order to mitigate numerical instability arising from the score $\nabla_x \log p_t$ as $t \downarrow 0$.

2.2 Conditional sampling through classifier guidance

Classifier guidance is a standard technique in diffusion models that enables conditional sampling by augmenting the drift term of the reverse process (2) with the gradient of a trained probabilistic classifier. Specifically, we use Y to represent the label of interest, which takes values in a finite label set $\mathcal{Y}$ with some prior. The label Y is correlated with $\overrightarrow{X}_0$. With a slight abuse of notation, we also use P_{data} to represent the joint distribution of $(\overrightarrow{X}_0, Y)$, and use p_t to represent the joint distribution of $(\overrightarrow{X}_t, Y)$ at time t. Since the OU forward process (1) operates solely on the X-coordinate, the label prior remains unchanged over time, so that $p_t(y) = P_{\text{data}}(y)$ for all $t \in [0,T]$. Our goal is to sample from the conditional distribution $\overrightarrow{X}_0 \mid Y = y$. While one could train a separate conditional diffusion model for this task, doing so would require retraining the entire model for each new condition, which can be computationally expensive. Instead, classifier guidance employs a probabilistic classifier associated with the target condition to steer an unconditional diffusion model toward the desired conditional distribution, incurring substantially lower training cost.

To motivate classifier guidance, we consider the exact conditional diffusion model, whose score function at time t is $\nabla_x \log p_t(x \mid y) = \nabla_x \log p_t(x) + \nabla_x \log p_t(y \mid x)$. The conditional score decomposes into an unconditional score plus a guidance term $\nabla_x \log p_t(y \mid x)$. In the idealized setting where the exact score and guidance terms are available, the corresponding reverse SDE is

$$\mathrm{d}\overleftarrow{X}_t^{(y)} = \Big\{\overleftarrow{X}_t^{(y)} + 2\nabla_x \log p_{T-t}(\overleftarrow{X}_t^{(y)}) + 2\,\nabla_x \log p_{T-t}(y \mid \overleftarrow{X}_t^{(y)})\Big\}\mathrm{d}t + \sqrt{2}\,\mathrm{d}B'_t, \qquad (3)$$

where $X_t^{(y)} \sim p_T(\cdot \mid y)$. In practice, $p_t(y \mid x)$ is unknown and must be approximated by a trained probabilistic classifier $\hat{p}_t(y \mid x)$. A common objective is to minimize the conditional label KL

$$\mathcal{L}_{\text{cls}}(\hat{p}_t; t) = \mathbb{E}_{X_t\sim p_t}\Big[D_{\text{KL}}\big(p_t(\cdot \mid X_t) \,\|\, \hat{p}_t(\cdot \mid X_t)\big)\Big],$$

which can be approximately achieved by minimizing the empirical cross-entropy loss. Given a trained probabilistic classifier, the classifier guidance approach uses the gradient of the log predicted probability to define an approximate guidance vector field, which is then incorporated into the classifier-guided reverse SDE together with the pre-trained score:

$$\mathrm{d}\hat{X}_t^{(y)} = \Big\{\hat{X}_t^{(y)} + 2s_\theta(\hat{X}_t^{(y)}, T-t) + 2\gamma\nabla_x \log \hat{p}_{T-t}(y \mid \hat{X}_t^{(y)})\Big\}\mathrm{d}t + \sqrt{2}\,\mathrm{d}B'_t, \qquad \hat{X}_0^{(y)} \sim \pi_d, \ (4)$$

where $\gamma \geq 0$ is a hyperparameter that controls the guidance strength. In practice, a discretized version of process (4) yields a classifier-guided diffusion model. When setting $\gamma = 1$, process (4) approximates process (3) and, with accurate drift estimates, approximately terminates at the target conditional distribution. In this case, a standard application of Girsanov's Theorem shows that the KL divergence between processes (3) and (4) is controlled by the averaged score MSE and the averaged guidance term MSE. In particular, the guidance term MSE at time t is given by

$$\mathcal{E}_{\text{guid}}(t,y) = \mathbb{E}_{X_t\sim p_t(\cdot|y)}\Big[\|\nabla_x \log p_t(y \mid X_t) - \nabla_x \log \hat{p}_t(y \mid X_t)\|_2^2\Big].$$

Notably, in practice, practitioners typically set $\gamma > 1$ to enhance sample fidelity [14, 27] at the expense of sample diversity. In such settings, the goal is not to sample exactly from $P_{\text{data}}(\cdot \mid y)$, but rather to generate visually convincing samples. We note that setting $\gamma = 1$ lets (4) approximate the conditional distribution $P_{\text{data}}(\cdot \mid y)$.

3 Main results

3.1 Controlling conditional KL does not guarantee effective guidance

We begin by showing that, without any regularity assumptions on the classifiers, a small conditional KL divergence between the true conditional label distribution and its approximation does not ensure convergence of the guidance term. Concretely, we construct a sequence of classifiers $(\hat{p}_n)_{n\geq 1}$ with vanishingly small conditional KL, yet the associated guidance vectors remain misaligned. We prove Theorem 3.1 in Appendix A.1, where a construction of $\{\hat{p}_n\}_{n=1}^{\infty}$ is also provided.

Theorem 3.1. *For any p such that $p(\{x : 0 < p(y \mid x) < 1\} \mid y) > 0$ and the map $x \mapsto p(y \mid x)$ is differentiable, there exists a sequence of classifiers $\{\hat{p}_n\}_{n=1}^{\infty}$, such that almost surely over $x \sim p$,*

$$D_{\mathrm{KL}}(p(\cdot \mid x) \,\|\, \hat{p}_n(\cdot \mid x)) \to 0 \qquad \text{as } n \to \infty,$$
$$\liminf_{n\to\infty} \mathbb{E}_{x\sim p(\cdot|y)}\big[\|\nabla_x \log p(y \mid x) - \nabla_x \log \hat{p}_n(y \mid x)\|_2^2\big] > 0.$$

In fact, we can choose $\{\hat{p}_n\}_{n=1}^{\infty}$, such that either (1) $\sup_{x\in\mathbb{R}^d} D_{\mathrm{KL}}(p(\cdot \mid x) \,\|\, \hat{p}_n(\cdot \mid x)) = O(1/n)$ and $\mathbb{E}_{x\sim p(\cdot|y)}[\|\nabla_x \log p(y \mid x) - \nabla_x \log \hat{p}_n(y \mid x)\|_2^2] = \Omega(1)$, or (2) $\sup_{x\in\mathbb{R}^d} D_{\mathrm{KL}}(p(\cdot \mid x) \,\|\, \hat{p}_n(\cdot \mid x)) = O(1/\sqrt{n})$ and $\mathbb{E}_{x\sim p(\cdot|y)}[\|\nabla_x \log p(y \mid x) - \nabla_x \log \hat{p}_n(y \mid x)\|_2^2] = \Omega(n)$.

Remark 3.2. The conditions required for Theorem 3.1 are quite general: they only require differentiability and that we cannot identify the target label y with probability one. Note that if $p(\{x : p(y \mid x) = 0 \text{ or } 1\} \mid y) = 1$, then it is impossible to construct such $\{\hat{p}_n\}_{n=1}^{\infty}$, since any modification of $p(y \mid x)$ on a set of positive $p(\cdot \mid y)$-probability would incur infinite conditional KL.

Theorem 3.1 shows that, for very general target distributions, controlling the conditional KL does not guarantee control of the guidance-term MSE. Any positive result must therefore rely on additional regularity assumptions on the classifier used. Such structural regularity arises naturally in diffusion models, since the OU forward process (1) smooths the data distribution at all positive times $t > 0$, which in turn induces smoothness in the corresponding conditional label probabilities. A more detailed discussion is provided in the next section. Our construction is based on introducing high-frequency oscillations and follows standard techniques.

3.2 Effective diffusion guidance under classifier smoothness assumptions

In this section, we prove our main contribution: in diffusion model settings where the intermediate distributions are naturally smooth, controlling the conditional KL divergence guarantees control of the guidance MSE, provided that the classifiers used for guidance satisfy standard smoothness conditions that hold for the true conditional label probabilities. We prove Theorem 3.3 in Appendix A.2.

Theorem 3.3. *Suppose we wish to sample from the conditional distribution $P_{\mathrm{data}}(\cdot \mid y)$ associated with a label y. Fix $T > 0$ and a reverse-time grid $0 = t_0 < t_1 < \cdots < t_N = T - \delta$. Let $\tau_k = t_{k+1} - t_k$ and $s_k = T - t_k$ for $k \in \{0\} \cup [N-1]$. Assume the following:*

1. *The data distribution P_{data} is supported on a bounded set $\mathcal{K} \subseteq \mathbb{R}^d$ with $R = \sup_{x\in\mathcal{K}} \|x\|_2 < \infty$, in the sense that $P_{\mathrm{data}}(x \in \mathcal{K}) = 1$.*
2. *For $k \in \{0\} \cup [N-1]$, $x \mapsto \hat{p}_{s_k}(y \mid x)$ belongs to $C^2(\mathbb{R}^d)$. In addition,*
$$\|\nabla_x \log \hat{p}_{s_k}(y \mid x)\|_2 \leq 2\lambda_{s_k}\sigma_{s_k}^{-2}R, \qquad \big|\mathrm{Tr}(\nabla_x^2 \log \hat{p}_{s_k}(y \mid x))\big| \leq 2\lambda_{s_k}^2\sigma_{s_k}^{-4}R^2.$$
3. *There is a universal constant $C > 0$, such that for $k \in \{0\} \cup [N-1]$ and small enough ϵ,*
   (a) $\mathbb{P}_{x\sim p_{s_k}(\cdot|y)}\left(\|\nabla_x \log \hat{p}_{s_k}(y \mid x)\|_2 \leq \frac{C\sqrt{d+\log(1/\epsilon)}}{\sigma_{s_k}}\right) \geq 1 - \epsilon/P_{\mathrm{data}}(y)$,
   (b) $\mathbb{P}_{x\sim p_{s_k}(\cdot|y)}\left(\left|\mathrm{Tr}\big(\nabla_x^2 \log \hat{p}_{s_k}(y \mid x)\big)\right| \leq \frac{C(d+\log(1/\epsilon))}{\sigma_{s_k}^2}\right) \geq 1 - \epsilon/P_{\mathrm{data}}(y)$.
4. *The target label satisfies $P_{\mathrm{data}}(y) > 0$.*
5. *For all $k \in \{0\} \cup [N-1]$, $\mathbb{E}_{x\sim p_{s_k}}[D_{\mathrm{KL}}(p_{s_k}(\cdot \mid x) \,\|\, \hat{p}_{s_k}(\cdot \mid x))] \leq \varepsilon_{s_k}^2$ for $\varepsilon_{s_k} \in [0, P_{\mathrm{data}}(y)/\sqrt{2}]$.*

Then for all $k \in \{0\} \cup [N-1]$, there exists a numerical constant $C_0 > 0$, such that

$$\mathbb{E}_{x\sim p_{s_k}(\cdot|y)}\Big[\big\|\nabla_x \log p_{s_k}(y \mid x) - \nabla_x \log \hat{p}_{s_k}(y \mid x)\big\|_2^2\Big]$$
$$\leq \frac{C_0\, \varepsilon_{s_k}}{\sigma_{s_k}^2 P_{\mathrm{data}}(y)}\Big(d + \log\Big(\frac{R^2+d}{\sigma_{s_k}^2}\Big) + \log\Big(\frac{1}{\varepsilon_{s_k}}\Big) + \log\Big(\frac{1}{P_{\mathrm{data}}(y)}\Big)\Big).$$

Remark 3.4. Ignoring logarithmic factors in $(R, d, 1/\delta, 1/P_{\mathrm{data}}(y))$, and additionally assume $d \gtrsim \log(1/\varepsilon_{s_k})$, we have $\mathbb{E}_{x\sim p_{s_k}(\cdot|y)}[\|\nabla_x \log p_{s_k}(y \mid x) - \nabla_x \log \hat{p}_{s_k}(y \mid x)\|_2^2] = O(\frac{d\varepsilon_{s_k}}{\sigma_{s_k}^2})$.

Remark 3.5. By Lemma B.1 and Corollary B.2 in the appendix, the second and the third assumptions of the lemma hold if we replace $\hat{p}_{s_k}$ with the true probability p_{s_k}. The first assumption holds for many datasets of practical interest to diffusion models, such as image datasets with bounded pixel values. The C^2 smoothness assumption is standard and can be verified for common classifiers such as Gaussian mixture models and logistic regression, as well as for neural networks via backtracking-type methods. We note that both the boundedness and C^2 assumptions are standard and have been widely adopted in the diffusion model literature [37, 28]. The high-probability bounds in the third assumption are weaker than their pointwise counterparts and hold for simple parametric families. We however note that the corresponding sub-Gaussian and sub-exponential conditions may not hold for typical deep neural network classifiers. The fourth assumption requires that the target label occurs with positive probability. The final assumption imposes an upper bound on the classification error as measured by the conditional KL.

The bound in Theorem 3.3 is related to, but distinct from, a reverse log-Sobolev inequality: Our goal is to control the guidance error, a quantity over the x-space conditioned on a single y, using a conditional KL over the label y-space without an absolute value operator. We note that establishing a reverse log-Sobolev inequality is usually highly non-trivial and density smoothness alone is generally insufficient to establish such inequalities. In particular, it does not follow directly from standard Gagliardo–Nirenberg–type inequalities.

In the next theorem, we assume a standard time-decaying step size for the diffusion model, as in [2], and under this choice derive a bound on the average MSE of the guidance term. With Theorem 3.3, the following lemma follows directly from the proof in [2]. We prove Lemma 3.1 in Appendix A.3.

Lemma 3.1. *Fix $T > 0$ and a reverse-time grid $0 = t_0 < t_1 < \cdots < t_N = T - \delta$. Let $\tau_k = t_{k+1} - t_k$ and $s_k = T - t_k$ for $k \in \{0\} \cup [N-1]$. Assume there exists $\kappa > 0$ such that $\tau_k \leq \kappa \min\{1,\, s_{k+1}\}$ for $k = 0, \ldots, N-1$. Define the average guidance term MSE*

$$\varepsilon_{\mathrm{guide}}^2 \;=\; \sum_{k=0}^{N-1} \tau_k\, \mathbb{E}_{x\sim p_{s_k}(\cdot|y)}\Big[\big\|\nabla_x \log p_{s_k}(y \mid x) - \nabla_x \log \hat{p}_{s_k}(y \mid x)\big\|_2^2\Big].$$

Under the assumptions of Theorem 3.3, there exists a positive numerical constant c_0, such that

$$\varepsilon_{\mathrm{guide}}^2 \leq c_0 \sum_{k=0}^{N-1} \frac{\kappa \varepsilon_{s_k}}{P_{\mathrm{data}}(y)}\Big(d + \log\Big(\frac{R^2+d}{\delta}\Big) + \log\Big(\frac{1}{\varepsilon_{s_k}}\Big) + \log\Big(\frac{1}{P_{\mathrm{data}}(y)}\Big)\Big).$$

The average guidance MSE $\varepsilon_{\mathrm{guide}}^2$ will later be used to control the sampling error in Section 3.4, and κ can be interpreted as controlling the maximum step size. Theorem 3.1 and Theorem 3.3 together provide a complete characterization of classifier guidance. On the one hand, in the absence of classifier smoothness, control of the conditional KL divergence does not guarantee control of the guidance term MSE. On the other hand, for bounded target distributions, if the classifier satisfies the same smoothness conditions as the ground-truth conditional label probabilities, this pathology disappears, and small conditional KL does imply control of the guidance MSE.

3.3 Rate optimality

The bound in Theorem 3.3 shows that as the conditional KL divergence vanishes at rate $O(\varepsilon^2)$, the guidance term MSE decays at order $O(\varepsilon)$ (ignoring dependence on other problem parameters). A natural question is whether this rate can be improved. Our next result shows that such an improvement is impossible. Specifically, we construct a data distribution and a family of classifiers that satisfy Assumptions (1)–(5) of Theorem 3.3, such that the conditional KL divergence scales as $O(\varepsilon^2)$, while the guidance MSE is $\Omega(\varepsilon)$. Our construction demonstrates that the ε-dependence in Theorem 3.3 is tight. We prove Theorem 3.6 in Appendix A.4. We also present our construction therein.

Theorem 3.6. *There exists a data distribution P_{data}, such that for any sufficiently small ε, there exists a sequence of probabilistic classifiers $x \mapsto \hat{p}_t(\cdot \mid x)$ for $t \in \{s_k : k \in \{0\} \cup [N-1]\}$ that satisfy Assumptions* (1)–(5) *of Theorem* 3.3*, such that*

(i) For a constant $C_1 > 0$ that is independent of ε,

$$\sup_{t \in \{s_k : k \in \{0\} \cup [N-1]\}} \mathbb{E}_{X_t \sim p_t(\cdot)} \Big[D_{\mathrm{KL}}\big(p_t(\cdot \mid X_t) \,\|\, \hat{p}_t(\cdot \mid X_t)\big) \Big] \;\le\; C_1\, \varepsilon^2.$$

(ii) For a constant $C_2 > 0$ that is independent of ε,

$$\inf_{t \in \{s_k : k \in \{0\} \cup [N-1]\}} \mathbb{E}_{X_t \sim p_t(\cdot \mid y)} \Big[\big\| \nabla_x \log p_t(y \mid X_t) - \nabla_x \log \hat{p}_t(y \mid X_t) \big\|_2^2 \Big] \;\ge\; C_2\, \varepsilon.$$

In summary, we not only provide an upper bound on the L^2 guidance error in terms of conditional KL, but also derive a nearly optimal rate. To our knowledge, such rate is not a trivial consequence of existing results, which we believe is a meaningful contribution.

3.4 Implications for conditional sampling error

In practice, diffusion models are implemented by discretizing continuous-time processes and using learned score functions. Consequently, the sampling error depends not only on the guidance error, but also on the discretization error and the score estimation error. For completeness, we analyze time-discretized guided diffusion in this section.

