Diffusion Models Observe Only Gradients: A Geometric Perspective on Score Matching Errors

Naïl B. Khelifa
University of Cambridge
nbk24@cam.ac.uk

Richard E. Turner
University of Cambridge
ret26@cam.ac.uk

Ramji Venkataramanan
University of Cambridge
rv285@cam.ac.uk

Abstract

Score-based diffusion models are typically trained by minimizing the L^2 score matching error, and standard theoretical analyses rely on this quantity to bound the sampling discrepancy between the learned and target distributions. Recent work has shown that the score admits an orthogonal decomposition into gradient and solenoidal components in which only the gradient part matters for exact sampling. We apply this decomposition to the score error of trained models and study its quantitative consequences. Since only the gradient component of the error enters the marginal Fokker–Planck dynamics, a model can incur arbitrarily large L^2 score error while perfectly matching the target. Starting from this observation, we present three results. First, we derive an upper bound on the Kullback-Leibler divergence that depends only on the observable gradient component of the error. It tightens the standard Girsanov bound for generic score networks and identifies that bound's looseness as the cost of working with path-space rather than marginal-space dynamics. Second, we prove that no non-trivial function of the L^2 score error can uniformly lower bound any divergence between the learned and target distributions. Third, we give an estimator of the gradient component, computable from denoising score matching residuals via a dual Sobolev identity, which empirically correlates substantially better with sample quality than the full L^2 error.

1 Introduction

Denoising Score Matching (DSM) and its variants [38, 39, 78, 72] have driven the success of score-based diffusion models [69, 35, 71, 73] by providing a scalable score-estimation objective. To approximate an unknown target distribution $p^\star$, a score-based diffusion model transports it to an easy-to-sample prior distribution (typically, a standard Gaussian), from which samples are brought back, via a reverse diffusion, towards $p^\star$. To implement this reverse diffusion, one must estimate the unknown scores $\nabla \log p_s^\star$ of the time-s noised marginals of the target distribution along the reverse path (precise definitions in Section 2). DSM produces such estimators by minimizing the time-integrated L^2 error from the true unknown score. The accuracy of this estimation critically determines the sampling quality of the diffusion model [16], as illustrated by the following fundamental upper bound (up to constants) on the KL divergence between the target $p^\star$ and the learned distribution $\hat{p}$:

$$\mathrm{KL}(p^\star \,\|\, \hat{p}) \;\lesssim\; \frac{1}{2}\int_0^T \mathbb{E}\big[\|\mathbf{e}_s\|_2^2\big]\; \mathrm{d}s, \qquad \text{(Full } L^2\text{)}$$

where $\mathbf{e}_s = \hat{\mathbf{s}}_s - \nabla \log p_s^\star$ is the time-s score estimation error, i.e. the difference between the estimated score $\hat{\mathbf{s}}_s$ and the true score $\nabla \log p_s^\star$. Inequality (Full L^2) underpins much of the current theory of diffusion models [70], from convergence analyses [16, 52, 14, 7], to minimax optimality results [60, 83, 21, 53], and interpretations of training objectives in terms of sample quality [44, 45].

Inequality (Full L^2) ties the full L^2 score error to the sampling quality of the diffusion model. However, as observed in the exact setting by Horvat and Pfister [37] and Vuong et al. [79], a

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

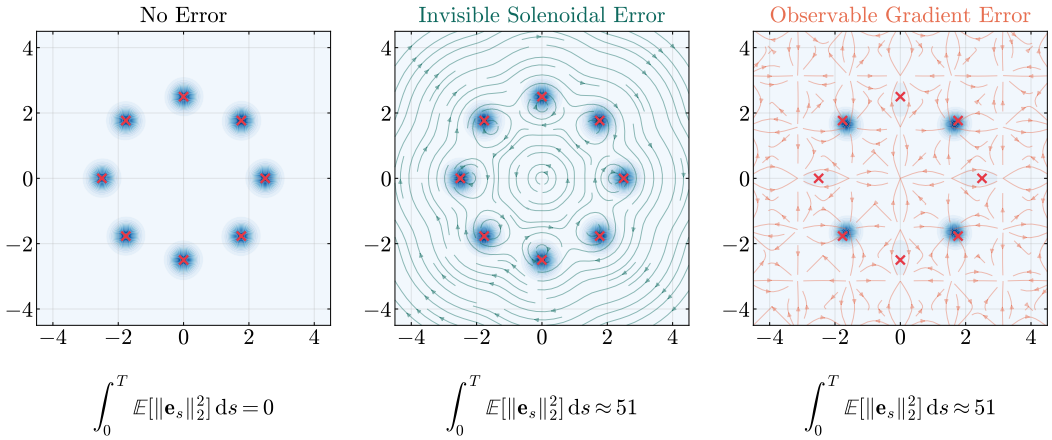

Figure 1: **Helmholtz-Hodge decomposition of score errors.** Two score estimates, both with total L^2 score matching error approximately equal to 51, manifest in qualitatively different generated distributions depending on their geometric structure. Red crosses represent true mode locations and blue centroids represent learned modes. *Center*: with purely solenoidal error fields (green arrows), the generated distribution is identical to the no-error case (left). *Right*: with gradient-like error fields (red arrows) of the same energy, the generated distribution visibly misses some modes. Standard score matching minimizes the full L^2 error, which cannot distinguish between these two cases, even though only the gradient component affects sampling quality.

learned diffusion model can incur arbitrarily large L^2 score error while perfectly matching the target distribution:

$$\mathrm{KL}(p^\star \,\|\, \hat{p}) = 0 \qquad \text{while} \qquad \int_0^T \mathbb{E}\big[\|\mathbf{e}_s\|_2^2\big]\ \mathrm{d}s \gg 0. \tag{1}$$

As a first contribution, we strengthen this into an impossibility result: no monotone function of the L^2 score error can uniformly lower-bound any meaningful divergence between $\hat{p}$ and $p^\star$ (Theorem 3.1 (iii)). The L^2 error is therefore not the intrinsic quantity controlling marginal sampling quality, raising a natural question:

Which components of the score error control sampling quality?

The key feature underlying observation (1) is that the marginal distributions of the diffusion process used to generate the sample solve a Fokker-Planck equation [40] that depends on the score error $\mathbf{e}_s$ only through a weighted divergence (Section 3) . Decomposing the error vector field into gradient and solenoidal components (a Helmholtz–Hodge decomposition [8], standard in scientific machine learning [64, 32, 27, 75, 54]), the weighted divergence term filters out the solenoidal component of the error field, making it structurally invisible to the marginal dynamics, while the gradient component affects marginals (see Figure 1). Therefore, a model with purely solenoidal errors satisfies (1), as illustrated in the middle panel of Figure 1. This decomposition and the resulting invisibility were established in the context of diffusion models in [37, 79] to study exact sampling and density estimation. We instead apply it, for trained models with nonzero error, to the score *error* and use it to obtain quantitative guarantees on sampling quality.

The discussion above suggests that, although DSM reduces the gradient component of the score error by minimizing the full L^2 objective, this overall score matching error also penalizes solenoidal components that do not affect marginals. Therefore, it is not always a reliable indicator of sampling quality (as in the first two panels of Figure 1). Instead, considering only the L^2 norm of the observable gradient component of errors (denoted $\Pi_{\mathcal{G}_s}\mathbf{e}_s$), we derive a new bound on sampling quality (Theorem 3.2):

$$\mathrm{KL}(p^\star \,\|\, \hat{p}) \ \lesssim\ \frac{1}{2}\int_{t_0}^T \mathbb{E}\big[\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2\big]\mathrm{d}s. \tag{Ours}$$

This new bound is strictly tighter than the standard one in (Full L^2) whenever the solenoidal component of errors is nonzero, motivating a better diagnostic of learning quality based on the geometry of diffusion. We propose a tractable surrogate to estimate the gradient component of the score error.

Contributions. Our results concern the approximate regime where the score estimation error is nonzero, i.e. the trained score network does not learn the true score perfectly.

- **An impossibility result.** Complementing the exact gauge-freedom results of [37, 79], we show that no strictly increasing function of the full L^2 score error can uniformly lower-bound any divergence between the learned and target distributions (Theorem 3.1 (iii)).
- **A tighter endpoint KL bound.** Applying the Helmholtz–Hodge decomposition to the score error, we prove that the Kullback-Leibler divergence between the learned and target models is controlled only by the $L^2(p_s^\star)$ norm of the gradient component of the error (Theorem 3.2), tightening the standard Girsanov bound [7, 14, 16]. This component is the minimum-energy error field producing the same marginal evolution.
- **A bridge with path-space analysis.** We recover the same bound by applying Girsanov's theorem to a marginally equivalent representative of the true reverse process (Appendix C). This identifies the looseness of the standard Girsanov bound in controlling a marginal-level divergence by a path-space one, which accumulates contributions from all errors, even those that do not affect the marginals.
- **A DSM-computable diagnostic.** Using a dual Sobolev identity, we give a tractable estimator for the norm of the gradient component of the score error, i.e. the right-hand side of our bound (Section 4.2), that requires only forward-noised data and the frozen score network to be computed. On image datasets, the observable error decreases by two to three orders of magnitude during training, while the residual L^2 error is dominated by the invisible component. The observable error also correlates substantially better with feature-FID than the full L^2 error on Fashion-MNIST.

Related work. Recent studies have found that standard denoising score matching can be inefficient in various settings. When the data concentrate near a low-dimensional manifold, score matching can waste capacity by forcing the score network to fit large directions normal to the data manifold [63, 19]. In another vein, a popular line of work modifies the relative importance of score errors across noise levels or frequency regimes through timestep weighting and training reweighting schemes [31, 18, 80, 51]. These studies reflect the empirical fact that not all score errors are equally relevant for sample quality. However, all these approaches treat the full score estimation error as the central quantity to analyze, without examining its geometric structure.

Another line of work has debated whether the score network should be constrained to be conservative (a gradient field) by construction. Such a conservative parametrization would indeed automatically lead to errors with no solenoidal components. While early evidence suggested unconstrained parametrizations lose little in practice [66], Chao et al. [13] reported that non-conservativity can degrade sampling and proposed a Jacobian-symmetrizing penalty to address this. On the theoretical side, [68] shows that an unconstrained network generically cannot represent an exact gradient, so a nonzero solenoidal component is the typical case. Horvat and Pfister [37] and Vuong et al. [79] showed that conservativity of the score network is not required to produce exact marginals, as the solenoidal component is invisible to the continuous-time marginal dynamics. Theorem 3.1 extends this in the approximate regime, and Theorem 3.2 quantifies the effect of the remaining, observable error on the sampling quality. In concurrent work, Dupuis et al. [22, Theorem 4.3] obtained an upper bound closely related to the one in Theorem 3.2, under a different set of assumptions. We provide an extensive literature review in Appendix A.