We consider DDPM-style samplers equipped with an exponential integrator. Specifically, we fix a time grid $0 = t_0 < t_1 < \cdots < t_N = T - \delta$, and define $\tau_k = t_{k+1} - t_k$ and $s_k = T - t_k$ for $k \in \{0\} \cup [N-1]$. For each step $k = 1, 2, \cdots, N$, we train a classifier $\hat{p}_{s_k}$ and a score estimate $s_\theta(\cdot, s_k)$. Given these components, a DDPM-type sampler starts from $\hat{X}_0^{(y)} \sim N(0, I_d)$ and, at the k-th update step for $k = 1, 2, \cdots, N$, set

$$\hat{X}_{t_{k+1}}^{(y)} = e^{\tau_k} \hat{X}_{t_k}^{(y)} + 2(e^{\tau_k} - 1)\big[s_\theta(\hat{X}_{t_k}^{(y)}, s_k) + \gamma\, \nabla \log \hat{p}_{s_k}(y \mid X_{t_k}^{(y)})\big] + \sqrt{e^{2\tau_k} - 1}\, Z_k, \tag{5}$$

where $Z_k \sim_{i.i.d.} N(0, I_d)$ are independent of the previous estimates, $\gamma > 0$ is a hyperparameter that controls the guidance strength. For $k = 0, 1, \ldots, N$, we denote by $\hat{q}_{\gamma, t_k}$ the marginal distribution of $\hat{X}_{t_k}^{(y)}$, so that $\hat{q}_{\gamma, T-\delta}$ denotes the output distribution of sampler (5). We compare $\hat{q}_{\gamma, T-\delta}$ against the corresponding ideal guided marginal distribution of $\overleftarrow{X}_t^{(y)}$. Specifically, for $t \in [0, T]$, define the γ-tilted distribution

$$p_{t,\gamma}(x) = \frac{p_t(x)\, p_t(y \mid x)^\gamma}{Z_{t,\gamma}}, \quad Z_{t,\gamma} = \mathbb{E}_{X \sim p_t}\left[p_t(y \mid X)^\gamma\right]. \tag{6}$$

We then consider the ideal classifier-guided reverse process

$$\mathrm{d}\overleftarrow{X}_t^{(y)} = \Big\{ \overleftarrow{X}_t^{(y)} + 2\nabla_x \log p_{T-t}(\overleftarrow{X}_t^{(y)}) + 2\gamma \nabla_x \log p_{T-t}(y \mid \overleftarrow{X}_t^{(y)}) \Big\} \mathrm{d}t + \sqrt{2}\, \mathrm{d}B'_t, \tag{7}$$

where $\overleftarrow{X}_0^{(y)} \sim p_{T,\gamma}$. Let $q_{\gamma,t}$ denote the marginal distribution of $\overleftarrow{X}_t^{(y)}$. Our analysis below bounds the sampling error between the output distribution $\hat{q}_{\gamma, T-\delta}$ and the corresponding ideal guided marginal $q_{\gamma, T-\delta}$ for $\gamma > 1$.

To state our result, we impose the following assumption on the score estimation error.

Assumption 3.1. *The score estimate s_θ satisfies $\sum_{k=0}^{N-1} \tau_k \mathbb{E}_{X \sim p_{T-t_k}}[\|\nabla_x \log p_{T-t_k}(X) - s_\theta(X, T-t_k)\|_2^2] \le \varepsilon_{\mathrm{score}}^2$.*

Under Assumption 3.1 and the assumptions of Theorem 3.3, we establish the following upper bound on the sampling error. We prove Theorem 3.7 in Appendix A.5.

Theorem 3.7. *Suppose $T \ge 1$ and assume there exists $\kappa > 0$ such that for each $k = 0, \ldots, N-1$, $\tau_k \le \kappa \min\{1,\, T - t_{k+1}\} = \kappa \min\{1,\, s_{k+1}\}$. Let $\Gamma_\gamma = 2\gamma^2 - 2\gamma + 1$. Then, under Assumption 3.1 and the assumptions of Theorem 3.3, it holds that*

$$D_{\mathrm{KL}}\big(q_{\gamma, T-\delta} \,\|\, \hat{q}_{\gamma, T-\delta}\big) \;\le\; C_{\mathrm{s}} \left(\frac{\varepsilon_{\mathrm{score}}^2 + \gamma^2 \varepsilon_{\mathrm{guide}}^2 + \Gamma_\gamma (d + R^2)(\kappa T + \kappa^2 N)}{P_{\mathrm{data}}(y)^\gamma} + \frac{\gamma^2 (d + R^2) e^{-2T}}{P_{\mathrm{data}}(y)^{2\gamma}} \right).$$

where we recall $\varepsilon_{\mathrm{guide}}^2$ is from Lemma 3.1, and C_{s} is a positive numerical constant.

Remark 3.8. The bound on $\varepsilon_{\mathrm{guide}}^2$ in Theorem 3.3 scales linearly with d, and this dependence cannot be improved for KL between d-dimensional distributions (e.g., for product distributions).

4 Experimental results

In this section, we present numerical experiments that support our theoretical findings, including a Gaussian mixture model (GMM) and MNIST experiment. Numerical validation of the counterexample (Theorem 3.1) is provided in Appendix C.

4.1 Guidance error control with classifier smoothness

We now validate Theorem 3.3 in the setting of a Gaussian mixture model (GMM).

Experimental setup. We consider a binary label space $\mathcal{Y} = \{0, 1\}$ and a sample space $\mathbb{R}^{50}$. We denote the target distribution by p. We set $p(y = 0) = p(y = 1) = 0.5$, $\mu_0 = [-2, 0, \ldots, 0]^\top \in \mathbb{R}^{50}$, $\mu_1 = [2, 0, \ldots, 0]^\top \in \mathbb{R}^{50}$, and assume the following GMM: $p(x \mid y = 0) = \mathcal{N}(\mu_0, 0.5\, I_{50})$ and $p(x \mid y = 1) = \mathcal{N}(\mu_1, 0.5\, I_{50})$. Consider the forward noising process of the diffusion model in (1) initialized at p, and denote by p_t the distribution at time t. The true conditional probabilities $p_t(y \mid x)$ can be computed analytically via Bayes' rule. For diffusion time $t = 1.0$, we generate independent training samples $\{(x_i, y_i)\}_{i=1}^N$ from p and fit a two-component Gaussian mixture model $\hat{p}_t$ using the EM algorithm with a diagonal covariance structure, which yields smooth classifiers and remains statistically reliable in high dimensions. We consider sample sizes $N \in \{200, 500, 1000, 2000, 3000, 5000\}$; sample sizes below 200 are excluded as diagonal GMM estimation becomes unreliable in $\mathbb{R}^{50}$ at very small N. Each experiment is repeated over 20 independent trials, and we report the mean and standard error. For each (t, N), we use Monte Carlo simulation with a test set of size 1,000 to calculate the conditional KL and the guidance MSE.

Simulation outcomes. Figure 1 plots the conditional KL divergence and the guidance MSE as functions of the sample size N, for each diffusion time $t = 1.0$. We observe that both the conditional KL divergence and the guidance MSE decrease consistently as N increases across all time points, with similar decay trends. Similar observations are also made for diffusion time points $t \in \{0.01, 0.1, 3.0\}$ This indicates that smaller classifier conditional KL indeed leads to smaller guidance MSE under classifier smoothness conditions, empirically validating Theorem 3.3 in the 50-dimensional setting.

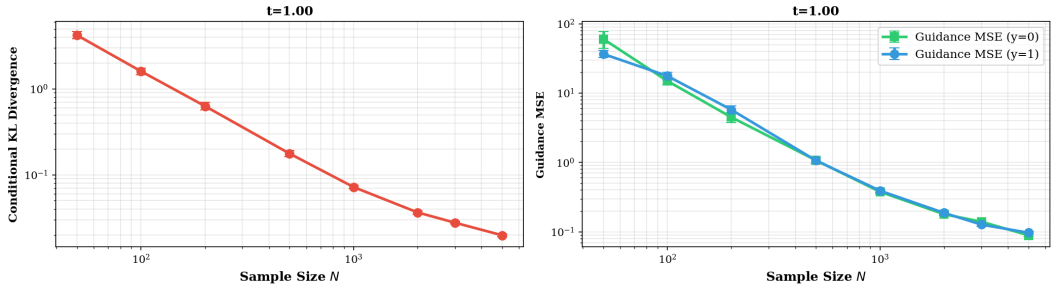


Figure 1: Numerical validation of Theorem 3.3 under a GMM. The left panel displays the conditional KL divergence, and the right panel shows the guidance MSE.

4.2 Real-world example using MNIST dataset

Finally, we validate our theoretical findings on the MNIST dataset [13]. We compare two classifiers trained using neural networks with similar classification accuracy but different smoothness properties. Our experiments show these classifiers exhibit different levels of effectiveness when used for guidance.

Experimental setup. We consider the 10-class MNIST digit classification task with label space $\mathcal{Y} = \{0, 1, \ldots, 9\}$ and sample space $\mathbb{R}^{28\times 28}$. We assume that the images are normalized so that pixel values lie in $[-1, 1]$. We take $\delta = 0.01$ and $T = 1.5$. We train two time-dependent neural classifiers $\hat{p}_t^{\text{smooth}}(y \mid x)$ and $\hat{p}_t^{\text{non-smooth}}(y \mid x)$ to predict labels from noise-corrupted images. Both classifiers are trained by minimizing the cross-entropy objective and use the same overall architecture, including the same convolutional layers, kernel sizes, normalization, and average pooling, differing only in their activation functions. In particular, the smooth classifier employs GELU and tanh activations, which have continuous higher-order derivatives, whereas the non-smooth classifier uses ReLU and Hardtanh activations, which introduce transition boundaries at which the gradient is discontinuous. Both classifiers are trained for 500 epochs on 60,000 MNIST training samples with a batch size of 256, and they achieve comparable classification accuracies of 0.867 and 0.861 and cross-entropy losses of 0.4013 and 0.4018, respectively.

Simulation outcomes. Figure 2 summarizes the numerical results. The top panel shows the fraction of generated images that are correctly classified by a held-out MNIST classifier trained independently on clean images, while the bottom panel reports the averaged gradient norm and Hessian trace as functions of diffusion time t. We observe that the smooth classifier exhibits significantly smaller gradient norms and Hessian traces than the non-smooth classifier. In addition, the smooth classifier achieves stronger guidance performance, as measured by the fraction of generated samples that are correctly classified, supporting our theoretical findings. Further details of these numerical experiments are provided in Appendix C.

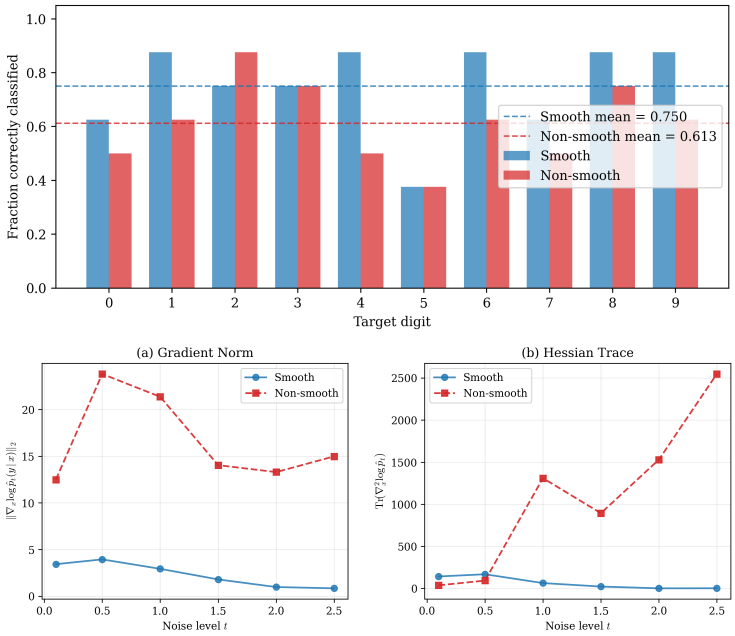


Figure 2: **Top panel:** Guidance effectiveness by target digit. **Bottom panel:** Averaged gradient norm and Hessian trace as functions of diffusion time t.

5 Conclusion

In this work, we identify tight conditions under which classification error (measured by conditional KL divergence) controls the effectiveness of diffusion guidance. Our negative result shows that low classification error alone is insufficient, by exhibiting a concrete example in which the classifier attains small conditional KL while the resulting guidance terms are severely misaligned. Complementarily, our positive result establishes that, under additional classifier smoothness assumptions, a probabilistic classifier with conditional KL ε^2 induces a guidance MSE of order $\widetilde{O}(d\varepsilon)$. We prove such dependency on ε is optimal, and further leverage our result to derive sampling error bounds for classifier-guided diffusion models. Several directions merit further investigation, including extending our analysis to non-compactly supported distributions and exploring alternative classifier training procedures. By rigorously connecting classifier training error to diffusion guidance quality, this work provides both theoretical foundations and practical insights for classifier-guided diffusion models.

References

[1] Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

[2] Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024.

[3] Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *The Twelfth International Conference on Learning Representations*, 2024.

[4] Cosmin I. Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A. Schnabel. Diffusion Models with Implicit Guidance for Medical Anomaly Detection . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15011. Springer Nature Switzerland, October 2024.

[5] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[6] Arwen Bradley and Preetum Nakkiran. Classifier-free guidance is a predictor-corrector. In *NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning*.

[7] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Wing Yin Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. *OpenAI*, Feb 2024.

[8] Patrick Cattiaux, Giovanni Conforti, Ivan Gentil, and Christian Léonard. Time reversal of diffusion processes under a finite entropy condition, 2022.

[9] Minshuo Chen, Song Mei, Jianqing Fan, and Mengdi Wang. An overview of diffusion models: Applications, guided generation, statistical rates and optimization. *arXiv preprint arXiv:2404.07771*, 2024.

[10] Minshuo Chen, Renyuan Xu, Yumin Xu, and Ruixun Zhang. Diffusion factor models: Generating high-dimensional returns with factor structure. *arXiv preprint arXiv:2504.06566*, 2025.

[11] Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023.

[12] Muthu Chidambaram, Khashayar Gatmiry, Sitan Chen, Holden Lee, and Jianfeng Lu. What does guidance do? a fine-grained analysis in a simple setting. *Advances in Neural Information Processing Systems*, 37:84968–85005, 2024.

[13] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[14] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[15] Anh-Dung Dinh, Daochang Liu, and Chang Xu. PixelAsParam: A gradient view on diffusion sampling with guidance. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 8120–8137. PMLR, 23–29 Jul 2023.

[16] Anh-Dung Dinh, Daochang Liu, and Chang Xu. Rethinking conditional diffusion sampling with progressive guidance. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[17] Zehao Dou, Subhodh Kotekal, Zhehao Xu, and Harrison H Zhou. From optimal score matching to optimal sampling. *arXiv preprint arXiv:2409.07032*, 2024.

[18] Ahmed El Alaoui, Andrea Montanari, and Mark Sellke. Sampling from the sherrington-kirkpatrick gibbs measure via algorithmic stochastic localization. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 323–334. IEEE, 2022.

[19] Ronen Eldan. Gaussian-width gradient complexity, reverse log-sobolev inequalities and nonlinear large deviations, 2018.

[20] Ronen Eldan and Michel Ledoux. A dimension-free reverse logarithmic sobolev inequality for low-complexity functions in gaussian space, 2019.

[21] G.B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013.

[22] Roy Ganz, Bahjat Kawar, and Michael Elad. Do perceptually aligned gradients imply robustness? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10628–10648. PMLR, 23–29 Jul 2023.

[23] Yingqing Guo, Hui Yuan, Yukang Yang, Minshuo Chen, and Mengdi Wang. Gradient guidance for diffusion models: An optimization perspective. *arXiv preprint arXiv:2404.14743*, 2024.

[24] Shivam Gupta, Linda Cai, and Sitan Chen. Faster diffusion-based sampling with randomized midpoints: Sequential and parallel. *arXiv e-prints*, pages arXiv–2406, 2024.

[25] Trevor Hastie, Robert Tibshirani, Jerome Friedman, et al. The elements of statistical learning, 2009.

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[27] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[28] Daniel Zhengyu Huang, Jiaoyang Huang, and Zhengjiang Lin. Convergence analysis of probability flow ode for score-based generative models. *IEEE Transactions on Information Theory*, 2025.

[29] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

[30] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[31] Yuchen Jiao, Yuxin Chen, and Gen Li. Towards a unified framework for guided diffusion models. *arXiv preprint arXiv:2512.04985*, 2025.

[32] OT Johnson and AR Barron. Fisher information inequalities and the central limit theorem. *Probability Theory and Related Fields*, 129 (3):391 – 409, July 2004. Publisher: Springer.

[33] Omid Kharazmi, Narayanaswamy Balakrishnan, and Hassan Jamali. Cumulative residual q-fisher information and jensen-cumulative residual χ^2 divergence measures. *Entropy*, 24(3):341, 2022.

[34] Gwanghyun Kim, Taesung Kwon, and {Jong Chul} Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings - 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 2416–2425. IEEE Computer Society, 2022. Publisher Copyright: © 2022 IEEE.; 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022 ; Conference date: 19-06-2022 Through 24-06-2022.

[35] 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.

[36] Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: The view from isoperimetry. *ICLR*, 2023.

[37] Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.

[38] Gen Li and Yuchen Jiao. Provable efficiency of guidance in diffusion models for general data distribution. *arXiv preprint arXiv:2505.01382*, 2025.

[39] Gen Li and Yuling Yan. $o(d/t)$ convergence theory for diffusion probabilistic models under minimal assumptions. In *The Thirteenth International Conference on Learning Representations*, 2025.

[40] Gen Li, Yuchen Zhou, Yuting Wei, and Yuxin Chen. Faster diffusion models via higher-order approximation. *arXiv preprint arXiv:2506.24042*, 2025.

[41] Runjia Li, Qiwei Di, and Quanquan Gu. Unified convergence analysis for score-based diffusion models with deterministic samplers. *arXiv preprint arXiv:2410.14237*, 2024.

[42] Yuchen Liang, Peizhong Ju, Yingbin Liang, and Ness Shroff. Broadening target distributions for accelerated diffusion models via a novel analysis approach. *arXiv preprint arXiv:2402.13901*, 2024.