Notation. Bold symbols (e.g. $\mathbf{f}, \mathbf{v}, \mathbf{w}, \ldots$) denote $\mathbb{R}^d$-valued objects, and $\mathbb{R}^d$-valued random variables are denoted $\mathbf{X}, \mathbf{Y}, \ldots$. $\mathcal{P}(\mathbb{R}^d)$ (resp. $\mathcal{P}_2(\mathbb{R}^d)$) denotes the space of Borel probability measures on $\mathbb{R}^d$ (resp. with finite second moment). For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the Kullback-Leibler divergence between these two measures is defined as $\mathrm{KL}(\mu \,\|\, \nu) = \int_{\mathbb{R}^d} \log(\frac{\mu}{\nu})\mu \mathrm{d}\mathbf{x}$ if μ is absolutely continuous with respect to ν and $+\infty$ otherwise. For a distribution $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, the space of square integrable vector fields relative to μ is defined as $L^2(\mu, \mathbb{R}^d) := \{\mathbf{v} : \mathbb{R}^d \to \mathbb{R}^d \ : \ \mathbb{E}_\mu[\|\mathbf{v}\|_2^2] < \infty\}$. It is a Hilbert space when equipped with its L^2 scalar product defined as follows. For $\mathbf{v}, \mathbf{w} \in L^2(\mu, \mathbb{R}^d)$,

we have $\langle \mathbf{v}\,, \mathbf{w} \rangle_{L^2(\mu,\mathbb{R}^d)} = \mathbb{E}_\mu[\mathbf{v} \cdot \mathbf{w}]$, where $\cdot$ denotes the standard Euclidean scalar product. This geometry yields the L^2-norm, defined as $\|\mathbf{v}\|_{L^2(\mu)} = (\mathbb{E}_\mu[\|\mathbf{v}\|_2^2])^{1/2}$.

The space of continuous functions from $[0,T]$ to $\mathbb{R}^d$ is denoted by $C([0,T],\mathbb{R}^d)$. For a stochastic process $(\mathbf{Z}_t)_{t\in[0,T]} \in C([0,T],\mathbb{R}^d)$, $\mathrm{Law}((\mathbf{Z}_t)_t)$ denotes its path-space distribution and $\mathrm{Law}(\mathbf{Z}_t)$ denotes the law of its marginal at time t. The space of differentiable scalar-valued functions on $\mathbb{R}^d$ with continuous gradient is denoted by $C^1(\mathbb{R}^d)$, and $C_c^\infty(\mathbb{R}^d)$ denotes the space of smooth (i.e. infinitely differentiable) scalar-valued functions with compact support on $\mathbb{R}^d$.

2 Background

Forward diffusion. A forward diffusion on $[0,T]$ ($T>0$) is defined by the stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{X}_t = \boldsymbol{f}_t(\mathbf{X}_t)\,\mathrm{d}t + \sigma_t\,\mathrm{d}\mathbf{B}_t, \qquad \mathbf{X}_0 \sim p^\star, \qquad t\colon 0 \to T, \tag{2}$$

where $\mathbf{B}_t$ is a standard d-dimensional Brownian motion, $\boldsymbol{f}_t : \mathbb{R}^d \to \mathbb{R}^d$ is the drift, and $\sigma_t > 0$ is the scalar diffusion coefficient. Under Assumption A2 below, this equation admits a non-explosive, pathwise unique solution [41], and its marginals $p_t^\star := \mathrm{Law}(\mathbf{X}_t)$ satisfy the forward Fokker–Planck equation [40, 9, 74]:

$$\partial_t p_t^\star = -\nabla\cdot\big(\boldsymbol{f}_t\, p_t^\star\big) + \frac{\sigma_t^2}{2}\,\Delta p_t^\star, \qquad p_0^\star = p^\star. \tag{3}$$

Because the data distribution $p^\star$ may be non-smooth or even singular, the score $\nabla \log p_t^\star$ can become unstable as $t \downarrow 0$ [15, 84]. Therefore, as is common in the literature [42, 7, 16], we work on a truncated interval $[t_0,T]$ with $t_0 > 0$, where the diffusion has already regularized the law: under suitable nondegeneracy and regularity assumptions [3, 4], $p_t^\star$ admits a smooth density for every $t > 0$, so that the score is well-defined and better behaved away from the singular endpoint.

Reverse-time diffusion. Under regularity conditions on the drift and diffusion coefficients (assumed throughout), the forward process admits a time reversal [33, 2, 12, 26]. The reverse-time SDE, running backwards from T down to t_0, is written in the standard backward-time convention $s : T \downarrow t_0$ (so that $\mathrm{d}s < 0$):

$$\mathrm{d}\mathbf{Y}_s = [\boldsymbol{f}_s(\mathbf{Y}_s) - \sigma_s^2\,\nabla_{\mathbf{x}} \log p_s^\star(\mathbf{Y}_s)]\,\mathrm{d}s + \sigma_s\,\mathrm{d}\bar{\mathbf{B}}_s, \qquad \mathbf{Y}_T \sim p_T^\star, \qquad s\colon T \downarrow t_0, \tag{4}$$

where $\bar{\mathbf{B}}_s$ is a Brownian motion under the reverse filtration. The marginals of this reverse SDE (4) exactly match those of the forward so that $\mathrm{Law}(\mathbf{Y}_s) = p_s^\star$ for all s, and they satisfy the backward Fokker-Planck equation (once again, backward in time *i.e.* $\mathrm{d}s < 0$):

$$\partial_s p_s^\star = -\nabla\cdot\Big(\big[\boldsymbol{f}_s - \sigma_s^2\,\nabla_{\mathbf{x}} \log p_s^\star\big]\, p_s^\star\Big) - \frac{\sigma_s^2}{2}\,\Delta p_s^\star, \qquad s\colon T \downarrow t_0, \tag{5}$$

Score estimation. In practice, the true score $\nabla_{\mathbf{x}} \log p_s^\star$ is unknown, and is estimated via score matching [38, 39] or denoising score matching (DSM) [78]. In modern diffusion and score-based generative modeling, DSM is used to learn a time-dependent score field $\mathbf{s}_\theta : \mathbb{R}^d \times [t_0,T] \to \mathbb{R}^d$ (parametrized by θ) of noise-perturbed marginals, by regressing toward the conditional score [71, 73, 72],

$$\mathcal{L}_{\mathrm{DSM}}(\theta) = \mathbb{E}_{s\sim\mathcal{U}[t_0,T]}\Big[\lambda(s)\mathbb{E}_{\mathbf{x}_0}\mathbb{E}_{\mathbf{x}_s|\mathbf{x}_0}\,\big\|\mathbf{s}_\theta(\mathbf{x}_s,s) - \nabla_{\mathbf{x}_s}\log p_s^\star(\mathbf{x}_s \mid \mathbf{x}_0)\big\|_2^2\Big], \tag{6}$$

where $\lambda : [0,T] \to \mathbb{R}_{>0}$ is a continuous positive time weighting. At the population level and under mild regularity, the minimizer of (6) approximates the true score field of the noised marginals.

Learned reverse process. Plugging the learned score into (4) defines the learned reverse process

$$\mathrm{d}\hat{\mathbf{Y}}_s = [\boldsymbol{f}_s(\hat{\mathbf{Y}}_s) - \sigma_s^2\,\mathbf{s}_\theta(\hat{\mathbf{Y}}_s,s)]\,\mathrm{d}s + \sigma_s\,\mathrm{d}\bar{\mathbf{B}}_s, \qquad \hat{\mathbf{Y}}_T \sim p_T^\star, \qquad s\colon T \to t_0. \tag{7}$$

Defining $\mathbf{e}_s(\mathbf{x}) := \mathbf{s}_\theta(\mathbf{x},s) - \nabla_{\mathbf{x}} \log p_s^\star(\mathbf{x})$ as the vector field of *score estimation error*, the marginals $\hat{p}_s := \mathrm{Law}(\hat{\mathbf{Y}}_s)$ of the reverse process (7) satisfy the backward Fokker-Planck equation:

$$\partial_s \hat{p}_s = -\nabla\cdot\Big(\big(\boldsymbol{f}_s - \sigma_s^2\,\nabla_{\mathbf{x}} \log p_s^\star\big)\,\hat{p}_s\Big) - \frac{\sigma_s^2}{2}\,\Delta\hat{p}_s + \sigma_s^2\,\nabla\cdot\big(\mathbf{e}_s\,\hat{p}_s\big), \quad \hat{p}_T = p_T^\star, \quad s\colon T \to t_0. \tag{8}$$

The learned generative distribution is $\hat{p}_{t_0} := \mathrm{Law}(\hat{\mathbf{Y}}_{t_0})$. Moreover, writing $\nabla_{\mathbf{x}} \log p_s^\star(\mathbf{x}) = \mathbf{s}_\theta(\mathbf{x}, s) - \mathbf{e}_s(\mathbf{x})$ in the true reverse Fokker–Planck (5) yields,

$$\partial_s p_s^\star = -\nabla\cdot\Big(\big[\boldsymbol{f}_s - \sigma_s^2\,\mathbf{s}_\theta(\cdot, s)\big]\,p_s^\star\Big) - \frac{\sigma_s^2}{2}\,\Delta p_s^\star - \sigma_s^2\nabla\cdot(p_s^\star \mathbf{e}_s), \qquad s\colon T \downarrow t_0. \tag{9}$$

Weighted divergence-free fields. For a density μ, a field $\mathbf{w} \in L^2(\mu; \mathbb{R}^d)$ is called divergence-free if $\nabla\cdot(\mu\mathbf{w}) = 0$ in the weak sense, namely $\int_{\mathbb{R}^d} \nabla\varphi(\mathbf{x}) \cdot \mathbf{w}(\mathbf{x})\,\mu(\mathbf{x})\,\mathrm{d}\mathbf{x} = 0$ for all $\varphi \in C_c^\infty(\mathbb{R}^d)$.

Assumptions. We make four assumptions, which are standard in the literature on diffusion-model convergence [15, 81, 83, 14] and required in the proofs of Theorems 3.1 and 3.2.

(A1) Marginal regularity. For all $s \in [t_0, T]$, the marginals $p_s^\star$ and $\hat{p}_s$ are strictly positive, C^1 probability densities on $\mathbb{R}^d$ with finite second moment.

(A2) Regularity of Drift, Diffusion coefficient, and Score Errors. The map $s \mapsto \sigma_s$ is continuous on $[t_0, T]$ with $0 < \underline{\sigma} \le \sigma_s \le \bar{\sigma}$. The maps $(s, \mathbf{x}) \mapsto \boldsymbol{f}_s(\mathbf{x})$, $\mathbf{s}_\theta(\mathbf{x}, s)$, $\mathbf{e}_s(\mathbf{x})$ are (i) jointly Borel measurable, (ii) locally Lipschitz in $\mathbf{x}$, and (iii) of linear growth: there exists $C < \infty$ such that, for a.e. $s \in [t_0, T]$ and all $\mathbf{x} \in \mathbb{R}^d$, $\|\boldsymbol{f}_s(\mathbf{x})\| + \|\mathbf{s}_\theta(\mathbf{x}, s)\| + \|\mathbf{e}_s(\mathbf{x})\| \le C(1 + \|\mathbf{x}\|)$.

(A3) Density regularity. The maps $(s, \mathbf{x}) \mapsto p_s^\star(\mathbf{x})$ and $(s, \mathbf{x}) \mapsto \hat{p}_s(\mathbf{x})$ are of class $C^{1,2}$ on $[t_0, T] \times \mathbb{R}^d$. With $r_s := p_s^\star/\hat{p}_s$,

$$\int_{t_0}^{T}\int_{\mathbb{R}^d} \Big(|\partial_s p_s^\star|\,\big(1 + |\log r_s|\big) + |\partial_s \hat{p}_s| r_s\Big)\mathrm{d}\mathbf{x}\,\mathrm{d}s < \infty.$$

(A4) Pointwise score integrability and tails. For a.e. $s \in [t_0, T]$, $\nabla \log \hat{p}_s$ has linear growth, and $\nabla \log p_s^\star,\ \nabla \log \hat{p}_s \in L^2(p_s^\star; \mathbb{R}^d)$.

Role of the assumptions Assumption A2 gives existence of unique non-explosive solutions to both reverse SDEs, and uniqueness of the associated Fokker–Planck curves. Together with the finite second moment in A1, it also implies $\mathbf{e}_s \in L^2(p_s^\star) \cap L^2(\hat{p}_s)$. Assumption A3 justifies differentiating $s \mapsto \mathrm{KL}(p_s^\star \| \hat{p}_s)$ under the integral (Lemma B.3). Assumption A4 ensures that the relative Fisher information $\mathcal{I}(p_s^\star \| \hat{p}_s)$ is finite for a.e. s, and that all integrations by parts are valid (Lemma B.3).

3 Geometric Structure of Score Errors: Observable and Invisible Components

The true densities $(p_s^\star)_{s\in[t_0,T]}$ evolve according to the backward Fokker-Planck given by (9), in which the score estimation errors $(\mathbf{e}_s)_{s\in[t_0,T]}$ enter exclusively through the scalar weighted divergence terms $(\nabla\cdot(p_s^\star \mathbf{e}_s))_{s\in[t_0,T]}$. This induces an invariance at the level of marginals: two errors with the same weighted divergences generate the same marginal dynamics. In particular, adding any divergence-free perturbation $\mathbf{w}_s \in L^2(p_s^\star; \mathbb{R}^d)$, i.e. any $\mathbf{w}_s$ such that $\nabla\cdot(p_s^\star \mathbf{w}_s) = 0$, leaves the backward Fokker-Planck unchanged. We therefore refer to these locally mass-preserving perturbations as *invisible*: they may contribute to the ambient L^2 score error, but they do not affect the reverse marginal dynamics.

For clarity, in this section we fix a time $s \in [t_0, T]$ in the reverse diffusion and study how quantities behave at that time.

Helmholtz-Hodge Decomposition. To characterize these invisible perturbations, we observe that for any differentiable vector field $\mathbf{w}_s$, the following are equivalent (by weak integration by parts):

$$\nabla\cdot\big(p_s^\star \mathbf{w}_s\big) = 0 \iff \text{for all } \varphi \in C_c^\infty(\mathbb{R}^d),\ \text{we have } \langle\nabla\varphi, \mathbf{w}_s\rangle_{L^2(p_s^\star;\mathbb{R}^d)} = 0 \iff \mathbf{w}_s \in \mathcal{G}_s^\perp, \tag{10}$$

where $\mathcal{G}_s := \mathrm{cl}(\{\nabla\varphi :\ \varphi \in C_c^\infty(\mathbb{R}^d)\})$ and cl denotes the closure with respect to $L^2(p_s^\star; \mathbb{R}^d)$; the inner product notation in (10) is defined on p.3. Thus, the set of vector fields $\mathbf{w}_s$ such that $\nabla\cdot\big(\mathbf{w}_s p_s^\star\big) = 0$ is the orthogonal complement of gradient fields in the space $L^2(p_s^\star; \mathbb{R}^d)$. We can decompose $L^2(p_s^\star; \mathbb{R}^d)$ into a gradient subspace $\mathcal{G}_s$ and its orthogonal complement $\mathcal{G}_s^\perp$. Expressing the error using this decomposition yields the *weighted Helmholtz–Hodge* decomposition [8]:

$$\mathbf{e}_s = \Pi_{\mathcal{G}_s}\mathbf{e}_s + \Pi_{\mathcal{G}_s^\perp}\mathbf{e}_s \in L^2(p_s^\star; \mathbb{R}^d) = \mathcal{G}_s \oplus \mathcal{G}_s^\perp \tag{11}$$

Capacity	Params	$\rho(\text{FID}, \hat{\mathcal{E}}_{\text{full}})$	$\rho(\text{FID}, \hat{\mathcal{E}}_{\text{grad}})$
tiny	1.76M	0.76 ± 0.14	$\mathbf{0.97 \pm 0.03}$
small	3.92M	0.89 ± 0.01	$\mathbf{0.96 \pm 0.02}$
full	6.27M	0.94 ± 0.03	$\mathbf{0.97 \pm 0.02}$

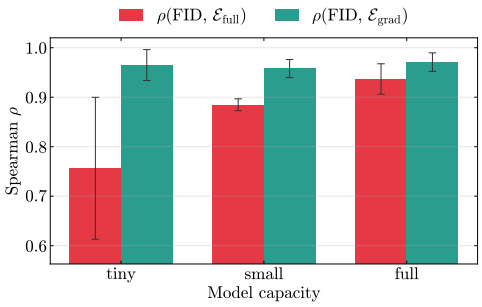

Figure 2: **Spearman correlation between feature-FID [34] and score-error norms (Fashion-MNIST [82]).** Mean $\pm$ standard-deviation across three seeds. Across all capacities, the finite-critic estimate $\hat{\mathcal{E}}_{\text{grad}} := \widehat{\mathcal{L}}_{\text{obs}}$ of $\mathbb{E}_{s,\mathbf{x}_s}\left[\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|_2^2\right]$ (Algorithm 1; a lower estimate at finite critic capacity) is uniformly more correlated with sample quality than the full error estimate $\hat{\mathcal{E}}_{\text{full}} := \mathbb{E}_{s,\mathbf{x}_s}\left[\|\mathbf{e}_s\|_2^2\right]$, with the largest gap at low capacity. Setup details in Appendix E.1.

where $\Pi_{\mathcal{G}_s}$ (resp. $\Pi_{\mathcal{G}_s^\perp}$) denotes the orthogonal projection operator on $\mathcal{G}_s$ (resp. on $\mathcal{G}_s^\perp$). The second component is naturally divergence-free and thus structurally invisible at the reference density since $\nabla\cdot(p_s^\star \mathbf{e}_s) = \nabla\cdot\big(p_s^\star \Pi_{\mathcal{G}_s}\mathbf{e}_s\big)$. Plugging (11) in the backward Fokker-Planck (9) implies,

$$\partial_s p_s^\star = -\nabla\cdot\Big(\big[\boldsymbol{f}_s - \sigma_s^2\,\mathbf{s}_\theta(\cdot,s)\big]\,p_s^\star\Big) - \frac{\sigma_s^2}{2}\,\Delta p_s^\star - \sigma_s^2\nabla\cdot\big(p_s^\star \Pi_{\mathcal{G}_s}\mathbf{e}_s\big), \qquad s : T \downarrow t_0. \tag{12}$$

Intuitively, the component $\Pi_{\mathcal{G}_s}\mathbf{e}_s$ corresponds to a *transport of mass across space*, which modifies the density. In contrast, $\Pi_{\mathcal{G}_s^\perp}\mathbf{e}_s$ preserves mass locally (e.g. rotation) without affecting the density. As a result, two error fields with identical L^2-norm can induce radically different effects on the generated distribution, as illustrated in Figure 1.

The following result summarizes the discussion above and its consequences (proof in Appendix B.2).

Theorem 3.1. *Suppose A1–A4 hold, and let*

$$\mathbf{e}_s = \mathbf{e}_{\text{obs},s} + \mathbf{e}_{\text{inv},s}, \qquad \mathbf{e}_{\text{obs},s} := \Pi_{\mathcal{G}_s}\mathbf{e}_s, \quad \mathbf{e}_{\text{inv},s} := \Pi_{\mathcal{G}_s^\perp}\mathbf{e}_s.$$

Consider the backward Fokker-Planck equation,

$$\partial_s q_s = -\nabla\cdot\Big(\big[\boldsymbol{f}_s - \sigma_s^2\,\mathbf{s}_\theta(\cdot,s)\big]\,q_s\Big) - \frac{\sigma_s^2}{2}\Delta q_s - \sigma_s^2\nabla\cdot\big(\mathbf{v}_s\,q_s\big), \qquad q_T = p_T^\star, \tag{13}$$

for $\mathbf{v}_s = \mathbf{e}_s$ and for $\mathbf{v}_s = \mathbf{e}_{\text{obs},s}$. Then:

(i) Only the observable component enters the marginal dynamics: *the target marginal curve $(p_s^\star)_s$ solves (13) both for $\mathbf{v}_s = \mathbf{e}_s$ and for $\mathbf{v}_s = \mathbf{e}_{\text{obs},s}$. In particular, it coincides with any solution for $\mathbf{v}_s = \mathbf{e}_{\text{obs},s}$ whenever that solution is unique.*

(ii) Purely invisible errors: *if $\mathbf{e}_{\text{obs},s} \equiv 0$ and $\mathbf{e}_{\text{inv},s} \not\equiv 0$, then $\hat{p}_s = p_s^\star$ for $s \in [t_0, T]$, while*

$$\int_{t_0}^{T} \|\mathbf{e}_s\|^2_{L^2(p_s^\star)}\,\mathrm{d}s = \int_{t_0}^{T} \|\mathbf{e}_{\text{inv},s}\|^2_{L^2(p_s^\star)}\,\mathrm{d}s > 0.$$

(iii) No lower bound: *Assume $d \geq 2$. For any divergence $\mathrm{Div}(\cdot\|\cdot)$ between probability distributions such that $\mathrm{Div}(\mu\|\nu) = 0$ iff $\mu = \nu$, there is no lower bound of the form*

$$\mathrm{Div}(p_{t_0}^\star \| \hat{p}_{t_0}) \;\geq\; F\left(\int_{t_0}^{T} \|\mathbf{e}_s\|^2_{L^2(p_s^\star)}\,\mathrm{d}s\right)$$

that holds uniformly over all score errors $\mathbf{e}$, for any strictly increasing function $F : [0,\infty) \to \mathbb{R}$ with $F(0) = 0$.

Theorem 3.1 implies that the ambient L^2 score error is not the intrinsic quantity controlling sampling quality. Indeed the theorem shows that purely solenoidal errors may have arbitrarily large L^2 norm

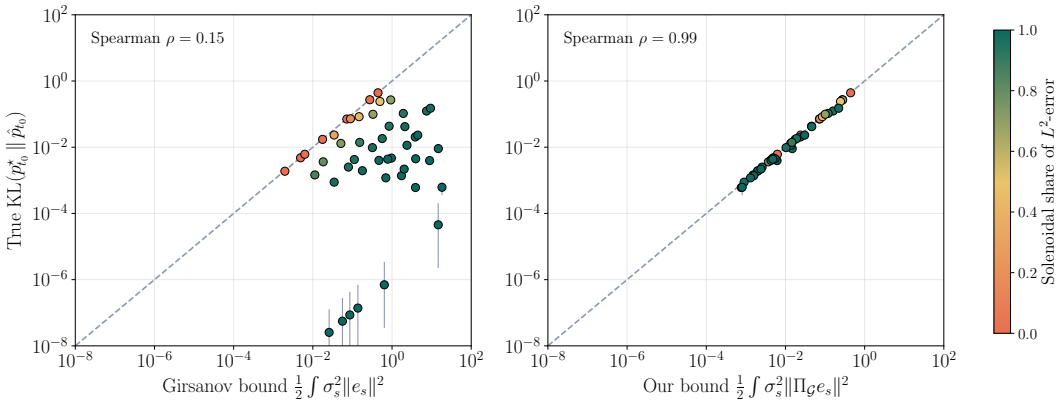

Figure 3: **Exact low-dimensional comparison of standard Girsanov bound with that of Theorem 3.2.** The target is an 8-mode anisotropic Gaussian mixture in $\mathbb{R}^2$ chosen so that forward marginals and scores are available in closed form. Each point is a synthetic score error $\mathbf{e}_s = a\,\nabla\psi_s + b\,J\nabla \log p^\star_s$, where the components are chosen so that the Helmholtz–Hodge decomposition is known: $\Pi_{\mathcal{G}_s}\mathbf{e}_s = a\nabla\psi_s$ and $\Pi_{\mathcal{G}_s^\perp}\mathbf{e}_s = b\,J\nabla \log p^\star_s$. The true $\mathrm{KL}(p^\star_{t_0}\|\hat{p}_{t_0})$ is computed by solving the learned reverse Fokker–Planck equation on a fine grid (error bars: discretisation error). Colour: solenoidal share of L^2-error. *Left:* The true KL has low Spearman correlation with the full-L^2 (Girsanov) bound, consistent with Theorem 3.1 (iii): purely solenoidal errors ($a = 0$, dark green dots) are taken into account by the Girsanov bound, while they do not affect the true KL (solver error $\approx 10^{-5}$). *Right:* Every point lies below the diagonal within numerical error, as guaranteed by Theorem 3.2, and the true KL is almost identical to our bound (Spearman $\rho = 0.99$). Setup in Appendix E.4.

while leaving all marginals unchanged, and that no distributional discrepancy can be uniformly lower-bounded by the full score error alone. Notably, this implies that the classical bound

$$\mathrm{KL}(p^\star_{t_0} \parallel \hat{p}_{t_0}) \;\le\; \frac{1}{2}\int_{t_0}^{T} \sigma_s^2 \|\mathbf{e}_s\|^2_{L^2(p^\star_s)}\mathrm{d}s \tag{14}$$

may be loose whenever score errors have large invisible (solenoidal) components.