[43] Siwei Lyu. Interpretation and generalization of score matching. *arXiv preprint arXiv:1205.2629*, 2012.

[44] Jiajun Ma, Tianyang Hu, Wenjia Wang, and Jiacheng Sun. Elucidating the design space of classifier-guided diffusion generation. In *The Twelfth International Conference on Learning Representations*, 2024.

[45] Song Mei and Yuchen Wu. Deep networks as denoising algorithms: Sample-efficient learning of diffusion models in high-dimensional graphical models. *IEEE Transactions on Information Theory*, 2025.

[46] Andrea Montanari and Yuchen Wu. Posterior sampling in high dimension via diffusion processes. *arXiv preprint arXiv:2304.11449*, 2023.

[47] Owen Oertell, Shikun Sun, Yiding Chen, Jin Peng Zhou, Zhiyong Wang, and Wen Sun. Efficient controllable diffusion via optimal classifier guidance, 2025.

[48] Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pages 26517–26582. PMLR, 2023.

[49] B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer Berlin Heidelberg, 2010.

[50] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

[51] Herbert E Robbins. An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and basic theory*, pages 388–394. Springer, 1992.

[52] P. Sánchez-Moreno, A. Zarzo, and J. S. Dehesa. Jensen divergence based on fisher's information. *Journal of Physics A: Mathematical and Theoretical*, 45(12):125305, 2012.

[53] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

[54] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[55] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[56] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[57] Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57):1–59, 2017.

[58] Zhicheng Sun, Zhenhao Yang, Yang Jin, Haozhe Chi, Kun Xu, Kun Xu, Liwei Chen, Hao Jiang, Yang Song, Kun Gai, and Yadong Mu. Rectifid: Personalizing rectified flow with anchored classifier guidance. In *Advances in Neural Information Processing Systems*, 2024.

[59] Wenpin Tang and Renyuan Xu. Conditional diffusion guidance. Technical report, Columbia University, 2024. Preprint, available at `https://www.columbia.edu/~wt2319/CDG.pdf`.

[60] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[61] Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Vijay Naik. End-to-end diffusion latent optimization improves classifier guidance. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7246–7256, 2023.

[62] Qimin Wang, Xinda Liu, and Guohua Geng. Guidpaint: Class-guided image inpainting with diffusion models, 2025.

[63] Yuqing Wang, Ye He, and Molei Tao. Evaluating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 37:19307–19352, 2024.

[64] Andre Wibisono, Yihong Wu, and Kaylee Yingxi Yang. Optimal score estimation via empirical bayes smoothing. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4958–4991. PMLR, 2024.

[65] Yuchen Wu, Minshuo Chen, Zihao Li, Mengdi Wang, and Yuting Wei. Theoretical insights for diffusion guidance: a case study for gaussian mixture models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 53291–53327, 2024.

[66] Yuchen Wu, Yuxin Chen, and Yuting Wei. Stochastic runge-kutta methods: Provable acceleration of diffusion models. *arXiv preprint arXiv:2410.04760*, 2024.

[67] Xingyu Xu and Yuejie Chi. Provably robust score-based diffusion posterior sampling for plug-and-play image reconstruction. *Advances in Neural Information Processing Systems*, 37:36148–36184, 2024.

[68] Kaihong Zhang, Caitlyn H Yin, Feng Liang, and Jingbo Liu. Minimax optimality of score-based diffusion models: beyond the density lower bound assumptions. In *Proceedings of the 41st International Conference on Machine Learning*, pages 60134–60178, 2024.

A Proofs of the main theoretical results

A.1 Proof of Theorem 3.1

For $\gamma \in (0,1)$, define $A_\gamma = \{x : 0 < p(y \mid x) < 1-\gamma\}$. By assumption, there exists $\gamma \in (0,1)$, such that $p(A_\gamma \mid y) > 0$. Denote by x_1 the first coordinate of x. Let $\{\delta_n\}_{n=1}^\infty$ be a sequence of amplitudes satisfying $\delta_n \downarrow 0$ as $n \to \infty$. For $x \in A_\gamma$, define the following perturbation:

$$\hat{p}_n(y \mid x) = p(y \mid x)\big(1 + \delta_n \sin(nx_1)\big),$$
$$\hat{p}_n(y' \mid x) = p(y' \mid x) \cdot \frac{1 - p(y \mid x) - \delta_n \sin(nx_1)\, p(y \mid x)}{1 - p(y \mid x)}, \qquad y' \in \mathcal{Y}\backslash\{y\}.$$

For $x \notin A_\gamma$, we let $\hat{p}_n(\cdot \mid x) = p(\cdot \mid x)$. For any $\delta_n \in [0, \gamma/(2-2\gamma)\,)$, the conditional density $\hat{p}_n(\cdot \mid x)$ is well defined and adds up to one almost surely with respect to $x \sim p$.

Note that for $x \notin A_\gamma$, by definition $D_{\mathrm{KL}}(p(\cdot \mid x) \,\|\, \hat{p}_n(\cdot \mid x)) = 0$. For $x \in A_\gamma$,

$$\begin{aligned} D_{\mathrm{KL}}(p(\cdot \mid x) \,\|\, \hat{p}_n(\cdot \mid x)) =& -p(y \mid x)\log(1 + \delta_n \sin(nx_1)) \\ &- \sum_{y' \in \mathcal{Y}\backslash\{y\}} p(y' \mid x) \log \frac{1 - p(y \mid x) - \delta_n \sin(nx_1)\, p(y \mid x)}{1 - p(y \mid x)} \\ \le& \delta_n + \frac{\delta_n p(y \mid x)}{1 - p(y \mid x) - \delta_n \sin(nx_1)\, p(y \mid x)} \le 2\delta_n \gamma^{-1}. \end{aligned}$$

where we have applied the standard bounds $\log(1+x) \le x$ for $x \ge 0$ and $-\log(1-x) \le \frac{x}{1-x}$ for $0 \le x < 1$. Therefore, as $n \to \infty$, almost surely over $x \sim p$

$$D_{\mathrm{KL}}(p(\cdot \mid x) \,\|\, \hat{p}_n(\cdot \mid x)) \le 2\delta_n \gamma^{-1} \to 0.$$

Next, we show that the guidance vector corresponding to $\hat{p}_n$ is misaligned with that of p. Note that for $x \in A_\gamma$,

$$\begin{aligned}\nabla_x \log \hat{p}_n(y \mid x) =& \nabla_x \log p(y \mid x) + \nabla_x \log(1 + \delta_n \sin(nx_1)) \\ =& \nabla_x \log p(y \mid x) + \frac{\delta_n n \cos(nx_1)}{1 + \delta_n \sin(nx_1)} e_1,\end{aligned}$$

where $e_1 \in \mathbb{R}^d$ denotes the first standard basis vector. Therefore, using the fact that $\cos^2 x = (1+\cos(2x))/2$ and $\delta_n \in [0, \gamma/(2-2\gamma)\,)$, we see that

$$\begin{aligned}&\mathbb{E}_{x \sim p(\cdot \mid y)}\left[\|\nabla_x \log p(y \mid x) - \nabla_x \log \hat{p}_n(y \mid x)\|_2^2\right] = \int_{A_\gamma} \frac{\delta_n^2 n^2 \cos(nx_1)^2}{(1+\delta_n \sin(nx_1))^2} p(x \mid y)\,\mathrm{d}x \\ &\ge \frac{(1-\gamma)^2 \delta_n^2 n^2}{2}\left[p(A_\gamma \mid y) + \int_{\mathbb{R}} \cos(2nx_1)\Big(\int_{\mathbb{R}^{d-1}} \mathbb{1}_{A_\gamma}(x) p(x_1, x_{2:d} \mid y)\mathrm{d}x_{2:d}\Big)\mathrm{d}x_1\right].\end{aligned}$$

Note that $x_1 \mapsto \int_{\mathbb{R}^{d-1}} \mathbb{1}_{A_\gamma}(x) p(x_1, x_{2:d} \mid y)\mathrm{d}x_{2:d}$ as a function over $\mathbb{R}$ is L^1 integrable. Therefore, by the Riemann-Lebesgue lemma, it holds that

$$I_n = \int_{\mathbb{R}} \cos(2nx_1)\Big(\int_{\mathbb{R}^{d-1}} \mathbb{1}_{A_\gamma}(x) p(x_1, x_{2:d} \mid y)\mathrm{d}x_{2:d}\Big)\mathrm{d}x_1 \to 0 \qquad \text{as } n \to \infty,$$

hence for a sufficiently large n, one can take $|I_n| \le \frac{1}{2} p(A_\gamma \mid y)$ which yields

$$\mathbb{E}_{x \sim p(\cdot \mid y)}\left[\|\nabla_x \log p(y \mid x) - \nabla_x \log \hat{p}_n(y \mid x)\|_2^2\right] \ge \frac{(1-\gamma)^2 p(A_\gamma \mid y) \delta_n^2 n^2}{4}.$$

It then suffices to choose a sequence $\{\delta_n\}_{n=1}^\infty$ such that $\delta_n n \ge c > 0$ while $\delta_n \downarrow 0$ as $n \to \infty$. We present below two illustrative scaling regimes.

Choice of δ_n	Cross entropy loss	Guidance vector error
$\delta_n = 1/n$	$O(1/n)$	$\Theta(1)$
$\delta_n = 1/\sqrt{n}$	$O(1/\sqrt{n})$	$\Theta(n)$

Table 1: Two illustrative scaling regimes.

In both regimes the cross entropy loss vanishes, yet the guidance error does not converge to zero.

A.2 Proof of Theorem 3.3

For $t \in \{s_0, s_1, \cdots, s_{N-1}\}$, let $u_t(x) = \log p_t(y \mid x) - \log \hat{p}_t(y \mid x)$. We aim to bound $\mathbb{E}_{x \sim p_t(\cdot \mid y)}\left[\|\nabla_x u_t(x)\|_2^2\right]$. Note that

$$\begin{aligned}&\nabla_x \cdot (u_t(x) p_t(x \mid y) \nabla_x u_t(x)) \\ &= p_t(x \mid y)\|\nabla_x u_t(x)\|_2^2 + u_t(x)\langle \nabla_x p_t(x \mid y), \nabla_x u_t(x)\rangle + p_t(x \mid y) u_t(x) \mathrm{Tr}(\nabla_x^2 u_t(x)).\end{aligned}$$

Consider a closed ball $\mathcal{B}_\rho$ with radius $\rho > 0$. Rearranging and using the fact that $p_t(x \mid y)\nabla_x \log p_t(x \mid y) = \nabla_x p_t(x \mid y)$, we find that

$$\begin{aligned}\int_{\mathcal{B}_\rho} \|\nabla_x u_t(x)\|_2^2 p_t(x \mid y)\mathrm{d}x =& \int_{\mathcal{B}_\rho} \nabla_x \cdot (u_t(x) p_t(x \mid y) \nabla_x u_t(x))\mathrm{d}x \\ &- \int_{\mathcal{B}_\rho} p_t(x \mid y) u_t(x)[\langle \nabla_x \log p_t(x \mid y), \nabla_x u_t(x)\rangle + \mathrm{Tr}(\nabla_x^2 u_t(x))]\mathrm{d}x.\end{aligned}$$

Letting $\rho \to \infty$, by Lemmas B.2 and B.3, we find that

$$\mathbb{E}_{x\sim p_t(\cdot|y)}\left[\|\nabla_x u_t(x)\|_2^2\right] = \underbrace{-\int_{\mathbb{R}^d} p_t(x \mid y)u_t(x)\langle \nabla_x \log p_t(x \mid y), \nabla_x u_t(x)\rangle \mathrm{d}x}_{\mathrm{I}}$$
$$\underbrace{-\int_{\mathbb{R}^d} p_t(x \mid y)u_t(x)\mathrm{Tr}(\nabla_x^2 u_t(x))\mathrm{d}x}_{\mathrm{II}}.$$

We then upper bound terms I and II, separately. We first bound term I. To this end, we define the set

$$\mathcal{G}_t(\epsilon) = \Big\{\|\nabla_x \log p_t(x \mid y)\|_2 \le r_t(\epsilon),\ \|\nabla_x \log p_t(y \mid x)\|_2 \le r_t(\epsilon),\ \|\nabla_x \log \hat{p}_t(y \mid x)\|_2 \le r_t(\epsilon)\Big\},$$

where $r_t(\epsilon) = \frac{C\sqrt{d+\log(1/\epsilon)}}{\sigma_t}$ is taken to be the maximum of the upper bounds from Corollaries B.1, B.2, and the theorem condition 3(a). We then have $p_t(\mathcal{G}_t(\epsilon)^c \mid y) \le 3\epsilon/P_{\mathrm{data}}(y)$. Note that

$$\begin{aligned}
&\mathbb{E}_{x\sim p_t(\cdot|y)}\left[|u_t(x)|\|\nabla_x \log p_t(x \mid y)\|_2\|\nabla_x u_t(x)\|_2\right] \\
&= \mathbb{E}_{x\sim p_t(\cdot|y)}[|u_t(x)|\|\nabla_x \log p_t(x \mid y)\|_2\|\nabla_x u_t(x)\|_2 \mathbb{1}_{\mathcal{G}_t(\epsilon)}] \\
&\quad + \mathbb{E}_{x\sim p_t(\cdot|y)}[|u_t(x)|\|\nabla_x \log p_t(x \mid y)\|_2\|\nabla_x u_t(x)\|_2 \mathbb{1}_{\mathcal{G}_t(\epsilon)^c}].
\end{aligned}$$

On $\mathcal{G}_t(\epsilon)$, by Lemma B.7 we have

$$\begin{aligned}
&\mathbb{E}_{x\sim p_t(\cdot|y)}[|u_t(x)|\|\nabla_x \log p_t(x \mid y)\|_2\|\nabla_x u_t(x)\|_2 \mathbb{1}_{\mathcal{G}_t(\epsilon)}] \\
&\le \frac{2C\varepsilon_t r_t(\epsilon)^2}{P_{\mathrm{data}}(y)} \le \frac{C'\varepsilon_t(d+\log(1/\epsilon))}{P_{\mathrm{data}}(y)\sigma_t^2},
\end{aligned} \tag{8}$$

where $C, C' > 0$ are numerical constants. On $\mathcal{G}_t(\epsilon)^c$, by Lemma B.1 we have

$$\begin{aligned}
&\mathbb{E}_{x\sim p_t(\cdot|y)}[|u_t(x)|\|\nabla_x \log p_t(x \mid y)\|_2\|\nabla_x u_t(x)\|_2 \mathbb{1}_{\mathcal{G}_t(\epsilon)^c}] \\
&\le \frac{4\lambda_t R}{\sigma_t^2}\mathbb{E}_{x\sim p_t(\cdot|y)}[|u_t(x)|\|\nabla_x \log p_t(x \mid y)\|_2 \mathbb{1}_{\mathcal{G}_t(\epsilon)^c}].
\end{aligned} \tag{9}$$

Since P_{data} is supported on the ball centered at the origin with radius R, then there exists $b \le C''(R+\sqrt{d})$ for some positive numerical constant C'', such that $p_t(\mathcal{B}_b \mid y) \ge 1/2$. By Lemma B.7, we know that $\mathbb{E}_{x\sim p_t(\cdot|y)}[|u_t(x)|] \le C\varepsilon_t/P_{\mathrm{data}}(y)$. Therefore, there exists $x_* \in \mathcal{B}_b$, such that $|u_t(x_*)| \le 2C\varepsilon_t/P_{\mathrm{data}}(y)$. Using the fundamental theorem of calculus and Lemma B.1, we see that

$$\begin{aligned}
|u_t(x)| \le& |u_t(x_*)| + \int_0^1 |\langle \nabla u_t(rx + (1-r)x_*), x - x_*\rangle| \mathrm{d}r \le |u_t(x_*)| \\
&+ 4\lambda_t \sigma_t^{-2} R\|x - x_*\|_2.
\end{aligned} \tag{10}$$

By Eq. (20), we have

$$\|\nabla_x \log p_t(x \mid y)\|_2 = \|\sigma_t^{-2}(\lambda_t m_t(x,y) - x)\|_2 \le \sigma_t^{-2}\lambda_t R + \sigma_t^{-2}\|x\|_2.$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{x\sim p_t(\cdot|y)}[|u_t(x)|\|\nabla_x \log p_t(x \mid y)\|_2 \mathbb{1}_{\mathcal{G}_t(\epsilon)^c}] \\
\lesssim& \mathbb{E}_{x\sim p_t(\cdot|y)}\big[(\lambda_t^2\sigma_t^{-4}R^2\|x - x_*\|_2 + \lambda_t\sigma_t^{-4}R\|x\|_2\|x - x_*\|_2 + \sigma_t^{-2}\lambda_t R\varepsilon_t/P_{\mathrm{data}}(y) \\
&+ \sigma_t^{-2}\|x\|_2\varepsilon_t/P_{\mathrm{data}}(y))\mathbb{1}_{\mathcal{G}_t(\epsilon)^c}\big] \\
\lesssim& \mathbb{E}_{x\sim p_t(\cdot|y)}\Big[\lambda_t^4\sigma_t^{-8}R^4\|x - x_*\|_2^2 + \lambda_t^2\sigma_t^{-8}R^2\|x\|_2^2\|x - x_*\|_2^2 + \frac{\varepsilon_t^2\sigma_t^{-4}\lambda_t^2 R^2}{P_{\mathrm{data}}(y)^2} + \frac{\varepsilon_t^2\sigma_t^{-4}\|x\|_2^2}{P_{\mathrm{data}}(y)^2}\Big]^{1/2} \\
&\times p_t(\mathcal{G}_t(\epsilon)^c \mid y)^{1/2} \\
\lesssim& \frac{R(R^2+d)\epsilon^{1/2}}{\sigma_t^4 P_{\mathrm{data}}(y)^{1/2}} + \frac{(R^2+d)^{1/2}\varepsilon_t\epsilon^{1/2}}{\sigma_t^2 P_{\mathrm{data}}(y)^{3/2}},
\end{aligned} \tag{11}$$

where "$\lesssim$" hides numerical constants. Combining Eqs. (8), (9) and (11), we get

$$|\mathrm{I}| \lesssim \frac{\varepsilon_t(d+\log(1/\epsilon))}{P_{\mathrm{data}}(y)\sigma_t^2} + \frac{R^2(R^2+d)\epsilon^{1/2}}{\sigma_t^6 P_{\mathrm{data}}(y)^{1/2}} + \frac{R(R^2+d)^{1/2}\varepsilon_t\epsilon^{1/2}}{\sigma_t^4 P_{\mathrm{data}}(y)^{3/2}}. \tag{12}$$