By contrast, Theorem 3.1 identifies the gradient projection of the error, $\Pi_{\mathcal{G}_s}\mathbf{e}_s$, as the component entering the Fokker-Planck dynamics at time s. Empirically, Figure 2 shows that the critic estimate of $\|\Pi_{\mathcal{G}}\mathbf{e}\|^2$ is a better proxy of sampling quality than the full error $\|\mathbf{e}\|^2$ as it correlates better with the feature Fréchet Inception Distance (FID) [34], widely used to measure sampling quality of generative models for images. (The code for all the experiments reported here is available on GitHub.)

Improved upper bound. Based on Theorem 3.1, we replace the ambient L^2 geometry by an observable geometry, to obtain the following sharper endpoint upper bound (proof in Appendix B.3).

Theorem 3.2. *Under Assumptions A1–A4,*

$$\mathrm{KL}(p^\star_{t_0} \parallel \hat{p}_{t_0}) \;\le\; \frac{1}{2}\int_{t_0}^{T} \sigma_s^2 \|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p^\star_s)}\mathrm{d}s. \tag{15}$$

Theorem 3.2 identifies the component of the error field that controls the endpoint divergence to be the $L^2(p^\star_s)$-projection of the error onto the subspace of gradient fields.

Remarks 1. *Link with Girsanov-based approach.* In contrast with the Girsanov-based approach, this new bound is a direct consequence of the corrected Fokker-Planck dynamics (12) (discussed in Appendix B.3). Appendix C reconciles the two by deriving the same bound via Girsanov's theorem applied to a *marginally-equivalent representative* of the learned process whose drift retains only the observable component. This identifies the standard bound's looseness as the cost of operating on path-space rather than marginal-space dynamics.

2. *Conservative parametrizations.* If the learned score vector field is parametrized as $\mathbf{s}_\theta(\mathbf{x}, s) = \nabla_{\mathbf{x}} E_\theta(\mathbf{x}, s)$ for a scalar potential E_θ (e.g. as in energy-based models [50]), then $\mathbf{e}_s = \nabla(E_\theta - \log p^\star_s)$

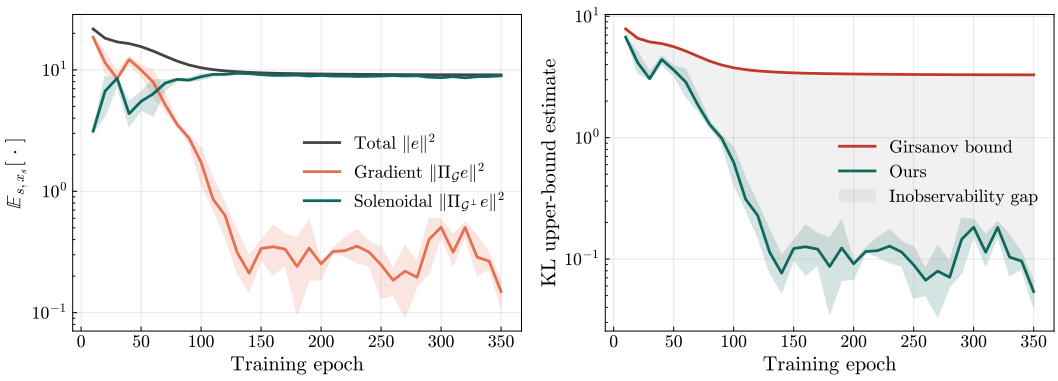

Figure 4: **DSM does not reduce the invisible component of the score error on CIFAR-10 [46].** *Left.* The gradient (observable) component of score estimation errors falls by roughly two orders of magnitude during training, while the solenoidal (invisible) component rises early and then plateaus, eventually accounting for nearly all of the residual error. *Right.* As a consequence, the standard Girsanov bound (14) saturates at the solenoidal floor while our bound (Theorem 3.2) drops by two orders of magnitude and stays below it. The shaded *inobservability gap* is what the standard analysis misses by penalizing error components invisible to marginals (Theorem 3.1). The gradient component is the finite-critic estimate $\widehat{\mathcal{L}}_{\text{obs}}$ (a lower estimate), and the solenoidal component is obtained as the full error minus this estimate (hence an upper estimate). Median (solid) and inter-quartile range (band) over four seeds. Setup details in Appendix E.3. The same observations hold for Fashion-MNIST (Figures 5, 6, Appendix E.2).

is a pure gradient field without any solenoidal component. In that case, the standard Girsanov bound coincides with (15). For unconstrained architectures (e.g. U-Nets [65, 35] or Transformers [76, 62]), which dominate practice, the gap between $\|\mathbf{e}_s\|_{L^2(p_s^\star)}$ and $\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|_{L^2(p_s^\star)}$ can be substantial and our new bound is tighter and correlates more strongly with the Fréchet Inception Distance, as shown in Figure 2.

3. *Gain with new upper bound.* The new upper bound (15) is expected to significantly improve on the standard one (14) when the gradient components of the errors have small norms compared to the solenoidal components. Figure 1 illustrates this: by artificially creating solenoidal errors, our upper bound is not affected while that given in (14) explodes.

Figure 3 makes this quantitative in a setting where the decomposition and the endpoint KL are known exactly and Figure 4 illustrates this effect during training: after an initial phase in which the observable gradient component is reduced, the remaining score error is dominated by the solenoidal component, creating a persistent gap between the classical full-error bound and the observable bound.

4 From Observable Bounds to a Computable Diagnostic

This section introduces a practical procedure to estimate the gradient component of score estimation errors. Theorem 3.2 bounds the divergence between the target and learned data distributions using the $L^2(p_s^\star)$-gradient projection $\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|_{L^2(p_s^\star)}$. Computing expectations with respect to $p_s^\star$ is practical since one has access to samples from the target distribution during training. However, explicitly projecting errors over the space of gradient fields is computationally hard. In this section, we present an approach that avoids explicit projections by characterizing the gradient component of $\mathbf{e}_s$ through an H^{-1} dual variational identity.

4.1 A Variational Reformulation

Recall from the Fokker-Planck equation (9) that at each time s, the dynamics of the true densities depend on $\mathbf{e}_s$ only through $\nabla\cdot(p_s^\star\,\mathbf{e}_s)$. Thus any vector field $\mathbf{v}_s$ such that $\nabla\cdot(p_s^\star\,\mathbf{e}_s) = \nabla\cdot(p_s^\star\,\mathbf{v}_s)$ induces the same effect on marginals; we refer to such fields as admissible. In Section 3, we observed that $\nabla\cdot(p_s^\star\mathbf{e}_s) = \nabla\cdot(p_s^\star\Pi_{\mathcal{G}_s}\mathbf{e}_s)$, making the gradient-projection of errors one such admissible vector

field, i.e. $\Pi_{\mathcal{G}_s}\mathbf{e}_s \in \{\mathbf{v}_s : \nabla\cdot(p_s^\star\mathbf{e}_s) = \nabla\cdot(p_s^\star\mathbf{v}_s)\}$. Moreover, since $\Pi_{\mathcal{G}_s}\mathbf{e}_s$ is the only observable part of $\mathbf{e}_s$, it contains the minimal and irreducible energy of $\mathbf{e}_s$ that is visible through $\nabla\cdot(p_s^\star\mathbf{e}_s)$.

Proposition 4.1. *For each $s \in [t_0, T]$,*

$$\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p_s^\star)} = \inf_{\mathbf{v}_s:\,\nabla\cdot(p_s^\star\mathbf{v}_s)=\nabla\cdot(p_s^\star\mathbf{e}_s)} \int_{\mathbb{R}^d} \|\mathbf{v}_s(\mathbf{x})\|_2^2\, p_s^\star(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{16}$$

Proof. Consider any admissible field $\mathbf{v}_s$, i.e. such that $\nabla\cdot(p_s^\star\mathbf{v}_s) = \nabla\cdot(p_s^\star\mathbf{e}_s)$. Then, $\nabla\cdot\big(p_s^\star(\mathbf{v}_s - \Pi_{\mathcal{G}_s}\mathbf{e}_s)\big) = 0$ hence $\mathbf{v}_s - \Pi_{\mathcal{G}_s}\mathbf{e}_s \in \mathcal{G}_s^\perp$. By orthogonality, $\|\mathbf{v}_s\|^2_{L^2(p_s^\star)} = \|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p_s^\star)} + \|\mathbf{v}_s - \Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p_s^\star)} \geq \|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p_s^\star)}$. Thus, $\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p_s^\star)} = \inf_{\mathbf{v}_s:\,\nabla\cdot(p_s^\star\mathbf{v}_s)=\nabla\cdot(p_s^\star\mathbf{e}_s)} \|\mathbf{v}_s\|^2_{L^2(p_s^\star)}$. □

In other words, among all vector fields $\mathbf{v}_s$ that induce the weighted divergence $\nabla\cdot(p_s^\star\mathbf{v}_s) = \nabla\cdot(p_s^\star\mathbf{e}_s)$, the field $\Pi_{\mathcal{G}_s}\mathbf{e}_s$ is the unique minimum-energy representative in $L^2(p_s^\star;\mathbb{R}^d)$. Notably, the right-hand term of (16) does not require computing a projection, and just requires minimizing $L^2(p_s^\star)$-norms. However, this right-hand term is still not satisfactory since it requires solving a Poisson equation $\nabla\cdot(p_s^\star\mathbf{v}_s) = \nabla\cdot(p_s^\star\mathbf{e}_s)$ which is known to be computationally intractable in high dimensions due to the curse of dimensionality inherent in standard numerical PDE solvers [30]. Fortunately, the optimization problem in (16) exactly corresponds to an H^{-1} norm [10], which enjoys a tractable dual characterization. The $H^{-1}(p_s^\star)$-norm of $\nabla\cdot(p_s^\star\mathbf{e}_s)$ is defined as

$$\|\nabla\cdot(p_s^\star\mathbf{e}_s)\|^2_{H^{-1}(p_s^\star)} := \inf_{\mathbf{v}_s:\,\nabla\cdot(p_s^\star\mathbf{e}_s)=\nabla\cdot(p_s^\star\mathbf{v}_s)} \int_{\mathbb{R}^d} \|\mathbf{v}_s(\mathbf{x})\|_2^2\, p_s^\star(\mathbf{x})\,\mathrm{d}\mathbf{x}, \tag{17}$$

and can be expressed in its dual form as [10],

$$\|\nabla\cdot(p_s^\star\mathbf{e}_s)\|^2_{H^{-1}(p_s^\star)} = \sup_{\varphi\in C_c^\infty(\mathbb{R}^d)} \left\{-2\int_{\mathbb{R}^d} \varphi\,\nabla\cdot(p_s^\star\mathbf{e}_s)\,\mathrm{d}\mathbf{x} - \int_{\mathbb{R}^d} \|\nabla\varphi\|_2^2\, p_s^\star\,\mathrm{d}\mathbf{x}\right\}. \tag{18}$$

From (16)-(18), we have

$$\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p_s^\star)} = \sup_{\varphi\in C_c^\infty(\mathbb{R}^d)} \left\{-2\int_{\mathbb{R}^d} \varphi\,\nabla\cdot(p_s^\star\mathbf{e}_s)\,\mathrm{d}\mathbf{x} - \int_{\mathbb{R}^d} \|\nabla\varphi\|_2^2\, p_s^\star\,\mathrm{d}\mathbf{x}\right\}. \tag{19}$$

These H^{-1}-norms have a natural geometric interpretation in optimal transport theory (discussed in Appendix D). Equation (19) is an unconstrained optimization problem over the set of smooth and scalar-valued test functions, and directly yields our method, described below, to estimate $\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p_s^\star)}$.

4.2 Estimation of the Observable Error

The variational characterization of $\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p_s^\star)}$ in (19) suggests a clear training procedure to compute this norm: aim to solve the optimization by learning a scalar-valued critic network φ_ψ with parameters ψ, via gradient ascent on the objective function in (19). In the remainder of this section, we adopt the standard framework of variance-preserving forward diffusion [73] (i.e. we fix $\boldsymbol{f}_s(\mathbf{x}) = -\frac{1}{2}\beta(s)\mathbf{x}$ and $\sigma_s = \sqrt{\beta(s)}$ in (2) for some schedule $\beta : [t_0, T] \to \mathbb{R}_{>0}$), widely used in practical training of diffusion models. Applying integration by parts to the first integral in (19), we obtain

$$\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p_s^\star)} = \sup_{\varphi\in C_c^\infty(\mathbb{R}^d)} \left\{2\,\mathbb{E}_{\mathbf{x}\sim p_s^\star}\big[\nabla\varphi(\mathbf{x})\cdot\mathbf{e}_s(\mathbf{x})\big] - \mathbb{E}_{\mathbf{x}\sim p_s^\star}\big[\|\nabla\varphi(\mathbf{x})\|_2^2\big]\right\}, \tag{20}$$

which only requires expectations over $p_s^\star$, accessible during training by sampling noised data $\mathbf{X}_s \sim p_s^\star$. We introduce a scalar-valued neural network $\varphi_\psi(\mathbf{x}, s)$ with parameters ψ, and define

$$\mathcal{J}(\psi) := \mathbb{E}_{s\sim\nu}\,\mathbb{E}_{\mathbf{x}\sim p_s^\star}\Big[2\,\nabla_{\mathbf{x}}\varphi_\psi(\mathbf{x}, s)\cdot\mathbf{e}_s(\mathbf{x}) - \|\nabla_{\mathbf{x}}\varphi_\psi(\mathbf{x}, s)\|_2^2\Big], \tag{21}$$

where ν is a distribution over $[t_0, T]$. A finite critic class yields a *lower* estimate of the observable energy, with equality for a sufficiently expressive class [36, 56] under exact optimization. Indeed, if $\nabla\varphi \in \mathcal{G}_s$, then $\nabla\varphi \perp \Pi_{\mathcal{G}_s^\perp}\mathbf{e}_s$ in $L^2(p_s^\star)$, so that $2\langle\nabla\varphi, \mathbf{e}_s\rangle - \|\nabla\varphi\|^2 = \|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2 - \|\nabla\varphi - \Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2$. Hence $\mathcal{J}(\psi) \leq \mathbb{E}_{s\sim\nu}\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p_s^\star)}$, with equality if and only if $\nabla_{\mathbf{x}}\varphi_\psi(\cdot, s) = \Pi_{\mathcal{G}_s}\mathbf{e}_s$ for ν-a.e. s.

To avoid the dependence of (21) on the unknown true score $\nabla_{\mathbf{x}} \log p^\star$ through $\mathbf{e}_s = \mathbf{s}_\theta(\cdot, s) - \nabla_{\mathbf{x}} \log p_s^\star$, we exploit the DSM identity [78]: for $\alpha(s) = \exp(-\frac{1}{2}\int_0^s \beta(u)\mathrm{d}u)$ and $\sigma(s)^2 = 1-\alpha(s)^2$, one has that $\mathbf{X}_s = \alpha(s)\mathbf{X}_0 + \sigma(s)\boldsymbol{\varepsilon}$ with $\mathbf{X}_0 \sim p^\star$, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I}_d)$, and

$$\nabla_{\mathbf{x}} \log p_s^\star(\mathbf{x}) = \mathbb{E}\big[\mathbf{s}_{\text{target}}(\mathbf{X}_s, s) \,\big|\, \mathbf{X}_s = \mathbf{x}\big], \quad \text{where} \quad \mathbf{s}_{\text{target}}(\mathbf{X}_s, s) = -\frac{\mathbf{X}_s - \alpha(s)\mathbf{X}_0}{\sigma(s)^2}. \tag{22}$$

Consequently, the residual $\mathbf{r}_\theta(\mathbf{x}, s) := \mathbf{s}_\theta(\mathbf{x}, s) - \mathbf{s}_{\text{target}}(\mathbf{x}, s)$ is an unbiased proxy for the score error in the sense that $\mathbb{E}[\mathbf{r}_\theta(\mathbf{X}_s, s) \mid \mathbf{X}_s = \mathbf{x}] = \mathbf{e}_s(\mathbf{x})$. Plugging this proxy in (21) yields,

$$\widehat{\mathcal{J}}(\psi) := \mathbb{E}_{s\sim\nu, \mathbf{X}_0, \boldsymbol{\varepsilon}}\Big[2\,\nabla_{\mathbf{x}}\varphi_\psi(\mathbf{X}_s, s) \cdot \mathbf{r}_\theta(\mathbf{X}_s, s) - \|\nabla_{\mathbf{x}}\varphi_\psi(\mathbf{X}_s, s)\|^2\Big], \quad \mathbf{X}_s = \alpha(s)\mathbf{X}_0 + \sigma(s)\boldsymbol{\varepsilon}. \tag{23}$$

The tower property gives $\mathbb{E}[\widehat{\mathcal{J}}(\psi)] = \mathcal{J}(\psi)$ for every *fixed* ψ. We then train our critic network φ_ψ to solve the following maximization problem,

$$\hat{\mathcal{L}}_{\text{obs}} := \sup_{\psi} \widehat{\mathcal{J}}(\psi). \tag{24}$$

Intuitively, optimizing (24) makes the critic search for the *steepest gradient direction* for the frozen score network (i.e., the direction that best exposes the gradient component of the score error). Algorithm 1 in Appendix E.3 summarizes the resulting estimator.

Uses of this diagnostic. We use $\widehat{\mathcal{L}}_{\text{obs}}$ as a diagnostic rather than as a training objective, decoupling critic estimation from score-network optimization. The diagnostic can be applied in two ways.

During training: one may periodically refit the critic φ_ψ against the frozen current score network $\hat{\mathbf{s}}_\theta$, so that $\widehat{\mathcal{L}}_{\text{obs}}$ tracks the gradient component of the score error along the optimization trajectory.

Post hoc: once fully trained, the score network can be frozen and the critic optimized more thoroughly to estimate the upper bound in Theorem 3.2. This is precisely the procedure that produces the gradient-component curves of Figures 2, 4, 5 and 6: at each checkpoint we freeze the score network, fit a fresh critic via (23), and report $\widehat{\mathcal{L}}_{\text{obs}}$ as our estimate of the observable error (full setup, hyperparameters, and a critic-suboptimality ablation in Appendices E.1–E.3). A single evaluation of $\widehat{\mathcal{L}}_{\text{obs}}$ takes a few minutes on one GPU and adds roughly 5–10% overhead to standard DSM training, with cost driven by critic optimization, since the estimator only requires forward-noised samples $\mathbf{X}_s = \alpha(s)\mathbf{X}_0 + \sigma(s)\boldsymbol{\varepsilon}$. Unlike sample-based metrics such as FID [34], the diagnostic $\widehat{\mathcal{L}}_{\text{obs}}$ targets a quantitative upper bound on $\mathrm{KL}(p_{t_0}^\star \| \hat{p}_{t_0})$ rather than perceptual quality and avoids reverse-SDE sampling entirely, making it complementary to FID and applicable in settings where sampling is expensive or where domain-specific feature extractors are unavailable.

5 Discussion and Future Work

The bound of Theorem 3.2 on $\mathrm{KL}(p_{t_0}^\star \parallel \hat{p}_{t_0})$ integrates the instantaneous gradient component of the errors along the path, while the sampling distribution $\hat{p}_{t_0}$ is shaped by how these gradient errors accumulate at the endpoint. We expect the diffusive part of the reverse SDE to have a smoothing effect, with some gradient errors injected at time s dissipated before reaching $\hat{p}_{t_0}$, and others amplified. Identifying which gradient components along the path affect the endpoint marginal is a fundamental open question, as pointed out in [42], and a sharper notion of endpoint observability building on our geometric framework would refine the bound further.

Beyond this refinement, our decomposition admits a natural interpretation in Wasserstein-2 geometry: $\Pi_{\mathcal{G}_s}\mathbf{e}_s$ is the component of the score error that lies in the tangent space at $p_s^\star$ of the Otto manifold; the solenoidal part lies in the orthogonal complement and does not move mass. In addition to providing geometric intuition, this perspective opens connections to gradient flows in Wasserstein space and to geometry-aware regularizers. Finally, our estimator of the norm of the gradient component of errors is currently used only as a training diagnostic. It freezes the score network to track gradient components during training. This could be extended into a training objective, either in combination with the standard DSM loss, or to make the critic guide the score network towards observable directions that translate into better learning and better capacity. We note that this raises subtle min-max optimization questions, since errors in estimating the critic could contaminate the score-network training signal if not carefully addressed.

Acknowledgements

NBK is supported by a G-Research Trinity College Studentship, and RET is supported by the EPSRC Probabilistic AI Hub (EP/Y028783/1). RV was supported in part by an EPSRC Mathematical Sciences Small Grant.

References

[1] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Birkhäuser, Basel, second edition, 2008.

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

[3] Donald G. Aronson. Bounds for the fundamental solution of a parabolic equation. *Bulletin of the American Mathematical Society*, 73(6):890–896, 1967.

[4] Donald G. Aronson. Non-negative solutions of linear parabolic equations. *Annali della Scuola Normale Superiore di Pisa, Classe di Scienze (3)*, 22(4):607–694, 1968.

[5] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[6] V. E. Beneš. Existence of optimal stochastic control laws. *SIAM Journal on Control*, 9(3):446–472, 1971.

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

[8] Harsh Bhatia, Gregory Norgard, Valerio Pascucci, and Peer-Timo Bremer. The Helmholtz-Hodge decomposition—a survey. *IEEE Transactions on Visualization and Computer Graphics*, 19(8):1386–1404, 2013.

[9] Vladimir I. Bogachev, Nicolai V. Krylov, Michael Röckner, and Stanislav V. Shaposhnikov. *Fokker–Planck–Kolmogorov Equations*, volume 207 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2015.

[10] Haïm Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer, New York, 2011.

[11] R. H. Cameron and W. T. Martin. Transformations of Wiener integrals under translations. *Annals of Mathematics*, 45(2):386–396, 1944.