We next bound term II. Similarly, we define the set

$$\begin{aligned}\mathcal{H}_t(\epsilon) = \Big\{ &\big|\mathrm{Tr}(\nabla_x^2 \log p_t(x \mid y))\big| \le R_t(\epsilon),\ \ \big|\mathrm{Tr}(\nabla_x^2 \log p_t(y \mid x))\big| \le R_t(\epsilon), \\ &\big|\mathrm{Tr}(\nabla_x^2 \log \hat{p}_t(y \mid x))\big| \le R_t(\epsilon)\Big\},\end{aligned}$$

where $R_t(\epsilon) = \frac{C'(d+\log(1/\epsilon))}{\sigma_t^2}$ is taken to be the maximum of the upper bound from Lemma B.5, Corollary B.2, and the theorem condition 3(b). In this case, $p_t(\mathcal{H}_t(\epsilon)^c \mid y) \le 3\epsilon/P_{\mathrm{data}}(y)$. On $\mathcal{H}_t(\epsilon)$, by Corollary B.2 and Lemma B.7,

$$\mathbb{E}_{x\sim p_t(\cdot|y)}[|u_t(x)| \cdot |\mathrm{Tr}(\nabla_x^2 u_t(x))| \mathbb{1}_{\mathcal{H}_t(\epsilon)}] \lesssim \frac{\varepsilon_t(d+\log(1/\epsilon))}{\sigma_t^2 P_{\mathrm{data}}(y)}.$$

On $\mathcal{H}_t(\epsilon)^c$, by Lemma B.1 and Eq. (10) we have

$$\begin{aligned}&\mathbb{E}_{x\sim p_t(\cdot|y)}[|u_t(x)| \cdot |\mathrm{Tr}(\nabla_x^2 u_t(x))| \mathbb{1}_{\mathcal{H}_t(\epsilon)^c}] \\ &\lesssim \frac{\lambda_t^2 R^2}{\sigma_t^4} \mathbb{E}_{x\sim p_t(\cdot|y)}\big[(|u_t(x_*)| + 4\lambda_t\sigma_t^{-2}R\|x-x_*\|_2)^2\big]^{1/2} \cdot \frac{\epsilon^{1/2}}{P_{\mathrm{data}}(y)^{1/2}} \\ &\lesssim \frac{R^2\epsilon^{1/2}\varepsilon_t}{\sigma_t^4 P_{\mathrm{data}}(y)^{3/2}} + \frac{R^2(R^2+d)\epsilon^{1/2}}{\sigma_t^6 P_{\mathrm{data}}(y)^{1/2}}.\end{aligned}$$

Next, we put together the above two equations, and conclude that

$$|\mathrm{II}| \lesssim \frac{\varepsilon_t(d+\log(1/\epsilon))}{\sigma_t^2 P_{\mathrm{data}}(y)} + \frac{R^2\epsilon^{1/2}\varepsilon_t}{\sigma_t^4 P_{\mathrm{data}}(y)^{3/2}} + \frac{R^2(R^2+d)\epsilon^{1/2}}{\sigma_t^6 P_{\mathrm{data}}(y)^{1/2}}. \tag{13}$$

Combining Eqs. (12) and (13), we conclude that

$$\mathbb{E}_{x\sim p_t(\cdot|y)}\left[\|\nabla_x u_t(x)\|_2^2\right] \lesssim \frac{\varepsilon_t(d+\log(1/\epsilon))}{P_{\mathrm{data}}(y)\sigma_t^2} + \frac{R^2(R^2+d)\epsilon^{1/2}}{\sigma_t^6 P_{\mathrm{data}}(y)^{1/2}} + \frac{R(R^2+d)^{1/2}\varepsilon_t\epsilon^{1/2}}{\sigma_t^4 P_{\mathrm{data}}(y)^{3/2}}.$$

Setting

$$\epsilon \ = \ \min\left\{ \frac{\sigma_t^8\varepsilon_t^2}{R^4(R^2+d)^2 P_{\mathrm{data}}(y)},\ \frac{\sigma_t^4 P_{\mathrm{data}}(y)}{R^2(R^2+d)} \right\},$$

we conclude that

$$\mathbb{E}_{x\sim p_t(\cdot|y)}\left[\|\nabla_x u_t(x)\|_2^2\right] \lesssim \frac{\varepsilon_t}{\sigma_t^2 P_{\mathrm{data}}(y)}\Big(d + \log\Big(\frac{R^2+d}{\sigma_t^2}\Big) + \log\Big(\frac{1}{\varepsilon_t}\Big) + \log\Big(\frac{1}{P_{\mathrm{data}}(y)}\Big)\Big).$$

The proof is done.

A.3 Proof of Lemma 3.1

By Theorem 3.3, we see that

$$\varepsilon_{\mathrm{guide}}^2 \lesssim \sum_{k=0}^{N-1} \tau_k \cdot \frac{\varepsilon_{s_k}}{\sigma_{s_k}^2 P_{\mathrm{data}}(y)}\Big(d + \log\Big(\frac{R^2+d}{\sigma_{s_k}^2}\Big) + \log\Big(\frac{1}{\varepsilon_{s_k}}\Big) + \log\Big(\frac{1}{P_{\mathrm{data}}(y)}\Big)\Big),$$

where "$\lesssim$" hides a numerical constant. When $\sigma_{s_k}^2 \in [1/2, 1)$, we have $\tau_k \le \kappa$, hence

$$\begin{aligned}&\tau_k \cdot \frac{\varepsilon_{s_k}}{\sigma_{s_k}^2 P_{\mathrm{data}}(y)}\Big(d + \log\Big(\frac{R^2+d}{\sigma_{s_k}^2}\Big) + \log\Big(\frac{1}{\varepsilon_{s_k}}\Big) + \log\Big(\frac{1}{P_{\mathrm{data}}(y)}\Big)\Big) \\ &\lesssim \frac{\kappa\varepsilon_{s_k}}{P_{\mathrm{data}}(y)}\Big(d + \log\Big(\frac{R^2+d}{\delta}\Big) + \log\Big(\frac{1}{\varepsilon_{s_k}}\Big) + \log\Big(\frac{1}{P_{\mathrm{data}}(y)}\Big)\Big).\end{aligned}$$

On the other hand, when $\sigma_{s_k}^2 \in (0, 1/2)$, we have $\tau_k \le \kappa(T - t_{k+1}) \le \kappa s_k$ and $\sigma_{s_k}^{-2} \lesssim s_k^{-1}$. Therefore,

$$\begin{aligned}
&\tau_k \cdot \frac{\varepsilon_{s_k}}{\sigma_{s_k}^2 P_{\mathrm{data}}(y)}\Big(d + \log\Big(\frac{R^2+d}{\sigma_{s_k}^2}\Big) + \log\Big(\frac{1}{\varepsilon_{s_k}}\Big) + \log\Big(\frac{1}{P_{\mathrm{data}}(y)}\Big)\Big) \\
&\lesssim \tau_k \cdot \frac{\varepsilon_{s_k}}{s_k P_{\mathrm{data}}(y)}\Big(d + \log\Big(\frac{R^2+d}{\sigma_{s_k}^2}\Big) + \log\Big(\frac{1}{\varepsilon_{s_k}}\Big) + \log\Big(\frac{1}{P_{\mathrm{data}}(y)}\Big)\Big) \\
&\lesssim \frac{\kappa\varepsilon_{s_k}}{P_{\mathrm{data}}(y)}\Big(d + \log\Big(\frac{R^2+d}{\delta}\Big) + \log\Big(\frac{1}{\varepsilon_{s_k}}\Big) + \log\Big(\frac{1}{P_{\mathrm{data}}(y)}\Big)\Big).
\end{aligned}$$

We complete the proof by combining these two upper bounds.

A.4 Proof of Theorem 3.6

We first construct a one-dimensional example and then extend it to $d > 1$.

Data distribution. Let the label space be $\mathcal{Y} = \{\pm 1\}$, and take the uniform label prior $P_{\mathrm{data}}(Y = 1) = P_{\mathrm{data}}(Y = -1) = 1/2$. Without loss, we assume that the guidance is towards label 1. Let $X_0 \in \mathbb{R}$ be a random variable that is independent of Y with distribution $X_0 \sim \mathrm{Unif}([-R, R])$ for some fixed $R > 2$. Let $(X_t)_{t\in[0,T]}$ be the OU process starting at X_0 (recall this is defined in Eq. (1)). For any $t > 0$, X_t has a smooth density p_t over $\mathbb{R}$ given by a Gaussian convolution of a compactly-supported density. In particular, for $t > 0$:

(1) p_t is symmetric about 0 since X_0 is symmetric and the OU kernel is Gaussian and also symmetric;

(2) $p_t(x) > 0$ for all $x \in \mathbb{R}$ and is continuous, so p_t is bounded away from zero on any compact interval.

Because Y and X_0 are independent, the OU process is also independent of the label. As a consequence, for every $t \in [0, T]$ and $x \in \mathbb{R}$,

$$p_t(Y = 1 \mid X_0 = x) = \frac{1}{2}, \qquad p_t(Y = -1 \mid X_0 = x) = \frac{1}{2}.$$

Thus, we conclude that:

- Assumption (1) holds: P_{data} is supported on the compact set $\mathcal{K} = [-R, R]$ with $R < \infty$.
- Assumption (4) holds with $P_{\mathrm{data}}(y) = 1/2$.

Moreover, $\log p_t(y \mid x)$ is constant in x, so $\frac{\mathrm{d}}{\mathrm{d}x} \log p_t(y \mid x) \equiv 0$ and $\frac{\mathrm{d}^2}{\mathrm{d}x^2} \log p_t(y \mid x) \equiv 0$.

Classifier construction. Fix $t_0 \in \{s_k : k \in \{0\} \cup [N-1]\}$. As noted above, p_{t_0} is smooth, symmetric and strictly positive on $\mathbb{R}$. For a sufficiently small $\varepsilon > 0$, define

$$\alpha_\varepsilon(x) = \varepsilon \sin\Big(\frac{x}{\sqrt{\varepsilon}}\Big), \qquad s_\varepsilon(x) = 2\alpha_\varepsilon(x).$$

Taking the first- and second-order derivatives, we get

$$\alpha_\varepsilon'(x) = \sqrt{\varepsilon}\,\cos\Big(\frac{x}{\sqrt{\varepsilon}}\Big), \qquad \alpha_\varepsilon''(x) = -\sin\Big(\frac{x}{\sqrt{\varepsilon}}\Big),$$

hence

$$\sup_{x\in\mathbb{R}} |\alpha_\varepsilon'(x)| \le \sqrt{\varepsilon} \le 1, \qquad \sup_{x\in\mathbb{R}} |\alpha_\varepsilon''(x)| \le 1$$

for all sufficiently small ε. This implies that α_ε has uniformly bounded first- and second-order derivatives.

Define a probabilistic classifier at time t_0 by

$$\hat{p}_{\varepsilon,t_0}(y = 1 \mid x) = \sigma\big(s_\varepsilon(x)\big), \qquad \hat{p}_{\varepsilon,t_0}(y = -1 \mid x) = 1 - \hat{p}_{\varepsilon,t_0}(y = 1 \mid x),$$

where $\sigma(z) = (1 + e^{-z})^{-1}$ is the logistic function. Note that the map $x \mapsto \hat{p}_{\varepsilon,t_0}(y \mid x)$ belongs to $C^\infty(\mathbb{R})$.

Verifying classifier smoothness. We then verify the classifier smoothness assumption. Note that

$$\log \hat{p}_{\varepsilon,t_0}(y=1 \mid x) = \log \sigma(s_\varepsilon(x)).$$

Differentiating and using the equality $\frac{d}{ds}\log\sigma(s) = 1-\sigma(s)$, we get

$$\frac{\mathrm{d}}{\mathrm{d}x}\log \hat{p}_{\varepsilon,t_0}(y=1 \mid x) = (1-\sigma(s_\varepsilon(x)))\, s'_\varepsilon(x).$$

Since $|s'_\varepsilon(x)| = 2|\alpha'_\varepsilon(x)| \le 2$ and $0 < 1-\sigma(s_\varepsilon(x)) < 1$, we have

$$\sup_{x\in\mathbb{R},\varepsilon\in(0,1/2),t_0\in\{s_k:k\in\{0\}\cup[N-1]\}} \Big|\frac{\mathrm{d}}{\mathrm{d}x}\log \hat{p}_{\varepsilon,t_0}(y=1 \mid x)\Big| \le 2.$$

Computing the second-order derivative gives

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\log \hat{p}_{\varepsilon,t_0}(y=1 \mid x) = (1-\sigma(s_\varepsilon(x)))\, s''_\varepsilon(x) - \sigma(s_\varepsilon(x))(1-\sigma(s_\varepsilon(x)))\,[s'_\varepsilon(x)]^2.$$

Since $|s''_\varepsilon(x)| = 2|\alpha''_\varepsilon(x)| \le 2$, $|s'_\varepsilon(x)| \le 2$ and $\sigma(s)(1-\sigma(s)) \le 1/4$, we get a uniform upper bound

$$\sup_{x\in\mathbb{R},\varepsilon\in(0,1/2),t_0\in\{s_k:k\in\{0\}\cup[N-1]\}} \Big|\frac{\mathrm{d}^2}{\mathrm{d}x^2}\log \hat{p}_{\varepsilon,t_0}(y=1 \mid x)\Big| \le 3.$$

Hence, the estimated probabilities have uniformly bounded first- and second-order derivatives, independent of ε, as claimed. Therefore,

- Assumptions (2) and (3) hold.

Scaling of the conditional KL. We now compute the conditional KL divergence achieved by $\hat{p}_{\varepsilon,t_0}$. For each $x \in \mathbb{R}$, the true label conditional distribution is $\mathrm{Bernoulli}(1/2)$, while the classifier $\hat{p}_{\varepsilon,t_0}$ yields a conditional distribution $\mathrm{Bernoulli}(\sigma(s_\varepsilon(x)))$. A Taylor expansion around 0 implies that, for sufficiently small s,

$$D_{\mathrm{KL}}\big(\mathrm{Bern}(1/2)\,\|\,\mathrm{Bern}(\sigma(s))\big) \le c\, s^2$$

for some numerical constant $c > 0$. Since $|s_\varepsilon(x)| \le 2\varepsilon$, hence for a sufficiently small ε,

$$D_{\mathrm{KL}}\big(p_{t_0}(\cdot \mid x)\,\|\,\hat{p}_{\varepsilon,t_0}(\cdot \mid x)\big) \le c\, s_\varepsilon(x)^2.$$

Taking the expectation over $X_{t_0} \sim p_{t_0}(\cdot)$ yields the following:

$$\mathbb{E}_{X_{t_0}\sim p_{t_0}(\cdot)}\Big[D_{\mathrm{KL}}\big(p_{t_0}(\cdot \mid X_{t_0})\,\|\,\hat{p}_{\varepsilon,t_0}(\cdot \mid X_{t_0})\big)\Big] \le c\,\mathbb{E}[s_\varepsilon(X_{t_0})^2] \le 4c\varepsilon^2.$$

This bound is independent of t_0, hence

- Assumption (5) holds and requirement (i) is satisfied.

Scaling of the guidance term MSE. We then analyze the guidance term MSE at time t_0. Note that

$$\log p_{t_0}(y=1 \mid x) \equiv \log\tfrac{1}{2} \quad\Longrightarrow\quad \frac{\mathrm{d}}{\mathrm{d}x}\log p_{t_0}(y=1 \mid x) = 0.$$

Since $|s_\varepsilon(x)| \le 2\varepsilon$, there exist absolute constants $0 < c_1 < c_2 < 1$, such that for all small enough ε,

$$c_1 \;\le\; 1-\sigma(s_\varepsilon(x)) \;\le\; c_2 \qquad \text{for all } x \in \mathbb{R}.$$

As a consequence,

$$\Big|\frac{\mathrm{d}}{\mathrm{d}x}\log p_{t_0}(y=1 \mid x) - \frac{\mathrm{d}}{\mathrm{d}x}\log \hat{p}_{\varepsilon,t_0}(y=1 \mid x)\Big| = \big|(1-\sigma(s_\varepsilon(x)))\, s'_\varepsilon(x)\big| \ge c_1|s'_\varepsilon(x)|.$$

We next compute the MSE under $X_{t_0} \sim p_{t_0}(\cdot \mid y=1)$. Recall that X_{t_0} is independent of Y, therefore,

$$\mathbb{E}_{X_{t_0}\sim p_{t_0}(\cdot\mid y=1)}\big[\|\frac{\mathrm{d}}{\mathrm{d}x}\log p_{t_0}(y=1 \mid X_{t_0}) - \frac{\mathrm{d}}{\mathrm{d}x}\log \hat{p}_{\varepsilon,t_0}(y=1 \mid X_{t_0})\|_2^2\big] \ge c_1^2\,\mathbb{E}_{X_{t_0}\sim p_{t_0}}\big[s'_\varepsilon(X_{t_0})^2\big].$$

Using the equality $\cos^2(z) = \frac{1}{2}(1+\cos(2z))$ and the Riemann-Lebesgue lemma, we have

$$\mathbb{E}_{X_{t_0}\sim p_{t_0}}\Big[\cos^2\Big(\frac{X_{t_0}}{\sqrt{\varepsilon}}\Big)\Big] \to \frac{1}{2}, \qquad \varepsilon\to 0 \qquad\Longrightarrow\qquad \frac{1}{\varepsilon}\mathbb{E}_{X_{t_0}\sim p_{t_0}}\big[s'_\varepsilon(X_{t_0})^2\big] \to 2, \qquad \varepsilon\to 0.$$

This verifies requirement (ii).

Extension to $d > 1$. For $d > 1$, take X_0 to be the uniform distribution over the cube $[-R, R]^d$ that is independent of Y. For $x \in \mathbb{R}^d$, define

$$\alpha_\varepsilon(x) = \varepsilon \sin\Big(\frac{x_1}{\sqrt{\varepsilon}}\Big), \qquad s_\varepsilon(x) = 2\alpha_\varepsilon(x),$$

which depend only on the first coordinate of x. Similarly, one can verify that by defining $\hat{p}_{\varepsilon,t_0} = \log \sigma(s_\varepsilon(x))$, all assumptions (1)–(5) and requirements (i) and (ii) continue to hold.

A.5 Proof of Theorem 3.7

Recall the γ-tilted distribution $p_{T,\gamma}$ from (6). Since the map $u \mapsto u^\gamma$ is convex for $\gamma > 1$, Jensen's inequality gives $Z_{T,\gamma} \geq \mathbb{E}_{p_T}[p_T(y \mid X)]^\gamma = P^\gamma_{\text{data}}$. We define the following path measures:

$$\mathbb{Q}_1 = \text{Law}\begin{cases} X_0 \sim p_{T,1} \\ \mathrm{d}\overleftarrow{X}^{(y)}_t = \Big\{\overleftarrow{X}^{(y)}_t + 2\nabla_x \log p_{T-t}(\overleftarrow{X}^{(y)}_t) + 2\nabla_x \log p_{T-t}(y \mid \overleftarrow{X}^{(y)}_t)\Big\}\mathrm{d}t + \sqrt{2}\,\mathrm{d}B'_t, \end{cases}$$

$$\mathbb{Q}_\gamma = \text{Law}\begin{cases} X_0 \sim p_{T,\gamma} \\ \mathrm{d}\overleftarrow{X}^{(y)}_t = \Big\{\overleftarrow{X}^{(y)}_t + 2\nabla_x \log p_{T-t}(\overleftarrow{X}^{(y)}_t) + 2\gamma\nabla_x \log p_{T-t}(y \mid \overleftarrow{X}^{(y)}_t)\Big\}\mathrm{d}t + \sqrt{2}\,\mathrm{d}B'_t. \end{cases}$$

We now construct a new path measure whose initial distribution satisfies $X_0 \sim p_{T,\gamma}$. To this end, we denote $H_t := p_{T-t}(y \mid x)$ and $G_t := \nabla_x \log p_{T-t}(y \mid x)$. Consider the following candidate change-of-measure process:

$$L_t := \frac{p_{T,\gamma}(X_0)}{p_{T,1}(X_0)} \exp\left\{\sqrt{2}(\gamma-1)\int_0^t \langle G_s, \mathrm{d}B'_s\rangle - (\gamma-1)^2 \int_0^t \|G_s\|^2 \mathrm{d}s\right\}. \tag{14}$$

Under $\mathbb{Q}_1$, the OU Fokker–Planck equation imply that H_t satisfies $-\partial_t H_t = (x + 2\nabla_x \log p_{T-t}(x))^\top \nabla_x H_t + \Delta_x H_t$. A simple application of Ito's lemma with Fokker-Planck yields

$$d\log H_t = \|G_t\|^2 \mathrm{d}t + \sqrt{2}\langle G_t, \mathrm{d}B'_t\rangle \iff \sqrt{2}\int_0^t \langle G_s, \mathrm{d}B'_s\rangle = \log\frac{H_t}{H_0} - \int_0^t \|G_s\|^2 \mathrm{d}s. \tag{15}$$

Using (15) with (14) gives us:

$$L_t := \frac{p_T(y \mid X_0)^{\gamma-1} Z_{T,1}}{Z_{T,\gamma}} \left(\frac{p_0(y \mid X_T)}{p_T(y \mid X_0)}\right)^{\gamma-1} \exp\left\{-\gamma(\gamma-1)\int_0^T \|G_s\|^2 \mathrm{d}s\right\}. \tag{16}$$

Assuming local square-integrability of G_t, the stochastic exponential in (16) is a nonnegative local martingale under $\mathbb{Q}_1$. Since the factor $p_{T,\gamma}(X_0)/p_{T,1}(X_0)$ is bounded and $\mathcal{F}_0$-measurable, L_t is also a nonnegative local martingale. Moreover, by (16) and the fact that $0 \leq H_t \leq 1$, $L_t \leq 1/P^\gamma_{\text{data}}$. Hence L_t is a bounded local martingale and therefore a uniformly integrable martingale. Since $\mathbb{E}_{\mathbb{Q}_1}[L_0] = 1$, it follows that $\mathbb{E}_{\mathbb{Q}_1}[L_t] = 1$ for all $t \leq T - \delta$. We have thus established sufficient conditions for Girsanov's theorem to hold (cf. Theorem 8 of [11]).

We now define a path measure $\widetilde{\mathbb{Q}}_\gamma$ through $\frac{\mathrm{d}\widetilde{\mathbb{Q}}_\gamma}{\mathrm{d}\mathbb{Q}_1} = L_t$. By Girsanov's theorem, under $\widetilde{\mathbb{Q}}_\gamma$, the process $B^\gamma_t = B'_t - \sqrt{2}(\gamma-1)\int_0^t G_s\,\mathrm{d}s$ is a $\widetilde{\mathbb{Q}}_\gamma$-Brownian motion. Therefore, we have $\mathrm{d}B'_t = \mathrm{d}B^\gamma_t + \sqrt{2}(\gamma-1)G_t\,\mathrm{d}t$. Plugging this into the SDE under $\mathbb{Q}_1$ gives us:

$$\mathrm{d}\overleftarrow{X}^{(y)}_t = \Big\{\overleftarrow{X}^{(y)}_t + 2\nabla_x \log p_{T-t}(\overleftarrow{X}^{(y)}_t) + 2\gamma\nabla_x \log p_{T-t}(y \mid \overleftarrow{X}^{(y)}_t)\Big\}\mathrm{d}t + \sqrt{2}\,\mathrm{d}B^\gamma_t. \tag{17}$$

Further notice that for any Borel measurable $A \subseteq \mathbb{R}^d$, if $X_0 \sim p_{T,1}$ then:

$$\widetilde{\mathbb{Q}}(X_0 \in A) = \mathbb{E}_{\mathbb{Q}_1}\left[L_t \mathbb{1}_{X_0 \in A}\right] = \mathbb{E}_{\mathbb{Q}_1}\left[\frac{p_{T,\gamma}(X_0)}{p_{T,1}(X_0)}\mathbb{1}_{X_0\in A}\right] = \int_A p_{T,\gamma}(x)\mathrm{d}x. \tag{18}$$

Thus $X_0 \sim p_{T,\gamma}$. Notably, this means the path measure $\widetilde{\mathbb{Q}}_\gamma$ is defined as follows:

$$\widetilde{\mathbb{Q}}_\gamma = \text{Law}\begin{cases} X_0 \sim p_{T,\gamma} \\ \mathrm{d}\overleftarrow{X}^{(y)}_t = \Big\{\overleftarrow{X}^{(y)}_t + 2\nabla_x \log p_{T-t}(\overleftarrow{X}^{(y)}_t) + 2\gamma\nabla_x \log p_{T-t}(y \mid \overleftarrow{X}^{(y)}_t)\Big\}\mathrm{d}t + \sqrt{2}\,\mathrm{d}B^\gamma_t. \end{cases} \tag{19}$$

Since $\delta > 0$, OU smoothing and the bounded-support assumption imply that the drift of the ideal guided reverse SDE is globally Lipschitz in x with at most linear growth uniformly on $[0, T-\delta]$. Thus by Theorem 5.2.1 of [49], the SDE admits a unique strong solution and is pathwise unique. In particular, uniqueness in law holds so $\widetilde{\mathbb{Q}}_\gamma = \mathbb{Q}_\gamma$.

Recall that $q_{\gamma,t}$ denotes the marginal distribution at reverse time t of the ideal classifier-guided process (7) with guidance strength γ and initialization $X_0 \sim p_{T,\gamma}$. In particular, for $\gamma = 1$, we have $q_{1,t} = p_{T-t}(\cdot \mid y)$. Following the same change-of-measure argument as in [2], we have

$$
\begin{aligned}
D_{\mathrm{KL}}\big(q_{\gamma,T-\delta} \,\|\, \hat{q}_{\gamma,T-\delta}\big) \lesssim\; & D_{\mathrm{KL}}\big(p_{T,\gamma} \,\|\, \pi_d\big) + \sum_{k=0}^{N-1} \tau_k \mathbb{E}_{q_{\gamma,t_k}} \left[\|\nabla \log p_{T-t_k}(X) - s_\theta(X, T-t_k)\|_2^2 \right] \\
& + \gamma^2 \sum_{k=0}^{N-1} \tau_k \mathbb{E}_{q_{\gamma,t_k}} \left[\|\nabla \log p_{T-t_k}(y \mid X) - \nabla \log \hat{p}_{T-t_k}(y \mid X)\|_2^2 \right] \\
& + \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_{q_{\gamma,t}} \left[\left\| \nabla_x \log \frac{p_{T-t}(x)}{p_{T-t_k}(x)} + \gamma \nabla_x \log \frac{p_{T-t}(y \mid x)}{p_{T-t_k}(y \mid x)} \right\|_2^2 \right] \mathrm{d}t \\
\le\; & \underbrace{D_{\mathrm{KL}}\big(p_{T,\gamma} \,\|\, \pi_d\big)}_{\text{(i)}} + \underbrace{\frac{1}{P_{\mathrm{data}}(y)^\gamma} \sum_{k=0}^{N-1} \tau_k \mathbb{E}_{q_{1,t_k}} \left[\|\nabla \log p_{T-t_k}(X) - s_\theta(X, T-t_k)\|_2^2 \right]}_{\text{(ii)}} \\
& + \underbrace{\frac{\gamma^2}{P_{\mathrm{data}}(y)^\gamma} \sum_{k=0}^{N-1} \tau_k \mathbb{E}_{q_{1,t_k}} \left[\|\nabla \log p_{T-t_k}(y \mid X) - \nabla \log \hat{p}_{T-t_k}(y \mid X)\|_2^2 \right]}_{\text{(iii)}} \\
& + \underbrace{\frac{1}{P_{\mathrm{data}}(y)^\gamma} \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_{q_{1,t}} \left[\left\| \nabla_x \log \frac{p_{T-t}(x)}{p_{T-t_k}(x)} + \gamma \nabla_x \log \frac{p_{T-t}(y \mid x)}{p_{T-t_k}(y \mid x)} \right\|_2^2 \right] \mathrm{d}t}_{\text{(iv)}} .
\end{aligned}
$$

where the second inequality follows from the uniform bound on the Radon-Nikodym derivative (16). Term (i) is the initialization error, (ii) is the score estimation error, (iii) is the guidance error, and (iv) is the discretization error. It therefore suffices to control terms (i) and (iv).

Following [2, Lemma 2], Lemma B.8 controls term (iv), while Lemma B.9 controls term (i). Under Assumption 3.1, and recalling that $\varepsilon^2_{\mathrm{guide}}$ is defined in Lemma 3.1, terms (ii) and (iii) are controlled by $\varepsilon^2_{\mathrm{score}}$ and $\varepsilon^2_{\mathrm{guide}}$, respectively. Combining these bounds concludes the proof.

B Supporting lemmas

We summarize in this section the technical lemmas required for our proof.

Lemma B.1. *Suppose P_{data} is supported on a bounded set $\mathcal{K} \subseteq \mathbb{R}^d$ with $R = \sup_{x \in \mathcal{K}} \|x\|_2 < \infty$. Recall that p_t is the marginal distribution of process* (1)*. Then the following statements hold:*

1. $\|\nabla_x \log p_t(y \mid x)\|_2 \le 2\lambda_t \sigma_t^{-2} R$
2. $|\mathrm{Tr}(\nabla_x^2 \log p_t(y \mid x))| \le \lambda_t^2 \sigma_t^{-4} R^2$

Proof. We prove Lemma B.1 in Appendix B.1. □

Lemma B.2. *Let $\mathcal{B}_\rho = \{x \in \mathbb{R}^d : \|x\|_2 \le \rho\}$. Recall that p_t is the distribution of $\overrightarrow{X}_t$ in process* (1)*, and $u_t(x) = \log p_t(y \mid x) - \log \hat{p}_t(y \mid x)$. Then under the assumptions of Theorem 3.3, for $t \in \{s_k : k \in \{0\} \cup [N-1]\}$ we have*

$$
\lim_{\rho \to \infty} \int_{\mathcal{B}_\rho} \nabla_x \cdot (u_t(x) p_t(x \mid y) \nabla_x u_t(x)) \, \mathrm{d}x = 0.
$$

Proof. We prove Lemma B.2 in Appendix B.2. □

Lemma B.3. *Under the assumptions of Theorem 3.3, the following statements hold for $t \in \{s_k : k \in \{0\} \cup [N-1]\}$:*

1. $\lim_{\rho\to\infty} \int_{\mathcal{B}_\rho} p_t(x \mid y)\|\nabla_x u_t(x)\|_2^2 \mathrm{d}x = \int_{\mathbb{R}^d} p_t(x \mid y)\|\nabla_x u_t(x)\|_2^2 \mathrm{d}x$.
2. *We define* $g_t(x) = u_t(x)p_t(x \mid y)[\langle \nabla_x \log p_t(x \mid y), \nabla_x u_t(x)\rangle + \mathrm{Tr}(\nabla_x^2 u_t(x))]$. *Then, it holds that*

$$\lim_{\rho\to\infty} \int_{\mathcal{B}_\rho} g_t(x)\mathrm{d}x = \int_{\mathbb{R}^d} g_t(x)\mathrm{d}x.$$

Proof. We prove Lemma B.3 in Appendix B.3. □

Lemma B.4. *Let X be a random vector in $\mathbb{R}^d$ defined on a probability space $(\Omega, \mathcal{T}, P)$, and let $\mathcal{F} \subseteq \mathcal{T}$ be a sub σ-algebra. If X is sub-Gaussian, then*

$$\mathbb{P}\left(\|\mathbb{E}[X \mid \mathcal{F}]\|_2 \geq 2\|X\|_{\psi_2}\sqrt{\log\left(\frac{2\cdot 5^d}{\epsilon}\right)}\right) \leq \epsilon,$$

where $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm (see [60]).

Proof. We prove Lemma B.4 in Appendix B.4. □

We apply Lemma B.4 to derive high-probability bounds on the gradients and Hessians, as stated in the following two corollaries.

Corollary B.1. *Recall that p_t is the law of $\overrightarrow{X}_t$ in the forward diffusion process* (1)*, then there exists a universal constant C, such that*

$$\mathbb{P}_{x\sim p_t(\cdot|y)}\left(\|\nabla_x \log p_t(x \mid y)\|_2 \leq \frac{C\sqrt{d+\log(1/\epsilon)}}{\sigma_t}\right) \geq 1-\epsilon.$$

Proof. We prove Corollary B.1 in Appendix B.5. □

Lemma B.5. *There exists a universal constant $C > 0$, such that for $t > 0$,*

$$\mathbb{P}_{x\sim p_t(\cdot|y)}\left(\left|\mathrm{Tr}\big(\nabla_x^2 \log p_t(x \mid y)\big)\right| \leq \frac{C(d+\log(1/\epsilon))}{\sigma_t^2}\right) \geq 1-\epsilon.$$

Proof. We prove Lemma B.5 in Appendix B.6. □

Similarly, we have the following results.

Corollary B.2. *There exists a universal constant $C > 0$, such that for $t > 0$,*

1. $\mathbb{P}_{x\sim p_t(\cdot|y)}\left(\|\nabla_x \log p_t(y \mid x)\|_2 \leq \frac{C\sqrt{d+\log(1/\epsilon)}}{\sigma_t}\right) \geq 1-\epsilon/P_{\mathrm{data}}(y)$,
2. $\mathbb{P}_{x\sim p_t(\cdot|y)}\left(\left|\mathrm{Tr}\big(\nabla_x^2 \log p_t(y \mid x)\big)\right| \leq \frac{C(d+\log(1/\epsilon))}{\sigma_t^2}\right) \geq 1-\epsilon/P_{\mathrm{data}}(y)$.

Proof. We prove Corollary B.2 in Appendix B.7. □

Lemma B.6. *For $t > 0$, we assume that*

$$\mathbb{E}_{x\sim p_t}\Big[D_{\mathrm{KL}}\big(p_t(\cdot \mid x)\,\|\,\hat{p}_t(\cdot \mid x)\big)\Big] \;\leq\; \varepsilon^2.$$

The probabilistic classifier $\hat{p}_t$ induces a conditional distribution over the label space given the covariates. Using the same notation, we define the corresponding joint distribution over (x, y) by $\hat{p}_t(x, y) = p_t(x)\hat{p}_t(y \mid x)$. Then for every label y with $P_{\mathrm{data}}(y) > 0$, it holds that

$$D_{\mathrm{KL}}\big(p_t(\cdot \mid y)\,\|\,\hat{p}_t(\cdot \mid y)\big) \;\leq\; \frac{\varepsilon^2}{P_{\mathrm{data}}(y)}.$$

Proof. We prove Lemma B.6 in Appendix B.8. □

Lemma B.7. *Recall that $u_t(x) = \log p_t(y \mid x) - \log \hat{p}_t(y \mid x)$. Under the conditions of Theorem 3.3, for $t \in \{s_k : k \in \{0\} \cup [N-1]\}$, there exists a numerical constant $C > 0$, such that*

$$\mathbb{E}_{x \sim p_t(\cdot|y)}[|u_t(x)|] = \int_{\mathbb{R}^d} p_t(x \mid y)|u_t(x)|\mathrm{d}x \le \frac{C\varepsilon_t}{P_{\mathrm{data}}(y)}.$$

Proof. We prove Lemma B.7 in Appendix B.9 □

Lemma B.8. *Under the assumptions of Theorem 3.7, we have*

$$\begin{aligned}
&\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_{x \sim p_{T-t}(\cdot|y)} \left[\left\| \nabla_x \log \frac{p_{T-t}(x)}{p_{T-t_k}(x)} + \gamma \nabla_x \log \frac{p_{T-t}(y \mid x)}{p_{T-t_k}(y \mid x)} \right\|_2^2 \right] \mathrm{d}t \\
&\lesssim (2\gamma^2 - 2\gamma + 1)\Big(\kappa^2(d + R^2)N + \kappa(d + R^2)T\Big),
\end{aligned}$$

where "$\lesssim$" hides a positive numerical constant.