[12] Patrick Cattiaux, Giovanni Conforti, Ivan Gentil, and Christian Léonard. Time reversal of diffusion processes under a finite entropy condition. *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 59(4):1844–1881, 2023.

[13] Chen-Hao Chao, Wei-Fang Sun, Bo-Wun Cheng, and Chun-Yi Lee. On investigating the conservative property of score-based generative models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 4076–4095. PMLR, 2023.

[14] Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 4735–4763. PMLR, 2023.

[15] Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 4672–4712. PMLR, 2023.

[16] 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 (ICLR)*, 2023.

[17] Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. *Statistical Optimal Transport: École d'Été de Probabilités de Saint-Flour XLIX – 2019*, volume 2364 of *Lecture Notes in Mathematics*. Springer, Cham, 2025.

[18] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11472–11481, 2022.

[19] Valentin De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022.

[20] C. Doléans-Dade. Quelques applications de la formule de changement de variables pour les semimartingales. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 16(3):181–194, 1970.

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

[22] Benjamin Dupuis, Tyler Farghly, Maxime Haddouche, Alain Oliviero Durmus, and Umut Simsekli. Tightening the score matching gap for diffusion models. In *Forty-third International Conference on Machine Learning*, 2026.

[23] Weinan E and Bing Yu. The Deep Ritz method: A deep learning-based numerical algorithm for solving variational problems. *Communications in Mathematics and Statistics*, 6(1):1–12, 2018.

[24] Lawrence C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.

[25] Lawrence C. Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. Textbooks in Mathematics. CRC Press, Boca Raton, FL, revised edition, 2015.

[26] H. Föllmer. An entropy approach to the time reversal of diffusion processes. In M. Métivier and E. Pardoux, editors, *Stochastic Differential Systems Filtering and Control*, volume 69 of *Lecture Notes in Control and Information Sciences*, pages 156–163. Springer, Berlin, Heidelberg, 1985.

[27] Wilfried Genuist, Éric Savin, Filippo Gatti, and Didier Clouteau. Divergence-free diffusion models for incompressible fluid flows. arXiv preprint arXiv:2601.19368, 2026.

[28] Igor V. Girsanov. On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Theory of Probability & Its Applications*, 5(3):285–301, 1960.

[29] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, and Richard Zemel. Learning the Stein discrepancy for training and evaluating energy-based models without sampling. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3732–3747. PMLR, 2020.

[30] Jiequn Han, Arnulf Jentzen, and Weinan E. Solving high-dimensional partial differential equations using deep learning. *Proceedings of the National Academy of Sciences*, 115(34):8505–8510, 2018.

[31] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via Min-SNR weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7441–7451, 2023.

[32] Derek Hansen, Danielle C. Maddix, Shima Alizadeh, Gaurav Gupta, and Michael W. Mahoney. Learning physical models that can respect conservation laws. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 12469–12510. PMLR, 2023.

[33] U. G. Haussmann and E. Pardoux. Time reversal of diffusions. *The Annals of Probability*, 14(4):1188–1205, 1986.

[34] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[35] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[36] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[37] Christian Horvat and Jean-Pascal Pfister. On gauge freedom, conservativity and intrinsic dimensionality estimation in diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

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

[39] Aapo Hyvärinen. Some extensions of score matching. *Computational Statistics & Data Analysis*, 51(5):2499–2512, 2007.

[40] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[41] Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer, New York, second edition, 1991.

[42] Naïl B. Khelifa, Richard E. Turner, and Ramji Venkataramanan. Quantifying error propagation and model collapse in diffusion models. In *Forty-third International Conference on Machine Learning*, 2026.

[43] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[44] Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[45] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[46] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[47] Chieh-Hsin Lai, Yuhta Takida, Naoki Murata, Toshimitsu Uesaka, Yuki Mitsufuji, and Stefano Ermon. FP-diffusion: Improving score-based diffusion models by enforcing the underlying score Fokker-Planck equation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 18365–18398. PMLR, 2023.

[48] Jean-François Le Gall. *Brownian Motion, Martingales, and Stochastic Calculus*, volume 274 of *Graduate Texts in Mathematics*. Springer, 2016.

[49] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[50] Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. In *Predicting Structured Data*. MIT Press, 2006.

[51] Haeil Lee, Hansang Lee, Seoyeon Gye, and Junmo Kim. Beta sampling is all you need: Efficient image generation strategy for diffusion models using stepwise spectral analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025.

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

[53] Rebecca M. Lewis, Oliver Y. Feng, Henry W. J. Reeve, Min Xu, and Richard J. Samworth. Learning the score under shape constraints. arXiv preprint arXiv:2512.14624, 2025.

[54] Xigui Li, Hongwei Zhang, Ruoxi Jiang, Deshu Chen, Chensen Lin, Limei Han, Yuan Qi, Xin Guo, and Yuan Cheng. Project and generate: Divergence-free neural operators for incompressible flows. arXiv preprint arXiv:2603.24500, 2026.

[55] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[56] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[57] Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev GAN. In *International Conference on Learning Representations (ICLR)*, 2018.

[58] Kirill Neklyudov, Rob Brekelmans, Daniel Severo, and Alireza Makhzani. Action matching: Learning stochastic dynamics from samples. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25858–25889. PMLR, 2023.

[59] A. A. Novikov. On an identity for stochastic integrals. *Theory of Probability & Its Applications*, 17(4):717–720, 1973.

[60] Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 26517–26582. PMLR, 2023.

[61] Felix Otto. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.

[62] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023.

[63] Jakiw Pidstrigach. Score-based generative models detect manifolds. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[64] Jack Richter-Powell, Yaron Lipman, and Ricky T. Q. Chen. Neural conservation laws: A divergence-free perspective. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[65] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015.

[66] Tim Salimans and Jonathan Ho. Should EBMs model the energy or the score? In *ICLR 2021 Workshop on Energy-Based Models: Current Perspectives, Challenges, and Opportunities*, 2021.

[67] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser, Cham, 2015.

[68] Saeed Saremi. On approximating ∇f with neural networks. arXiv preprint arXiv:1910.12744, 2019.

[69] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265. PMLR, 2015.

[70] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[71] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[72] Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 574–584. PMLR, 2020.

[73] 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 (ICLR)*, 2021.

[74] Daniel W. Stroock and S. R. Srinivasa Varadhan. *Multidimensional Diffusion Processes*, volume 233 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1979.

[75] Noah Trupin, Rahul Ghosh, and Aadi Jangid. Learning flow distributions via projection-constrained diffusion on manifolds. arXiv preprint arXiv:2602.17773, 2026.

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[77] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 2009.

[78] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

[79] An B. Vuong, Michael T. McCann, Javier E. Santos, and Yen Ting Lin. Are we really learning the score function? Reinterpreting diffusion models through Wasserstein gradient flow matching. *Transactions on Machine Learning Research*, 2025.

[80] Binxu Wang and Cengiz Pehlevan. An analytical theory of spectral bias in the learning dynamics of diffusion models. In *Advances in Neural Information Processing Systems*, 2025.

[81] Andre Wibisono, Yihong Wu, and Kaylee Yingxi Yang. Optimal score estimation via empirical Bayes smoothing. In *Proceedings of the Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 4958–4991. PMLR, 2024.

[82] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747, 2017.

[83] 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*, volume 235 of *Proceedings of Machine Learning Research*, pages 60134–60178. PMLR, 2024.

[84] Pengze Zhang, Hubery Yin, Chen Li, and Xiaohua Xie. Tackling the singularities at the endpoints of time intervals in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6945–6954, 2024.

A Extended Related Work

Score Decomposition. Horvat and Pfister [37] used the Helmholtz-Hodge decomposition of the score field into a gradient and a divergence-free part to show that exact sampling and density estimation only depend on the gradient part, and illustrated this on low-dimensional Gaussian examples. Vuong et al. [79] measured the non-conservativity of trained networks through loop integrals and reinterpreted diffusion training as matching a Wasserstein gradient flow. Chao et al. [13] measured non-conservativity through Jacobian-asymmetry statistics, reported that it can degrade sampling, and proposed a regularizer to address this. Richter-Powell et al. [64] propose divergence-free architectures based on differential forms, while Lai et al. [47] regularize the score network towards the score Fokker–Planck equation satisfied by the true family of scores. Salimans and Ho [66] found that constraining the network to be a gradient costs little or nothing in sample quality, and Saremi [68] showed that unconstrained networks generically cannot represent exact gradients.

Our work differs from existing work in three ways. We (i) decompose the score *error* rather than the score field, (ii) derive quantitative consequences in the approximate regime: an endpoint KL bound (Theorem 3.2) and an impossibility result (Theorem 3.1 (iii)), and (iii) estimate the observable error energy in trained image models.

Minimum-energy velocities and Wasserstein geometry. In the Benamou–Brenier and Otto optimal transport frameworks [5, 61, 1], a curve of marginals is represented by many velocity fields, and the minimum-kinetic-energy one is the gradient field in the tangent space (Appendix D). Action Matching [58] learns this gradient velocity from samples using a scalar neural potential and a variational objective. Our estimator in Section 4.2 uses the same tangent-space geometry, but applies it to the perturbation generated by the score error, $\nabla \cdot (p_s^\star \mathbf{e}_s)$, which involves the unknown true score.

Neural Helmholtz–Hodge decompositions and variational critics. Classical Helmholtz–Hodge methods [8] solve elliptic equations for scalar and vector potentials on grids. They are accurate in low dimension but suffer from the curse of dimensionality [30]. Learned Stein discrepancies [29], Sobolev GANs [57] and the Deep Ritz method [23] optimize neural critics or potentials in variational formulations related to ours. None of these targets the quantity in our bound. The specific difficulty is that the field to be projected, $\mathbf{e}_s$, contains the unknown score $\nabla \log p_s^\star$. Because $\mathbf{e}_s$ enters the dual objective (20) linearly, it can be replaced by the conditionally unbiased DSM residual. The resulting estimator only needs forward-noised data and the frozen score network.

Error weighting and manifold effects. Timestep and frequency reweighting schemes [31, 18, 80, 51] and analyses of score matching near low-dimensional data manifolds [63, 19] also reflect that not all score errors matter equally for sample quality. These works act on the full error across noise levels; the decomposition studied here is complementary to them and can be combined with any time weighting.

B Proofs of Main Results

B.1 Preliminary results

Cutoff functions. The proofs of the two following results (Lemma B.1 and Proposition B.2) rely on cutoff functions, which we define as follows. Fix $\chi \in C_c^\infty(\mathbb{R}^d)$ with $0 \le \chi \le 1$, $\chi(\mathbf{x}) = 1$ for $\|\mathbf{x}\| \le 1$ and $\chi(\mathbf{x}) = 0$ for $\|\mathbf{x}\| \ge 2$. Let $C_\chi := \sup_{\mathbf{y}\in\mathbb{R}^d} \|\nabla\chi(\mathbf{y})\| < \infty$, and set $\chi_R := \chi(\cdot/R)$ for $R \ge 1$, so that $0 \le \chi_R(\mathbf{x}) \le 1$ and $\chi_R(\mathbf{x}) = 1$ if $\|\mathbf{x}\| \le R$. Since χ is constant on the open sets $\{\|\mathbf{y}\| < 1\}$ and $\{\|\mathbf{y}\| > 2\}$, $\nabla\chi(\mathbf{y}) \ne 0$ only if $1 \le \|\mathbf{y}\| \le 2$. As $\nabla\chi_R(\mathbf{x}) = R^{-1}\nabla\chi(\mathbf{x}/R)$, for every $R \ge 1$ and $\mathbf{x} \in \mathbb{R}^d$,

$$\|\nabla\chi_R(\mathbf{x})\| \le \frac{C_\chi}{R}\,\mathbf{1}_{\{R\le\|\mathbf{x}\|\le 2R\}}.$$

In particular $\chi_R \to 1$ pointwise as $R \to \infty$. Moreover, if $R \le \|\mathbf{x}\| \le 2R$, then $1 + \|\mathbf{x}\| \le 1 + 2R \le 3R$ and $1 + \|\mathbf{x}\|^2 \le 1 + 4R^2 \le 5R^2$, so that, for all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\nabla\chi_R(\mathbf{x})\| \le \frac{3C_\chi}{1+\|\mathbf{x}\|}\,\mathbf{1}_{\{R\le\|\mathbf{x}\|\le 2R\}}, \qquad \|\nabla\chi_R(\mathbf{x})\|^2\,(1+\|\mathbf{x}\|^2) \le 5C_\chi^2\,\mathbf{1}_{\{R\le\|\mathbf{x}\|\le 2R\}}. \quad (25)$$

Lemma B.1. *Let $p, g \in C^1(\mathbb{R}^d)$ satisfy*

$$\int_{\mathbb{R}^d} \Big(|g| + (1 + \|\mathbf{x}\|)\, \|\nabla g\| \Big)\, |p|\, \mathrm{d}\mathbf{x} < \infty, \tag{26}$$

and let $\mathbf{F} := p\,\mathbf{v}$, where $\mathbf{v} : \mathbb{R}^d \to \mathbb{R}^d$ is locally Lipschitz and satisfies $\|\mathbf{v}(\mathbf{x})\| \le C_{\mathbf{v}}(1 + \|\mathbf{x}\|)$ for all $\mathbf{x} \in \mathbb{R}^d$. Then $\mathbf{F} \cdot \nabla g \in L^1(\mathbb{R}^d)$,i.e. is Lebesgue integrable. If, moreover, $g\, \nabla \cdot \mathbf{F} \in L^1(\mathbb{R}^d)$, then

$$\int_{\mathbb{R}^d} g\, \nabla \cdot \mathbf{F}\, \mathrm{d}\mathbf{x} = -\int_{\mathbb{R}^d} \mathbf{F} \cdot \nabla g\, \mathrm{d}\mathbf{x}. \tag{27}$$

Proof. Since p is C^1, on each ball $B_\rho := \{\|\mathbf{x}\| \le \rho\}$, p is bounded by $M_p := \sup_{B_\rho} |p|$ and, by the mean value theorem, Lipschitz with constant $L_p := \sup_{B_\rho} \|\nabla p\|$, while $\mathbf{v}$ is bounded by $C_{\mathbf{v}}(1 + \rho)$ and Lipschitz with some constant $L_{\mathbf{v}}$. Hence, for all $\mathbf{x}, \mathbf{y} \in B_\rho$,

$$\begin{aligned} \|\mathbf{F}(\mathbf{x}) - \mathbf{F}(\mathbf{y})\| &\le |p(\mathbf{x})|\, \|\mathbf{v}(\mathbf{x}) - \mathbf{v}(\mathbf{y})\| + \|\mathbf{v}(\mathbf{y})\|\, |p(\mathbf{x}) - p(\mathbf{y})| \\ &\le \big(M_p L_{\mathbf{v}} + C_{\mathbf{v}}(1 + \rho) L_p \big) \|\mathbf{x} - \mathbf{y}\|. \end{aligned} \tag{28}$$

As a consequence, $\mathbf{F} = p\mathbf{v}$ is locally Lipschitz and, by Rademacher's theorem [25, Section 3.1.2], $\nabla \cdot \mathbf{F}$ exists almost everywhere. By the Cauchy–Schwarz inequality and the growth bound $\|\mathbf{v}(\mathbf{x})\| \le C_{\mathbf{v}}(1 + \|\mathbf{x}\|)$, for all $\mathbf{x} \in \mathbb{R}^d$,

$$|\mathbf{F} \cdot \nabla g| = |p|\, |\mathbf{v} \cdot \nabla g| \le |p|\, \|\mathbf{v}\|\, \|\nabla g\| \le C_{\mathbf{v}}\, (1 + \|\mathbf{x}\|)\, \|\nabla g\|\, |p|,$$

and

$$\frac{|g|\, \|\mathbf{F}\|}{1 + \|\mathbf{x}\|} = \frac{|g|\, |p|\, \|\mathbf{v}\|}{1 + \|\mathbf{x}\|} \le \frac{C_{\mathbf{v}}(1 + \|\mathbf{x}\|)\, |g|\, |p|}{1 + \|\mathbf{x}\|} = C_{\mathbf{v}}\, |g|\, |p|,$$

and both right-hand sides are integrable by (26). Hence $\mathbf{F} \cdot \nabla g \in L^1(\mathbb{R}^d)$, which is the first claim, and $|g|\, \|\mathbf{F}\| / (1 + \|\cdot\|) \in L^1(\mathbb{R}^d)$.

Assume now that $g\, \nabla \cdot \mathbf{F} \in L^1(\mathbb{R}^d)$, and fix $R \ge 1$. The vector field $\mathbf{G}_R := \chi_R\, g\, \mathbf{F}$ (for a cutoff function χ_R as described in the paragraph above Lemma B.1) is locally Lipschitz as the product of a locally Lipschitz function and a locally Lipschitz vector field by the same two-term estimate as for $\mathbf{F}$ in (28). Moreover, $\mathbf{G}_R$ vanishes outside $\{\|\mathbf{x}\| \le 2R\}$ because χ_R does. Choose $\psi \in C_c^\infty(\mathbb{R}^d)$ with $\psi = 1$ on $\{\|\mathbf{x}\| \le 2R + 1\}$. By definition of the weak derivative [24], and since $\nabla \psi = 0$ wherever $\mathbf{G}_R \ne 0$,

$$\int_{\mathbb{R}^d} \nabla \cdot \mathbf{G}_R\, \mathrm{d}\mathbf{x} = \int_{\mathbb{R}^d} \psi\, \nabla \cdot \mathbf{G}_R\, \mathrm{d}\mathbf{x} = -\int_{\mathbb{R}^d} \mathbf{G}_R \cdot \nabla \psi\, \mathrm{d}\mathbf{x} = 0.$$

Expanding $\nabla \cdot \mathbf{G}_R$ with the product rule gives, for every $R \ge 1$,

$$0 = \int_{\mathbb{R}^d} \chi_R\, \mathbf{F} \cdot \nabla g\, \mathrm{d}\mathbf{x} + \int_{\mathbb{R}^d} \chi_R\, g\, \nabla \cdot \mathbf{F}\, \mathrm{d}\mathbf{x} + \int_{\mathbb{R}^d} g\, \mathbf{F} \cdot \nabla \chi_R\, \mathrm{d}\mathbf{x}. \tag{29}$$

As $R \to \infty$, the first two integrals converge to $\int \mathbf{F} \cdot \nabla g\, \mathrm{d}\mathbf{x}$ and $\int g\, \nabla \cdot \mathbf{F}\, \mathrm{d}\mathbf{x}$ by dominated convergence, with the integrable dominating functions $|\mathbf{F} \cdot \nabla g|$ and $|g\, \nabla \cdot \mathbf{F}|$. For the third, by (25),

$$\Big| \int_{\mathbb{R}^d} g\, \mathbf{F} \cdot \nabla \chi_R\, \mathrm{d}\mathbf{x} \Big| \le 3 C_\chi \int_{\{\|\mathbf{x}\| \ge R\}} \frac{|g(\mathbf{x})|\, \|\mathbf{F}(\mathbf{x})\|}{1 + \|\mathbf{x}\|}\, \mathrm{d}\mathbf{x} \xrightarrow[R \to \infty]{} 0$$

by dominated convergence. Letting $R \to \infty$ in (29) yields (27). □

We now prove a specialized result that applies Lemma B.1 in our context.

Proposition B.2. *Under A1–A4 and for $r_s := p^\star_s / \hat{p}_s$, the following hold for a.e. $s \in [t_0, T]$:*

- *(i) Both $(p, g) = (p^\star_s, \log r_s)$ and $(p, g) = (\hat{p}_s, r_s)$ satisfy (26). Moreover, for each of these pairs and every $\mathbf{F} = p\mathbf{v}$ as in Lemma B.1 with $|\nabla \cdot \mathbf{F}| = |\partial_s p|$ almost everywhere on $\mathbb{R}^d$, we have $g\, \nabla \cdot \mathbf{F} \in L^1(\mathbb{R}^d)$, so that (27) holds.*
- *(ii) $\nabla \log r_s \in \mathcal{G}_s$, where $\mathcal{G}_s$ is the closure in $L^2(p^\star_s; \mathbb{R}^d)$ of $\{\nabla \varphi : \varphi \in C_c^\infty(\mathbb{R}^d)\}$.*

Proof. Fix $s \in [t_0, T]$ such that the linear-growth bound of A2(iii) holds at s, the conclusions of A4 hold at s, and

$$\int_{\mathbb{R}^d} \Big(|\partial_s p_s^\star| \left(1 + |\log r_s|\right) + |\partial_s \hat{p}_s|\, r_s \Big)\, \mathrm{d}\mathbf{x} < \infty. \tag{30}$$

The integrand in (30) is nonnegative and, by A3, its integral over $[t_0, T] \times \mathbb{R}^d$ is finite, so (30) also holds for a.e. s by Tonelli's theorem. We prove (i) and (ii) for such an s, which is fixed from now on.

Growth of the log-ratio. By A1, $p_s^\star$ and $\hat{p}_s$ are strictly positive and of class C^1, so $\log r_s = \log p_s^\star - \log \hat{p}_s$ and r_s belong to $C^1(\mathbb{R}^d)$. By definition of $\mathbf{e}_s$, $\nabla \log p_s^\star = \mathbf{s}_\theta(\cdot, s) - \mathbf{e}_s$, so A2(iii) gives $\|\nabla \log p_s^\star(\mathbf{x})\| \le C(1 + \|\mathbf{x}\|)$, and A4 gives $\|\nabla \log \hat{p}_s(\mathbf{x})\| \le C_s(1 + \|\mathbf{x}\|)$ for some $C_s < \infty$. Hence $\|\nabla \log r_s(\mathbf{x})\| \le (C + C_s)(1 + \|\mathbf{x}\|)$. Applying the fundamental theorem of calculus to $t \mapsto \log r_s(t\mathbf{x})$ on $[0, 1]$ and using $\|\mathbf{x}\| \le \frac{1}{2}(1 + \|\mathbf{x}\|^2)$,

$$\begin{aligned} |\log r_s(\mathbf{x})| &\le |\log r_s(\mathbf{0})| + \int_0^1 \|\nabla \log r_s(t\mathbf{x})\|\, \|\mathbf{x}\|\, \mathrm{d}t \le |\log r_s(\mathbf{0})| + (C + C_s)\Big(\|\mathbf{x}\| + \tfrac{1}{2}\|\mathbf{x}\|^2\Big) \\ &\le |\log r_s(\mathbf{0})| + (C + C_s)(1 + \|\mathbf{x}\|^2). \end{aligned}$$

Thus, with $c_s := |\log r_s(\mathbf{0})| + C + C_s$, for all $\mathbf{x} \in \mathbb{R}^d$,

$$\|\nabla \log r_s(\mathbf{x})\| \le c_s(1 + \|\mathbf{x}\|), \qquad |\log r_s(\mathbf{x})| \le c_s(1 + \|\mathbf{x}\|^2). \tag{31}$$