Proof of Lemma B.8. We prove Lemma B.8 in Appendix B.10. □

Lemma B.9. *Under the assumptions of Theorem 3.7, for $T \ge 1$ and $\gamma > 1$, we have*

$$D_{\mathrm{KL}}\big(p_{T,\gamma} \,\|\, \pi_d\big) \lesssim \left(d + \frac{\gamma^2 R^2}{P_{\mathrm{data}}(y)^{2\gamma}}\right) e^{-2T},$$

where "$\lesssim$" hides a positive numerical constant.

Proof of Lemma B.9. We prove Lemma B.9 in Appendix B.11. □

B.1 Proof of Lemma B.1

Recall p_t represents the joint distribution of $(Y, \overrightarrow{X}_t)$ (see Section 2.2). With process (1), we define

$$m_t(x) = \mathbb{E}[\overrightarrow{X}_0 \mid \overrightarrow{X}_t = x], \quad m_t(x, y) = \mathbb{E}[\overrightarrow{X}_0 \mid \overrightarrow{X}_t = x, Y = y].$$

Note that $\overrightarrow{X}_t \mid \overrightarrow{X}_0 = x_0 \sim \mathcal{N}(\lambda_t x_0, \sigma_t^2 I_d)$, hence we have

$$\nabla_x p_t(x \mid x_0) = -\left(\frac{x - \lambda_t x_0}{\sigma_t^2}\right) p_t(x \mid x_0).$$

Note that $p_t(x, y) = \int_{\mathbb{R}^d} p_t(x \mid x_0) p(x_0, y) \mathrm{d}x_0$, hence (one can check that differentiation under the integral sign is valid)

$$\begin{aligned}
\nabla_x p_t(x, y) &= \int_{\mathbb{R}^d} \nabla_x p_t(x \mid x_0) p_0(x_0, y) \mathrm{d}x_0 \\
&= \frac{1}{\sigma_t^2} \int_{\mathbb{R}^d} (\lambda_t x_0 - x) p_t(x \mid x_0) p_0(x_0, y) \mathrm{d}x_0 \\
&= \sigma_t^{-2} p_t(x, y) \left(\lambda_t m_t(x, y) - x\right).
\end{aligned} \tag{20}$$

Thus $\nabla_x \log p_t(x, y) = \sigma_t^{-2} (\lambda_t m_t(x, y) - x)$. Similarly, $\nabla_x \log p_t(x) = \sigma_t^{-2} (\lambda_t m_t(x) - x)$. Therefore,

$$\begin{aligned}
\nabla_x \log p_t(y \mid x) &= \nabla_x \log p_t(x, y) - \nabla_x \log p_t(x) \\
&= \sigma_t^{-2} (\lambda_t m_t(x, y) - x) - \sigma_t^{-2} (\lambda_t m_t(x) - x) \\
&= \lambda_t \sigma_t^{-2} (m_t(x, y) - m_t(x)).
\end{aligned}$$

By assumption, P_{data} is supported on a bounded set $\mathcal{K} \subseteq \mathbb{R}^d$ with $R = \sup_{x \in \mathcal{K}} \|x\|_2 < \infty$, hence $\|m_t(x_t, y)\|_2, \|m_t(x_t)\|_2 \le R$. Using these upper bounds, we establish via triangle inequality that $\|\nabla_x \log p_t(y \mid x)\|_2 \le 2\lambda_t \sigma_t^{-2} R$. This proves the first point of the lemma.

Next, we compute the Hessian of $p_t(x, y)$ with respect to x. Denote $\Sigma_t(x) = \mathrm{Cov}[\overrightarrow{X}_0 \mid \overrightarrow{X}_t = x]$ and $\Sigma_t(x, y) = \mathrm{Cov}[\overrightarrow{X}_0 \mid \overrightarrow{X}_t = x, Y = y]$. Under the current regularity assumptions, we have

$$
\begin{aligned}
&\nabla_x^2 p_t(x, y) \\
&= \int_{\mathbb{R}^d} \nabla_x^2 p_t(x \mid x_0)\, p_0(x_0, y)\, \mathrm{d}x_0 \\
&= \sigma_t^{-4} \int_{\mathbb{R}^d} \Big[(x - \lambda_t x_0)(x - \lambda_t x_0)^\top - \sigma_t^2 I_d\Big] p_t(x \mid x_0)\, p_0(x_0, y)\, \mathrm{d}x_0 \\
&= \sigma_t^{-4} p_t(x, y) \Big(\mathbb{E}\big[(x - \lambda_t \overrightarrow{X}_0)(x - \lambda_t \overrightarrow{X}_0)^\top \mid \overrightarrow{X}_t = x, Y = y\big] - \sigma_t^2 I_d \Big) \\
&= \sigma_t^{-4} p_t(x, y) \Big((x - \lambda_t m_t(x, y))(x - \lambda_t m_t(x, y))^\top + \lambda_t^2\, \mathrm{Cov}[\overrightarrow{X}_0 \mid \overrightarrow{X}_t = x, Y = y] - \sigma_t^2 I_d \Big).
\end{aligned}
$$

We then compute the Hessian of the log-density. Note that

$$
\nabla_x^2 \log p_t(x, y) = \frac{1}{p_t(x, y)} \nabla_x^2 p_t(x, y) - \big(\nabla_x \log p_t(x, y)\big)\big(\nabla_x \log p_t(x, y)\big)^\top,
$$

Using the expressions we derived for $\nabla_x \log p_t(x, y)$ and $\nabla_x^2 p_t(x, y)$, we have

$$
\nabla_x^2 \log p_t(x, y) = \lambda_t^2 \sigma_t^{-4}\, \mathrm{Cov}[\overrightarrow{X}_0 \mid \overrightarrow{X}_t = x, Y = y] - \sigma_t^{-2} I_d.
$$

An analogous calculation for $p_t(x)$ gives

$$
\nabla_x^2 \log p_t(x) = \lambda_t^2 \sigma_t^{-4}\, \mathrm{Cov}[\overrightarrow{X}_0 \mid \overrightarrow{X}_t = x] - \sigma_t^{-2} I_d,
$$

so that

$$
\nabla_x^2 \log p_t(y \mid x) = \nabla_x^2 \log p_t(x, y) - \nabla_x^2 \log p_t(x) = \lambda_t^2 \sigma_t^{-4} \Big(\Sigma_t(x, y) - \Sigma_t(x) \Big).
$$

By assumption $\mathrm{Tr}(\Sigma_t(x_t, y)), \mathrm{Tr}(\Sigma_t(x_t)) \in [0, R^2]$, thus we have

$$
\big| \mathrm{Tr}\left(\nabla_x^2 \log p_t(y \mid x)\right) \big| = \Big| \lambda_t^2 \sigma_t^{-4} \Big(\mathrm{Tr}(\Sigma_t(x, y)) - \mathrm{Tr}(\Sigma_t(x)) \Big) \Big| \le \lambda_t^2 \sigma_t^{-4} R^2.
$$

The proof is done.

B.2 Proof of Lemma B.2

Note that $p_t(x \mid y) = \int_{\mathcal{K}} P_{\mathrm{data}}(x_0 \mid y) \varphi_t(x \mid x_0) \mathrm{d}x_0$, where φ_t is the normal density associated with the conditional distribution $\overrightarrow{X}_t \mid \overrightarrow{X}_0 = x_0 \sim \mathcal{N}(\lambda_t x_0, \sigma_t^2 I_d)$. Therefore, $p_t(x \mid y) \le \sup_{z \in \mathcal{K}} \varphi_t(x \mid z)$. For $\rho \ge 4\lambda_t R$, $\|x\|_2 = \rho$ and $z \in \mathcal{K}$, by the triangle inequality we have $\|x - \lambda_t z\|_2 \ge \rho - \lambda_t R$. Thus we have

$$
\varphi_t(x \mid z) = \frac{1}{(2\pi\sigma_t^2)^{d/2}} \exp\left(-\frac{\|x_t - \lambda_t z\|_2^2}{2\sigma_t^2} \right) \le \frac{1}{(2\pi\sigma_t^2)^{d/2}} \exp\left(-\frac{(\rho - \lambda_t R)^2}{2\sigma_t^2} \right).
$$

Let $C_t = (2\pi\sigma_t^2)^{-d/2}$ and $c_t = 1/(2\sigma_t^2)$. We can then conclude that for $\|x\|_2 = \rho$, it holds that $p_t(x \mid y) \le C_t \exp(-c_t\, (\rho - \lambda_t R)^2)$. For $\rho \ge 4\lambda_t R$, we have $(\rho - \lambda_t R)^2 \ge \rho^2/2$, which implies that $\sup_{\|x\|_2 = \rho} p_t(x \mid y) \le C_t \exp\big(-c_t \rho^2/2\big)$.

By Proposition B.1 (a), we have $\|\nabla_x \log p_t(y \mid x)\|_2 \le 2\lambda_t \sigma_t^{-2} R$. By triangle inequality, this implies that $\|\nabla_x u_t(x)\|_2 \le 4\lambda_t \sigma_t^{-2} R$. By the fundamental theorem of calculus we have

$$
|u_t(x)| \le |u_t(0_d)| + \int_0^1 |\langle \nabla_x u_t(rx), x \rangle| \mathrm{d}r \le |u_t(0_d)| + 4\lambda_t \sigma_t^{-2} R \|x\|_2.
$$

Since $p_t(y \mid x), \hat{p}_{\phi,t}(y \mid x) \in C^2(\mathbb{R}^d)$, the function $x \mapsto u_t(x) p_t(x \mid y) \nabla_x u_t(x) \in C^2(\mathbb{R}^d)$. By the divergence theorem, it holds that

$$
I_\rho = \int_{\mathcal{B}_\rho} \nabla_x \cdot (u_t(x) p_t(x \mid y) \nabla_x u_t(x))\, \mathrm{d}x = \int_{\partial \mathcal{B}_\rho} u_t(x) p_t(x \mid y) \nabla_x u_t(x) \cdot n(x) \mathrm{d}S
$$

where $n(x)$ is the outward pointing unit normal at almost each point on the boundary $\partial\mathcal{B}_\rho$. On $\partial\mathcal{B}_\rho$, we have

$$|u_t(x)|\, p_t(x \mid y)\, \|\nabla_x u_t(x)\|_2 \le 4\lambda_t \sigma_t^{-2} R\Big(|u_t(0_d)| + 4\lambda_t\sigma_t^{-2}R\|x\|_2\Big) C_t \exp\left(-c_t\rho^2/2\right).$$

A classical result in measure theory states that the Lebesgue measure of a d-dimensional sphere of radius ρ satisfies $\mu(\partial\mathcal{B}_\rho) \asymp \rho^{d-1}$ [21, Section 2.7, Corollary 2.55]. Combining these results, we conclude that

$$\begin{aligned}
|I_\rho| &\le \int_{\partial\mathcal{B}_\rho} |u_t(x)|\, p_t(x \mid y) \|\nabla_x u_t(x)\|_2 \mathrm{d}S \\
&\lesssim 4\lambda_t\sigma_t^{-2} R\Big(|u_t(\vec{0}_d)| + 4\lambda_t\sigma_t^{-2}R\|x\|_2\Big) C_t \exp\left(-c_t\rho^2/2\right) \rho^{d-1} \to 0 \qquad \text{as } \rho\to\infty.
\end{aligned}$$

The proof is done.

B.3 Proof of Lemma B.3

Define $f_\rho(x) = p_t(x \mid y)\|\nabla_x u_t(x)\|_2^2 \mathbb{1}_{\mathcal{B}_\rho}(x)$, then as $\rho\to\infty$,

$$\lim_{\rho\to\infty} \mathbb{1}_{\mathcal{B}_\rho} = 1, \qquad \lim_{\rho\to\infty} p_t(x \mid y)\|\nabla_x u_t(x)\|_2^2 \mathbb{1}_{\mathcal{B}_\rho}(x) = p_t(x \mid y)\|\nabla_x u_t(x)\|_2^2.$$

By the monotone convergence theorem,

$$\lim_{\rho\to\infty} \int_{\mathcal{B}_\rho} p_t(x \mid y)\|\nabla_x u_t(x)\|_2^2 \mathrm{d}x = \int_{\mathbb{R}^d} p_t(x \mid y)\|\nabla_x u_t(x)\|_2^2 \mathrm{d}x.$$

Now consider $g_t(x) = u_t(x)p_t(x \mid y)[\langle \nabla_x \log p_t(x \mid y), \nabla_x u_t(x)\rangle + \mathrm{Tr}(\nabla_x^2 u_t(x))]$. Note that

$$|g_t(x)| \lesssim C_t\lambda_t^2\sigma_t^{-4}R^2 \exp\left(-c_t\|x\|_2^2/2\right)\Big(|u_t(0_d)| + 4\lambda_t\sigma_t^{-2}R\|x\|_2\Big) := h_t(x).$$

Notice that $h_t \in L^1(\mathbb{R}^d)$, hence by the dominated convergence theorem,

$$\lim_{\rho\to\infty} \int_{\mathcal{B}_\rho} g_t(x)\mathrm{d}x = \int_{\mathbb{R}^d} g_t(x)\mathrm{d}x.$$

The proof is done.

B.4 Proof of Lemma B.4

Fix $u \in \mathbb{S}^{d-1}$. By Jensen's inequality and Markov's inequality, we have

$$\begin{aligned}
\mathbb{P}\left(\mathbb{E}[\langle X, u\rangle \mid \mathcal{F}] \ge \lambda\right) &\le \inf_{c>0} \mathbb{P}\left(e^{\mathbb{E}[\langle X,u\rangle|\mathcal{F}]^2/c^2} \ge e^{\lambda^2/c^2}\right) \\
&\le \inf_{c>0} \mathbb{P}\left(\mathbb{E}\left[e^{\langle X,u\rangle^2/c^2} \mid \mathcal{F}\right] \ge e^{\lambda^2/c^2}\right) \\
&\le \inf_{c>0} \frac{\mathbb{E}\left[\mathbb{E}\left[e^{\langle X,u\rangle^2/c^2} \mid \mathcal{F}\right]\right]}{e^{\lambda^2/c^2}} \\
&= \inf_{c>0} \frac{\mathbb{E}[e^{\langle X,u\rangle^2/c^2}]}{e^{\lambda^2/c^2}} \\
&\le 2e^{-\lambda^2/\|X\|_{\psi_2}^2}
\end{aligned}$$

where the last inequality holds by taking $c = \|X\|_{\psi_2}$ [60, Definition 2.6.4]. Let $\mathcal{N} \subseteq \mathbb{S}^{d-1}$ be a $1/2$-net of $\mathbb{S}^{d-1}$. Standard volumetric arguments then imply that $|\mathcal{N}| \le 5^d$ [60, Corollary 4.2.11]. Then a union bound implies that

$$\mathbb{P}\left(\exists u \in \mathcal{N} : \mathbb{E}[\langle X, u\rangle \mid \mathcal{F}] \ge \lambda\right) \le 5^d \cdot 2e^{-\lambda^2/\|X\|_{\psi_2}^2}.$$

By [60, Exercise 4.35], we have $\|x\|_2 \le 2\sup_{u\in\mathcal{N}}\langle u, x\rangle$. Using this and setting $\lambda = \|X\|_{\psi_2}\sqrt{\log\left(\frac{2\cdot 5^d}{\epsilon}\right)}$, we get

$$\mathbb{P}\left(\|\mathbb{E}[X \mid \mathcal{F}]\|_2 \ge 2\|X\|_{\psi_2}\sqrt{\log\left(\frac{2\cdot 5^d}{\epsilon}\right)}\right) \le \epsilon.$$

The proof is done.

B.5 Proof of Corollary B.1

Inspecting the proof of Lemma B.1, we have

$$\begin{aligned}\nabla_x \log p_t(x \mid y) &= \sigma_t^{-2}\left(\lambda_t \mathbb{E}[X_0 \mid X_t = x, Y = y] - x\right) \\ &= -\sigma_t^{-2}\mathbb{E}[X_t - \lambda_t X_0 \mid X_t = x, Y = y],\end{aligned}$$

where $X_0 \sim P_{\text{data}}$ and $X_t = \lambda_t X_0 + \sigma_t \xi$ for $\xi \sim \mathcal{N}(0, I_d)$ that is independent of X_0. Noting that $X_t - \lambda_t X_0 = \sigma_t \xi$ is sub-Gaussian with $\|X_t - \lambda_t X_0\|_{\psi_2} \le C_1 \sigma_t$ for some universal constant $C_1 > 0$. The proof is done by Lemma B.4.

B.6 Proof of Lemma B.5

Inspecting the proof of Lemma B.1, we have

$$\begin{aligned}\nabla_x^2 \log p_t(x \mid y) &= \lambda_t^2 \sigma_t^{-4}\ \mathrm{Cov}[X_0 \mid X_t = x, Y = y] - \sigma_t^{-2} I_d \\ &= \sigma_t^{-4}\ \mathrm{Cov}[X_t - \lambda_t X_0 \mid X_t = x, Y = y] - \sigma_t^{-2} I_d,\end{aligned}$$

where $X_0 \sim P_{\text{data}}$ and $X_t = \lambda_t X_0 + \sigma_t \xi$ for $\xi \sim \mathcal{N}(0, I_d)$ that is independent of X_0. Note that $X_t - \lambda_t X_0 = \sigma_t \xi$. Therefore,

$$\begin{aligned}\left|\mathrm{Tr}\big(\nabla_x^2 \log p_t(x \mid y)\big)\right| &\le \sigma_t^{-2}\Big(d + \mathrm{Tr}\big(\mathrm{Cov}[\xi \mid X_t = x, Y = y]\big)\Big) \\ &\le \sigma_t^{-2}\Big(d + \mathrm{Tr}\big(\mathbb{E}[\xi\xi^\top \mid X_t = x, Y = y]\big)\Big).\end{aligned}$$

Note that $\mathbb{E}_{x \sim p_t(\cdot \mid y)}[\mathrm{Tr}(\mathbb{E}[\xi\xi^\top \mid X_t = x, Y = y])] = d$, and that for $\lambda \ge 0$,

$$\mathbb{E}_{x \sim p_t(\cdot \mid y)}\left[e^{\lambda(\mathrm{Tr}(\,\mathbb{E}[\xi\xi^\top \mid X_t = x, Y = y]) - d)}\right] \le \mathbb{E}_{x \sim p_t(\cdot \mid y)}\left[\mathbb{E}\left[e^{\lambda(\mathrm{Tr}(\xi\xi^\top) - d)} \mid X_t = x, Y = y\right]\right],$$

where the upper bound is by Jensen's inequality. Therefore, $\|\mathrm{Tr}(\,\mathbb{E}[\xi\xi^\top \mid X_t = x, Y = y]) - d\|_{\psi_1} \le c_1 \|\mathrm{Tr}(\xi\xi^\top) - d\|_{\psi_1} \le c_2\sqrt{d}$ (note that $(\mathrm{Tr}(\xi\xi^\top) - d)$ is distributed as a centered chi-squared distribution with d degrees of freedom). Here, $c_1, c_2 > 0$ are numerical constants, and $\|\cdot\|_{\psi_1}$ denotes the sub-exponential norm [60]. Therefore, by sub-exponential tail bound, there exists a numerical constant $C > 0$, such that

$$\mathbb{P}_{x \sim p_t(\cdot \mid y)}\left(\left|\mathrm{Tr}\big(\nabla_x^2 \log p_t(x \mid y)\big)\right| \le \frac{C(d + \log(1/\epsilon))}{\sigma_t^2}\right) \ge 1 - \epsilon.$$

The proof is done.

B.7 Proof of Corollary B.2

We first prove the first point. By Bayes rule and the proof of Lemma B.1,

$$\begin{aligned}\nabla_x \log p_t(y \mid x) &= \nabla_x \log p_t(x, y) - \nabla_x \log p_t(x) \\ &= -\sigma_t^{-2}\left(\mathbb{E}[X_t - \lambda_t X_0 \mid X_t = x, Y = y] - \mathbb{E}[X_t - \lambda_t X_0 \mid X_t = x]\right),\end{aligned}$$

where $X_0 \sim P_{\text{data}}$ and $X_t = \lambda_t X_0 + \sigma_t \xi$ for $\xi \sim \mathcal{N}(0, I_d)$ that is independent of X_0. By Corollary B.1, there exists a numerical constant $C_1 > 0$, such that

$$\mathbb{P}_{x \sim p_t(\cdot \mid y)}\left(\|\nabla_x \log p_t(x \mid y)\|_2 \le \frac{C_1\sqrt{d + \log(1/\epsilon)}}{\sigma_t}\right) \ge 1 - \epsilon.$$

Likewise, there exists a numerical constant $C_2 > 0$, such that

$$\begin{aligned}&\mathbb{P}_{x \sim p_t}\left(\|\nabla_x \log p_t(x)\|_2 \le \frac{C_2\sqrt{d + \log(1/\epsilon)}}{\sigma_t}\right) \ge 1 - \frac{\epsilon}{2} \\ \implies &\mathbb{P}_{x \sim p_t(\cdot \mid y)}\left(\|\nabla_x \log p_t(x)\|_2 \le \frac{C_2\sqrt{d + \log(1/\epsilon)}}{\sigma_t}\right) \ge 1 - \frac{\epsilon}{2P_{\text{data}}(y)}.\end{aligned}$$

Point (a) then follows by taking $C = C_1 + C_2$. The proof of point (b) relies on Lemma B.5 and is similar to that of point (a), and we skip it for compactness.

B.8 Proof of Lemma B.6

Note that

$$\begin{aligned} D_{\mathrm{KL}}\big(p_t \,\|\, \hat{p}_t\big) =& \mathbb{E}_{(x,y')\sim p_t}\left[\log \frac{p_t(y',x)}{\hat{p}_t(y',x)}\right] = \mathbb{E}_{(x,y')\sim p_t}\left[\log \frac{p_t(y' \mid x)}{\hat{p}_t(y' \mid x)}\right] \\ =& \mathbb{E}_{x\sim p_t}\left[D_{\mathrm{KL}}\big(p_t(\cdot \mid x) \,\|\, \hat{p}_t(\cdot \mid x)\big)\right]. \end{aligned}$$

By assumption we have $D_{\mathrm{KL}}\big(p_t \,\|\, \hat{p}_t\big) \;\le\; \varepsilon^2$. Using the chain rule for KL divergence, we get

$$D_{\mathrm{KL}}\big(p_t \,\|\, \hat{p}_t\big) = \mathbb{E}_{y\sim p_t}\big[D_{\mathrm{KL}}\big(p_t(\cdot \mid y) \,\|\, \hat{p}_t(\cdot \mid y)\big)\big] + D_{\mathrm{KL}}\big(p_t(y) \,\|\, \hat{p}_t(y)\big),$$

which further implies that

$$\sum_{y\in\mathcal{Y}} p_t(y)\, D_{\mathrm{KL}}\big(p_t(\cdot \mid y) \,\|\, \hat{p}_t(\cdot \mid y)\big) \;\le\; \varepsilon^2 \Longrightarrow D_{\mathrm{KL}}\big(p_t(\cdot \mid y) \,\|\, \hat{p}_t(\cdot \mid y)\big) \;\le\; \frac{\varepsilon^2}{p_t(y)} = \frac{\varepsilon^2}{P_{\mathrm{data}}(y)}.$$

The proof is done.

B.9 Proof of Lemma B.7

Define $A = \{x : p_t(y \mid x) < \hat{p}_t(y \mid x)\}$ and $B = \{x : p_t(y \mid x) \ge \hat{p}_t(y \mid x)\}$. Note that

$$\mathbb{E}_{x\sim p_t(\cdot|y)}[|u_t(x)|] = \underbrace{\int_A p_t(x \mid y)|u_t(x)|\mathrm{d}x}_{\mathrm{I}} + \underbrace{\int_B p_t(x \mid y)|u_t(x)|\mathrm{d}x}_{\mathrm{II}}.$$

We first bound term I. On set A we have $\log p_t(y \mid x) - \log \hat{p}_t(y \mid x) < 0$. Since $\log(1+x) \le x$, we have

$$\left|\log\left(\frac{p_t(y \mid x)}{\hat{p}_t(y \mid x)}\right)\right| = \log\left(\frac{\hat{p}_t(y \mid x)}{p_t(y \mid x)}\right) \le \frac{\hat{p}_t(y \mid x) - p_t(y \mid x)}{p_t(y \mid x)}.$$

Using this upper bound, we can bound term I as follows:

$$\mathrm{I} \le \int_A p_t(x \mid y)\, \frac{\hat{p}_t(y \mid x) - p_t(y \mid x)}{p_t(y \mid x)}\, \mathrm{d}x.$$

Applying Bayes' rule $p_t(x \mid y) = p_t(x)\, p_t(y \mid x)/p_t(y)$, the integrand simplifies to:

$$p_t(x \mid y)\, \frac{\hat{p}_t(y \mid x) - p_t(y \mid x)}{p_t(y \mid x)} = \frac{p_t(x)}{p_t(y)}\, \big(\hat{p}_t(y \mid x) - p_t(y \mid x)\big).$$

Therefore, using the equality $p_t(y) = P_{\mathrm{data}}(y)$, it holds that

$$\mathrm{I} \le \frac{1}{p_t(y)} \int_A p_t(x)\, \big(\hat{p}_t(y \mid x) - p_t(y \mid x)\big)\, \mathrm{d}x \le \frac{1}{P_{\mathrm{data}}(y)} \int p_t(x)\, \big|\hat{p}_t(y \mid x) - p_t(y \mid x)\big|\, \mathrm{d}x. \tag{21}$$

For each x, let $P_x = p_t(\cdot \mid x)$ and $Q_x = \hat{p}_t(\cdot \mid x)$. Then

$$\big|\hat{p}_t(y \mid x) - p_t(y \mid x)\big| \le \mathrm{TV}\big(P_x, Q_x\big) \le \sqrt{\frac{1}{2}\, D_{\mathrm{KL}}\big(P_x \,\|\, Q_x\big)},$$

where the last upper bound follows from Pinsker's inequality. Taking expectations and using Jensen's inequality, we have

$$\begin{aligned} \mathbb{E}_{X_t\sim p_t}\big[\, \big|\hat{p}_t(y \mid X_t) - p_t(y \mid X_t)\big|\, \big] &\le \mathbb{E}_{X_t\sim p_t}\left[\sqrt{\frac{1}{2}\, D_{\mathrm{KL}}\big(p_t(\cdot \mid X_t) \,\|\, \hat{p}_t(\cdot \mid X_t)\big)}\right] \\ &\le \sqrt{\frac{1}{2}\, \mathbb{E}_{X_t\sim p_t}\big[D_{\mathrm{KL}}\big(p_t(\cdot \mid X_t) \,\|\, \hat{p}_t(\cdot \mid X_t)\big)\big]}. \end{aligned} \tag{22}$$

By assumption, we have $\mathbb{E}_{X_t\sim p_t}[D_{\mathrm{KL}}(p_t(\cdot \mid X_t) \,\|\, \hat{p}_t(\cdot \mid X_t))] \;\le\; \varepsilon_t^2$. Combining Eqs. (21) and (22) gives

$$\mathrm{I} \;\le\; \frac{1}{P_{\mathrm{data}}(y)} \cdot \frac{\varepsilon_t}{\sqrt{2}}. \tag{23}$$

We next upper bound term II. Define $\mathcal{Z} = B \cup \{*\}$, and a mapping $\mathcal{A}^B : \mathbb{R}^d \to \mathcal{Z}$ by

$$\mathcal{A}^B(x) = \begin{cases} x & x \in B, \\ * & x \notin B. \end{cases}$$

Let $P_y = p_t(\cdot \mid y)$ and $\hat{P}_y = \hat{p}_t(\cdot \mid y)$. Define the pushforward measures $P_{y,B} = (\mathcal{A}^B)_\sharp P_y$ and $\hat{P}_{y,B} = (\mathcal{A}^B)_\sharp \hat{P}_y$. By the data processing inequality for KL divergences, we have

$$D_{\mathrm{KL}}(P_{y,B} \mid\mid \hat{P}_{y,B}) \le D_{\mathrm{KL}}(P_y \mid\mid \hat{P}_y) = D_{\mathrm{KL}}(p_t(\cdot \mid y) \mid\mid \hat{p}_t(\cdot \mid y)).$$

We then compute $D_{\mathrm{KL}}(P_{y,B} \,\|\, \hat{P}_{y,B})$ explicitly. On the set B, the measures $P_{y,B}$ and $\hat{P}_{y,B}$ admit Lebesgue densities $p_t(x \mid y)$ and $\hat{p}_t(x \mid y)$, while at the atom $\{*\}$ they have masses $P_y(B^c)$ and $\hat{P}_y(B^c)$, respectively. As a consequence,

$$\begin{aligned}
&D_{\mathrm{KL}}(P_{y,B} \mid\mid \hat{P}_{y,B}) \\
&= \int_B p_t(x \mid y) \log\left(\frac{p_t(x \mid y)}{\hat{p}_t(x \mid y)}\right) \mathrm{d}x + P_y(B^c) \log\left(\frac{P_y(B^c)}{\hat{P}_y(B^c)}\right) \\
&= \int_B p_t(x \mid y) \log\left(\frac{p_t(y \mid x)}{\hat{p}_t(y \mid x)}\right) \mathrm{d}x + P_y(B) \log\left(\frac{\hat{p}_t(y)}{p_t(y)}\right) + P_y(B^c) \log\left(\frac{P_y(B^c)}{\hat{P}_y(B^c)}\right),
\end{aligned}$$

where the second equality uses the decompositions $p_t(x \mid y) = p_t(x)p_t(y \mid x)/p_t(y)$ and $\hat{p}_t(x \mid y) = p_t(x)\hat{p}_t(y \mid x)/\hat{p}_t(y)$ (using the notations from Lemma B.6). Leveraging the above equality, we have

$$\begin{aligned}
&\int_B p_t(x \mid y) \log\left(\frac{p_t(y \mid x)}{\hat{p}_t(y \mid x)}\right) \mathrm{d}x \\
&= D_{\mathrm{KL}}(P_{y,B} \mid\mid \hat{P}_{y,B}) - P_y(B^c) \log\left(\frac{P_y(B^c)}{\hat{P}_y(B^c)}\right) - P_y(B) \log\left(\frac{\hat{p}_t(y)}{p_t(y)}\right) \\
&\le D_{\mathrm{KL}}(p_t(\cdot \mid y) \mid\mid \hat{p}_t(\cdot \mid y)) - P_y(B^c) \log\left(\frac{P_y(B^c)}{\hat{P}_y(B^c)}\right) - P_y(B) \log\left(\frac{\hat{p}_t(y)}{p_t(y)}\right).
\end{aligned}$$

By Lemma B.6, it holds that $D_{\mathrm{KL}}(p_t(\cdot \mid y) \mid\mid \hat{p}_t(\cdot \mid y)) \le \frac{\varepsilon_t^2}{P_{\mathrm{data}}(y)}$. In addition, $\left|\hat{P}_y(B^c) - P_y(B^c)\right| \le \mathrm{TV}(P_y, \hat{P}_y) \le \sqrt{D_{\mathrm{KL}}(P_y \,\|\, \hat{P}_y)/2} \le \varepsilon_t/\sqrt{2P_{\mathrm{data}}(y)}$, and by Eq. (22) we have $\left|p_t(y) - \hat{p}_t(y)\right| \le \mathbb{E}_{X_t \sim p_t}\left[\left|\hat{p}_t(y \mid X_t) - p_t(y \mid X_t)\right|\right] \le \varepsilon_t/\sqrt{2}$. Using these upper bounds and the inequality $\log(1+x) < x$, we have

$$\begin{aligned}
\mathrm{II} =& \int_B p_t(x \mid y) \log\left(\frac{p_t(y \mid x)}{\hat{p}_t(y \mid x)}\right) \mathrm{d}x \\
\le& \frac{\varepsilon_t^2}{P_{\mathrm{data}}(y)} + P_y(B^c) \cdot \frac{|P_y(B^c) - \hat{P}_y(B^c)|}{P_y(B^c)} + \frac{|p_t(y) - \hat{p}_t(y)|}{p_t(y) - |p_t(y) - \hat{p}_t(y)|} \\
\le& \frac{\varepsilon_t^2}{P_{\mathrm{data}}(y)} + \frac{\varepsilon_t}{\sqrt{2}P_{\mathrm{data}}(y)^{1/2}} + \frac{\varepsilon_t}{\sqrt{2}(P_{\mathrm{data}}(y) - \varepsilon_t/\sqrt{2})} \\
\le& \frac{\varepsilon_t^2}{P_{\mathrm{data}}(y)} + \frac{\varepsilon_t}{\sqrt{2}P_{\mathrm{data}}(y)^{1/2}} + \frac{\sqrt{2}\varepsilon_t}{P_{\mathrm{data}}(y)}.
\end{aligned} \tag{24}$$

The proof is done by putting together Eqs. (23) and (24).

B.10 Proof of Lemma B.8

By Bayes' rule, $\nabla_x \log p_s(y \mid x) = \nabla_x \log p_s(x \mid y) - \nabla_x \log p_s(x)$. Therefore,

$$\begin{aligned}
&\sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_{x\sim p_{T-t}(\cdot|y)} \left[\left\| \nabla_x \log \frac{p_{T-t}(x)}{p_{T-t_k}(x)} + \gamma \nabla_x \log \frac{p_{T-t}(y \mid x)}{p_{T-t_k}(y \mid x)} \right\|_2^2 \right] \mathrm{d}t \\
&= \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} \mathbb{E}_{x\sim p_{T-t}(\cdot|y)} \left[\left\| (1-\gamma) \nabla_x \log \frac{p_{T-t}(x)}{p_{T-t_k}(x)} + \gamma \nabla_x \log \frac{p_{T-t}(x \mid y)}{p_{T-t_k}(x \mid y)} \right\|_2^2 \right] \mathrm{d}t \\
&\lesssim (2\gamma^2 - 2\gamma + 1)(d + R^2)(\kappa^2 N + \kappa T),
\end{aligned}$$

where the second inequality follows by applying $\|a+b\|_2^2 \lesssim \|a\|_2^2 + \|b\|_2^2$ and the same argument as [2, Lemma 2] to the unconditional and conditional score increments separately.

B.11 Proof of Lemma B.9

Let $h_T(x) := p_T(y \mid x)$ and $p_y := P_{\mathrm{data}}(y)$. Recall that $p_{T,\gamma}(x)/p_T(x) = h_T(x)^\gamma / Z_{T,\gamma}$ and, by Jensen's inequality, $Z_{T,\gamma} \geq p_y^\gamma$. Using the inequality $D_{\mathrm{KL}}(P\|Q) \leq \chi^2(P\|Q)$, we obtain

$$D_{\mathrm{KL}}\big(p_{T,\gamma} \,\|\, p_T\big) \leq \frac{\mathrm{Var}_{X\sim p_T}\big(h_T(X)^\gamma\big)}{Z_{T,\gamma}^2} \leq \frac{\gamma^2}{p_y^{2\gamma}} \mathbb{E}_{X\sim p_T}\big[(h_T(X) - p_y)^2\big],$$

where the second inequality follows since $u \mapsto u^\gamma$ is γ-Lipschitz on $[0,1]$.

Let $Y_y := \mathbb{1}\{Y = y\}$. Since $h_T(X_T) = \mathbb{P}(Y_y = 1 \mid X_T)$, Pinsker's inequality and the data-processing inequality give $\mathbb{E}[(h_T(X_T) - p_y)^2] \leq \frac{1}{2} I(Y_y; X_T) \leq \frac{1}{2} I(X_0; X_T)$. Writing the forward OU process as $X_T = e^{-T} X_0 + \sqrt{1 - e^{-2T}}\, Z$, where $Z \sim N(0, I_d)$, and using $\|X_0\|_2 \leq R$ almost surely, the standard Gaussian-channel bound yields

$$I(X_0; X_T) \leq \frac{d}{2} \log\left(1 + \frac{e^{-2T} R^2}{d(1 - e^{-2T})}\right) \leq \frac{e^{-2T} R^2}{2(1 - e^{-2T})}.$$

Consequently,

$$D_{\mathrm{KL}}\big(p_{T,\gamma} \,\|\, p_T\big) \leq \frac{\gamma^2 R^2 e^{-2T}}{4 p_y^{2\gamma} (1 - e^{-2T})}.$$

It remains to compare p_T with π_d. Applying the Donsker–Varadhan variational inequality with $P = p_{T,\gamma}$, $Q = p_T$, and $f(x) = \log\big(p_T(x)/\pi_d(x)\big)$,

$$\begin{aligned}
D_{\mathrm{KL}}\big(p_{T,\gamma} \,\|\, \pi_d\big) &= D_{\mathrm{KL}}\big(p_{T,\gamma} \,\|\, p_T\big) + \mathbb{E}_{p_{T,\gamma}}\left[\log \frac{p_T(X)}{\pi_d(X)}\right] \\
&\leq 2 D_{\mathrm{KL}}\big(p_{T,\gamma} \,\|\, p_T\big) + \log \int_{\mathbb{R}^d} \frac{p_T(x)^2}{\pi_d(x)}\, \mathrm{d}x.
\end{aligned}$$

Using again the OU representation, let $a := e^{-T}$ and $\sigma_T^2 := 1 - a^2$. Then $p_T(x) = \mathbb{E}_{X_0}\big[\phi_{\sigma_T^2 I_d}(x - aX_0)\big]$, where ϕ_Σ denotes the density of $N(0, \Sigma)$. Thus, for independent $X_0, X_0' \sim P_{\mathrm{data}}$, we have $(X_0, X_0') \sim P_{\mathrm{data}} \otimes P_{\mathrm{data}}$. This allows us to express $p_T(x)^2$ as an expectation over two independent copies, yielding:

$$\int_{\mathbb{R}^d} \frac{p_T(x)^2}{\pi_d(x)}\, \mathrm{d}x = \mathbb{E}_{X_0, X_0'}\left[\int_{\mathbb{R}^d} \frac{\phi_{\sigma_T^2 I_d}(x - aX_0)\phi_{\sigma_T^2 I_d}(x - aX_0')}{\phi_{I_d}(x)}\, \mathrm{d}x\right].$$

Completing the square, for fixed $u, v \in \mathbb{R}^d$, the inner integral is

$$\int_{\mathbb{R}^d} \frac{\phi_{\sigma_T^2 I_d}(x - au)\phi_{\sigma_T^2 I_d}(x - av)}{\phi_{I_d}(x)}\, \mathrm{d}x = (1 - a^4)^{-d/2} \exp\left\{\frac{a^2}{2(1 - a^4)}\left(2u^\top v - a^2\|u\|_2^2 - a^2\|v\|_2^2\right)\right\}.$$

Since $\|u\|_2, \|v\|_2 \leq R$ and $2u^\top v \leq \|u\|_2^2 + \|v\|_2^2$, we have $2u^\top v - a^2\|u\|_2^2 - a^2\|v\|_2^2 \leq 2(1 - a^2)R^2$. Therefore,

$$\int_{\mathbb{R}^d} \frac{p_T(x)^2}{\pi_d(x)}\, \mathrm{d}x \leq (1 - a^4)^{-d/2} \exp\left\{\frac{a^2 R^2}{1 + a^2}\right\}.$$

Taking logarithms and substituting $a = e^{-T}$ yields

$$\log \int_{\mathbb{R}^d} \frac{p_T(x)^2}{\pi_d(x)} \, \mathrm{d}x \leq -\frac{d}{2} \log(1 - e^{-4T}) + \frac{R^2 e^{-2T}}{1 + e^{-2T}}.$$

Since $T \geq 1$, we have $(1 - e^{-2T})^{-1} \lesssim 1$ and $-\log(1 - e^{-4T}) \lesssim e^{-4T} \leq e^{-2T}$. Combining the preceding bounds yields

$$D_{\mathrm{KL}}\big(p_{T,\gamma} \,\|\, \pi_d\big) \lesssim \left(d + \frac{\gamma^2 R^2}{P_{\mathrm{data}}^{2\gamma}}\right) e^{-2T},$$

which proves the result.

C Additional numerical experiments

C.1 Counterexample: controlling conditional KL does not guarantee effective guidance

We numerically validate Theorem 3.1 by constructing a counterexample in a binary classification setting. We use our construction to demonstrate that a classifier can achieve vanishingly small conditional KL, while exhibiting substantial guidance term misalignment, thereby demonstrating that controlling conditional KL does not necessarily guarantee effective guidance.

Experimental setup. We consider a binary classification problem with label space $\mathcal{Y} = \{0, 1\}$ and a 50-dimensional input space. We denote by p the target distribution, and assume the guidance is towards the conditional distribution $p(\cdot \mid y = 1)$. We set $p(x) = \mathcal{N}(0, I_{50})$. Conditioned on x, we take $p(y = 1 \mid x) = 0.5 + 0.3 \tanh(x_1)$, which yields probabilities in the range $[0.2, 0.8]$. We set $\gamma = 0.3$ and define the set $A_\gamma = \{x : p(y = 1 \mid x) < 1 - \gamma\}$. Following the construction in the proof of Theorem 3.1 (Appendix A.1), we construct the following probabilistic classifier by perturbing the true conditional label probabilities, with the perturbation indexed by n:

$$\hat{p}_n(y = 1 \mid x) = \begin{cases} p(y = 1 \mid x) \cdot [1 + \delta_n \sin(n x_1)], & \text{if } x \in A_\gamma, \\ p(y = 1 \mid x), & \text{if } x \notin A_\gamma, \end{cases}$$

with $\hat{p}_n(y = 0 \mid x) = 1 - \hat{p}_n(y = 1 \mid x)$ to ensure normalization. The perturbation acts only on the first coordinate x_1, so the remaining 49 coordinates do not contribute to the guidance error. The parameter $n \in \{10, 30, \ldots, 1000\}$ determines the oscillation frequency of the perturbation, and the amplitude δ_n also scales with n. For the Monte Carlo estimation of the conditional KL divergence and the guidance term MSE, we utilize $N = 10{,}000$ independent sample points. We examine the two regimes discussed in the proof of Theorem 3.1(Appendix A.1): (1) In Regime 1 we set $\delta_n = 1/n$, and the theorem predicts that the conditional KL decays at rate $O(1/n)$, while the guidance MSE is $\Omega(1)$; (2) In Regime 2 we set $\delta_n = 1/\sqrt{n}$. In this case, the theorem shows that the conditional KL decays at a slower rate $O(1/\sqrt{n})$, and the guidance MSE diverges to infinity at rate $\Omega(n)$.

Simulation outcomes. Figure 3 presents the simulation outcomes for both regimes. We measure two quantities: the conditional KL divergence $\mathbb{E}_{x \sim p}\left[\mathrm{KL}(p(\cdot \mid x) \,\|\, \hat{p}_n(\cdot \mid x))\right]$, and the guidance term MSE $\mathbb{E}_{x \sim p(\cdot \mid y=1)}\left[\|\nabla_x \log p(y = 1 \mid x) - \nabla_x \log \hat{p}_n(y = 1 \mid x)\|_2^2\right]$. We discuss the results from these two regimes separately in the following.

Regime 1 ($\delta_n = 1/n$): As shown in the first plot of Figure 3, the conditional KL (top-left) decreases from approximately 2×10^{-3} at $n = 10$ to approximately 2×10^{-7} at $n = 1000$, decaying faster than the $O(1/n)$ upper bound. On the other hand, the guidance MSE (second plot of Figure 3) remains nearly constant at around 0.343 across all frequencies, confirming the predicted $\Omega(1)$ behavior.

Regime 2 ($\delta_n = 1/\sqrt{n}$): As shown in the third plot of Figure 3, the conditional KL divergence (bottom-left) decreases from approximately 2×10^{-2} at $n = 10$ to approximately 2×10^{-3} at $n = 1000$, also decaying to zero faster than the upper bound $O(1/\sqrt{n})$. In contrast, the guidance error (fourth plot of Figure 3) increases dramatically from approximately 4 at $n = 10$ to approximately 300 at $n = 1000$, exhibiting clear $\Omega(n)$ growth.

C.2 Effect of classifier smoothness using counterexample construction

To further investigate the role of classifier smoothness in guidance accuracy, we extend the preceding counterexample experiment by varying the perturbation amplitude δ_n. Recall that the perturbation in

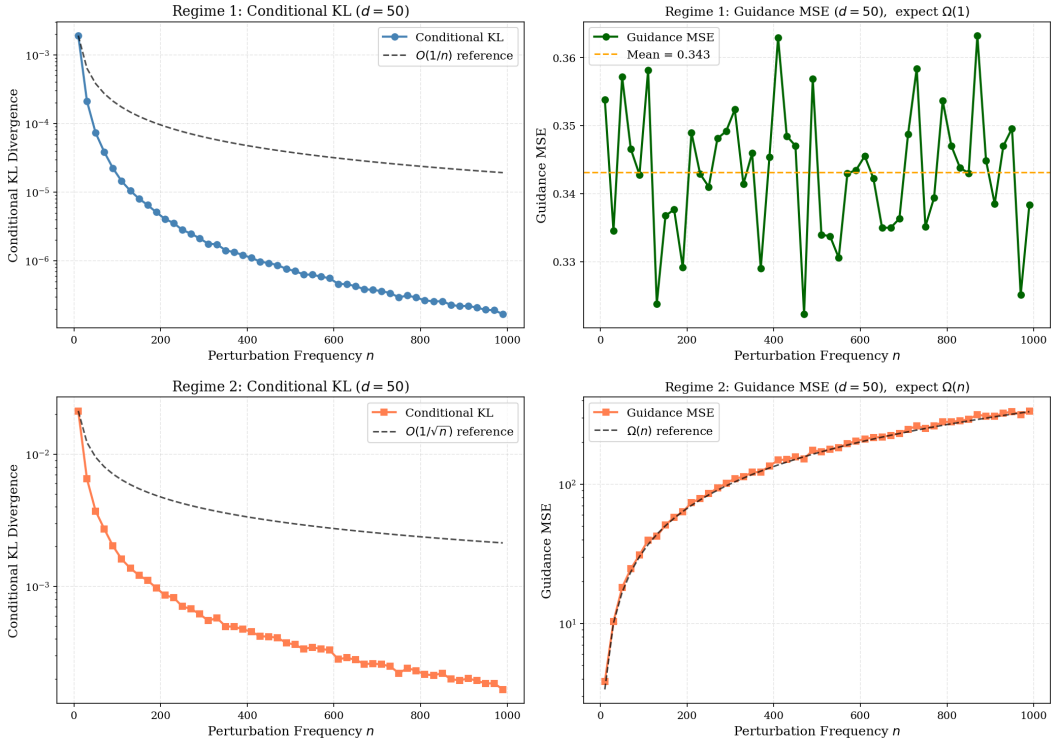


Figure 3: The first and second rows correspond to Regimes 1 and 2, respectively. The first column displays the conditional KL divergence, and the second column shows the guidance MSE.

our construction is of the form $\delta_n \sin(nx_1)$, whose first derivative has magnitude proportional to $n\delta_n$. We consider four choices

$$\delta_n \in \left\{ \frac{1}{n^2}, \frac{1}{n}, \frac{1}{\sqrt{n}}, \frac{1}{n^{1/4}} \right\}.$$

These choices correspond to progressively less regular perturbations as n increases: $n\delta_n = 1/n$ vanishes for $\delta_n = 1/n^2$, remains bounded for $\delta_n = 1/n$, and grows for $\delta_n = 1/\sqrt{n}$ and $\delta_n = 1/n^{1/4}$.

To isolate the effect of this regularity from classifier accuracy, we compare the four perturbation regimes at approximately matched levels of conditional KL divergence. Since the conditional KL values attained at the discrete frequencies do not coincide exactly across the four regimes, we evaluate the curves on a shared conditional-KL grid using interpolation. Figure 4 plots the resulting guidance MSE against conditional KL.

Figure 4 demonstrates that classifier accuracy as measured by conditional KL alone does not determine the quality of the guidance vector. At comparable conditional KL levels, the guidance MSE differs substantially across the four perturbation regimes. In particular, the $\delta_n = 1/n^2$ regime, for which the perturbation derivative vanishes as n increases, yields the smallest guidance error, whereas the less regular choices $\delta_n = 1/\sqrt{n}$ and $\delta_n = 1/n^{1/4}$ produce substantially larger guidance errors. The intermediate choice $\delta_n = 1/n$ yields a guidance error that remains approximately constant. These results empirically illustrate that increasing classifier regularity suppresses the high-frequency gradient errors underlying the counterexample in Theorem 3.1, and provide complementary motivation for the smoothness conditions imposed in Theorem 3.3.

C.3 Additional results for MNIST experiment

This section presents additional numerical results for the neural classifier experiment described in Section 4.2. Figure 5 shows the training loss and accuracy curves, indicating that both classifiers achieve similar performance.

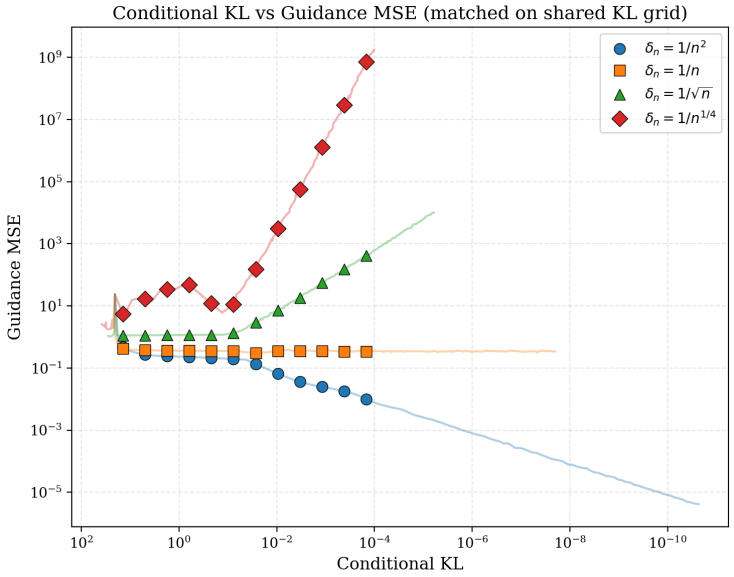


Figure 4: Effect of perturbation regularity on guidance error at comparable levels of conditional KL divergence. We consider $\delta_n \in \{1/n^2, 1/n, 1/\sqrt{n}, 1/n^{1/4}\}$ and evaluate the corresponding guidance MSEs on a shared conditional-KL grid.

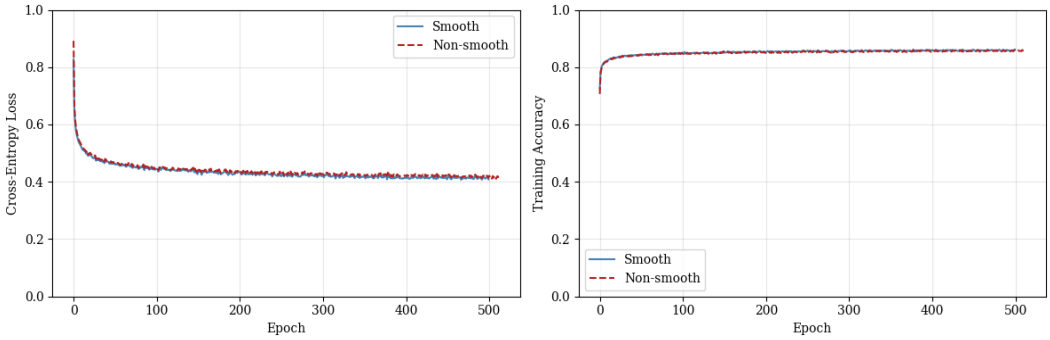


Figure 5: Training loss and accuracy curves for smooth and non-smooth classifiers. Both achieve similar final performance in terms of cross-entropy and accuracy.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state all main claims with their assumptions and scope, including the negative result (Theorem 3.1), the positive result under smoothness (Theorem 3.3), the rate optimality result (Theorem 3.6), and the sampling error bound (Theorem 3.7).

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The conclusion discusses limitations and future directions, including the restriction to compactly supported distributions and alternative classifier training procedures.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: All assumptions are stated explicitly in Theorem 3.3. Complete proofs for all theorems and lemmas appear in the appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: All experimental setups are described in detail in Section 4, including data distributions, sample sizes, classifier architectures, training procedures, and evaluation metrics.

Guidelines:

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

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

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

   Answer: [No]

   Justification: Experiments use synthetic data (GMM) and the MNIST dataset. Full experimental details are provided in the paper and appendix, but no anonymized code release is currently provided.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section 4 specify data distributions, sample sizes, classifier architectures (GELU/AvgPool vs. Hardtanh/MaxPool), training epochs, batch sizes, Monte Carlo sample sizes, and evaluation procedures.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The GMM experiments in Section 4.2 report results averaged over 20 independent trials with mean and standard error reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

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

8. **Experiments compute resources**

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

   Answer: [No]

   Justification: The experiments are small-scale: the counterexample and GMM validations rely on Monte Carlo estimation over synthetic data in $\mathbb{R}^{50}$, and the MNIST classifier training uses a standard small convolutional network trained for 400 epochs on 60,000 samples.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: This is a theoretical paper on diffusion model guidance with no foreseeable ethical concerns.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [N/A]

Justification: This is a theoretical paper on diffusion model guidance with no foreseeable societal impact.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: The paper poses no such risks.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: MNIST is cited.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

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

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs only used for writing and editing

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.