Proof of (i). Both pairs consist of C^1 functions. We use (31), $(1 + \|\mathbf{x}\|)^2 \le 2(1 + \|\mathbf{x}\|^2)$, and the finite second moment of $p_s^\star$ from A1. For $(p, g) = (p_s^\star, \log r_s)$,

$$\int_{\mathbb{R}^d} \Big(|\log r_s| + (1 + \|\mathbf{x}\|)\, \|\nabla \log r_s\| \Big)\, p_s^\star\, \mathrm{d}\mathbf{x} \le 3c_s \int_{\mathbb{R}^d} (1 + \|\mathbf{x}\|^2)\, p_s^\star\, \mathrm{d}\mathbf{x} < \infty.$$

For $(p, g) = (\hat{p}_s, r_s)$, we have $\nabla r_s = r_s \nabla \log r_s$ and $\hat{p}_s r_s = p_s^\star$, so

$$\begin{aligned} \int_{\mathbb{R}^d} \Big(r_s + (1 + \|\mathbf{x}\|)\, \|\nabla r_s\| \Big)\, \hat{p}_s\, \mathrm{d}\mathbf{x} &= \int_{\mathbb{R}^d} \Big(1 + (1 + \|\mathbf{x}\|)\, \|\nabla \log r_s\| \Big)\, p_s^\star\, \mathrm{d}\mathbf{x} \\ &\le 1 + 2c_s \int_{\mathbb{R}^d} (1 + \|\mathbf{x}\|^2)\, p_s^\star\, \mathrm{d}\mathbf{x} < \infty. \end{aligned}$$

Hence both pairs satisfy (26). Finally, if $|\nabla \cdot \mathbf{F}| = |\partial_s p|$ almost everywhere, then $|g\, \nabla \cdot \mathbf{F}|$ equals $|\partial_s p_s^\star|\, |\log r_s|$ for the first pair and $|\partial_s \hat{p}_s|\, r_s$ for the second, both integrable by (30). Lemma B.1 then gives (27).

Proof of (ii). We first show that $\nabla\psi \in \mathcal{G}_s$ for $\psi \in C_c^1(\mathbb{R}^d)$. Let $h \in C_c^\infty(\mathbb{R}^d)$ with $h \ge 0$, $h = 0$ outside $\{\|\mathbf{y}\| \le 1\}$ and $\int h = 1$, and set $h_\varepsilon := \varepsilon^{-d} h(\cdot/\varepsilon)$ for $\varepsilon \in (0, 1]$. Then $\psi_\varepsilon := h_\varepsilon * \psi \in C_c^\infty(\mathbb{R}^d)$ and $\nabla\psi_\varepsilon = h_\varepsilon * \nabla\psi$. Since $\nabla\psi$ is continuous with compact support, it is uniformly continuous; let ω be its modulus of continuity. Then, for every $\mathbf{x}$,

$$\|\nabla\psi_\varepsilon(\mathbf{x}) - \nabla\psi(\mathbf{x})\| \le \int_{\mathbb{R}^d} h_\varepsilon(\mathbf{y})\, \|\nabla\psi(\mathbf{x} - \mathbf{y}) - \nabla\psi(\mathbf{x})\|\, \mathrm{d}\mathbf{y} \le \omega(\varepsilon),$$

so, $p_s^\star$ being a probability density, $\|\nabla\psi_\varepsilon - \nabla\psi\|_{L^2(p_s^\star)} \le \omega(\varepsilon) \to 0$ as $\varepsilon \to 0$. Since each $\nabla\psi_\varepsilon$ lies in $\mathcal{G}_s$ and $\mathcal{G}_s$ is closed, $\nabla\psi \in \mathcal{G}_s$.

Let $\varphi_R := \chi_R \log r_s \in C_c^1(\mathbb{R}^d)$, where χ_R is a cutoff function as defined above Lemma B.1. Then

$$\nabla\varphi_R - \nabla \log r_s = (\chi_R - 1)\, \nabla \log r_s + \log r_s\, \nabla\chi_R.$$

Since $|\chi_R - 1| \le \mathbf{1}_{\{\|\mathbf{x}\| > R\}}$ and $\nabla \log r_s = \nabla \log p_s^\star - \nabla \log \hat{p}_s \in L^2(p_s^\star; \mathbb{R}^d)$ by A4, dominated convergence gives $\|(\chi_R - 1)\nabla \log r_s\|_{L^2(p_s^\star)} \to 0$. For the second term, (31) and (25) give

$$\begin{aligned} \int_{\mathbb{R}^d} |\log r_s|^2\, \|\nabla\chi_R\|^2\, p_s^\star\, \mathrm{d}\mathbf{x} &\le c_s^2 \int_{\mathbb{R}^d} (1 + \|\mathbf{x}\|^2)\, \left[\|\nabla\chi_R(\mathbf{x})\|^2 (1 + \|\mathbf{x}\|^2)\right] p_s^\star(\mathbf{x})\, \mathrm{d}\mathbf{x} \\ &\le 5C_\chi^2 c_s^2 \int_{\{\|\mathbf{x}\| \ge R\}} (1 + \|\mathbf{x}\|^2)\, p_s^\star(\mathbf{x})\, \mathrm{d}\mathbf{x}, \end{aligned}$$

which tends to 0 as $R \to \infty$ because $p_s^\star$ has a finite second moment by A1. By the triangle inequality, $\|\nabla\varphi_R - \nabla \log r_s\|_{L^2(p_s^\star)} \to 0$. Since $\varphi_R \in C_c^1(\mathbb{R}^d)$ by the above, $\nabla\varphi_R \in \mathcal{G}_s$. By closedness of $\mathcal{G}_s$, this implies $\nabla \log r_s \in \mathcal{G}_s$. □

B.2 Proof of Theorem 3.1

Proof. Proof of (i): Recall that the score estimation error field $\mathbf{e}_s$ is decomposed as:

$$\mathbf{e}_s = \mathbf{e}_{\text{obs},s} + \mathbf{e}_{\text{inv},s}, \qquad \mathbf{e}_{\text{obs},s} := \Pi_{\mathcal{G}_s}\mathbf{e}_s, \quad \mathbf{e}_{\text{inv},s} := \Pi_{\mathcal{G}_s^\perp}\mathbf{e}_s, \tag{32}$$

where $\mathcal{G}_s^\perp$ is the orthogonal complement in $L^2(p_s^\star;\mathbb{R}^d)$. Also recall from (9) that the marginals of the true backward SDE $(p_s^\star)_{s\in[t_0,T]}$ satisfy,

$$\partial_s p_s^\star = -\nabla\cdot\Big(\big[\boldsymbol{f}_s - \sigma_s^2\mathbf{s}_\theta(\cdot,s)\big]p_s^\star\Big) - \frac{\sigma_s^2}{2}\Delta p_s^\star - \sigma_s^2\nabla\cdot\big(p_s^\star\,\mathbf{e}_s\big). \tag{33}$$

Applying the decomposition (32), (33) can be written as

$$\partial_s p_s^\star = -\nabla\cdot\Big(\big[\boldsymbol{f}_s - \sigma_s^2\mathbf{s}_\theta(\cdot,s)\big]p_s^\star\Big) - \frac{\sigma_s^2}{2}\Delta p_s^\star - \sigma_s^2\nabla\cdot\big(p_s^\star\,\mathbf{e}_{\text{obs},s}\big), \tag{34}$$

since, by definition of $\mathcal{G}_s$ (see (10)), $\nabla\cdot(p_s^\star\,\mathbf{e}_{\text{inv},s}) = 0$. Therefore $(p_s^\star)_{s\in[t_0,T]}$ solves (13) with $\mathbf{v}_s = \mathbf{e}_{\text{obs},s}$ and terminal condition $p_T^\star$. This proves (i).

Proof of (ii): Assume $\mathbf{e}_{\text{obs}} \equiv 0$. Then (33), or equivalently (34), reduces to the following PDE, backward in time s, from T to t_0:

$$\partial_s p_s^\star = -\nabla\cdot\Big(\big[\boldsymbol{f}_s - \sigma_s^2\mathbf{s}_\theta(\cdot,s)\big]p_s^\star\Big) - \frac{\sigma_s^2}{2}\Delta p_s^\star. \tag{35}$$

Writing $\nabla_{\mathbf{x}}\log p_s^\star = \mathbf{s}_\theta(\cdot,s) - \mathbf{e}_s$ in (8), the $\mathbf{e}_s$ terms cancel, and $(\hat{p}_s)_s$ solves the same equation (35) with the same terminal condition $p_T^\star$. Since the drift $\boldsymbol{f}_s - \sigma_s^2\mathbf{s}_\theta(\cdot,s)$ is locally Lipschitz with linear growth by A2, this equation has a unique solution [9].

Thus, $p_s^\star = \hat{p}_s$ for all $s\in[t_0,T]$, meaning that any divergence between these two marginals is 0. Moreover, if $\mathbf{e}_{\text{obs}} \equiv 0$, then $\mathbf{e} = \mathbf{e}_{\text{inv}}$, hence

$$\int_{t_0}^T \|\mathbf{e}_s\|^2_{L^2(p_s^\star)}\,\mathrm{d}s = \int_{t_0}^T \|\mathbf{e}_{\text{inv},s}\|^2_{L^2(p_s^\star)}\,\mathrm{d}s.$$

If $\mathbf{e}_{\text{inv}} \not\equiv 0$, the right-hand side is strictly positive. This proves (ii).

Proof of (iii): no lower bound from the full L^2 score error.

Assume by contradiction that there exists a strictly increasing function $F:[0,\infty)\to\mathbb{R}$ with $F(0)=0$ such that, for every score error field $\mathbf{e}$,

$$\mathrm{Div}(p_{t_0}^\star\|\hat{p}_{t_0}) \;\geq\; F\left(\int_{t_0}^T \|\mathbf{e}_s\|^2_{L^2(p_s^\star)}\,\mathrm{d}s\right). \tag{36}$$

We first exhibit an admissible purely invisible error. Let $J\in\mathbb{R}^{d\times d}$ be a nonzero antisymmetric matrix (this requires $d\geq 2$), let $\psi\in C_c^\infty(\mathbb{R}^d)$ be non-constant, and for $\lambda>0$ set

$$\mathbf{u}_s^\lambda := \lambda\,\frac{J\nabla\psi}{p_s^\star}.$$

Then $\nabla\cdot(p_s^\star\mathbf{u}_s^\lambda) = \lambda\,\nabla\cdot(J\nabla\psi) = \lambda\sum_{i,j}J_{ij}\partial_i\partial_j\psi = 0$ by antisymmetry, so $\mathbf{u}_s^\lambda\in\mathcal{G}_s^\perp$. Moreover $\mathbf{u}_s^\lambda$ is bounded, compactly supported, locally Lipschitz and jointly measurable, so the score $\mathbf{s}_\theta(\cdot,s) = \nabla\log p_s^\star + \mathbf{u}_s^\lambda$ satisfies A1–A4 (this uses that $\nabla\log p_s^\star$ is locally Lipschitz with linear growth, which holds as soon as one score network satisfies A2, since $\nabla\log p_s^\star = \mathbf{s}_\theta - \mathbf{e}_s$), with $\hat{p}_s = p_s^\star$ by part (ii). Define $\mathbf{e}_s := \mathbf{u}_s^\lambda$ for some $\lambda>0$. Since $\mathbf{u}_s^\lambda\in\mathcal{G}_s^\perp$, we have for almost every $s\in[t_0,T]$ that $\Pi_{\mathcal{G}_s}\mathbf{e}_s = 0$. Therefore, by part (ii), the corresponding learned marginal curve coincides with the true one, and in particular $\hat{p}_{t_0} = p_{t_0}^\star$. Hence, using that $\mathrm{Div}(\mu\|\nu) = 0$ if and only if $\mu=\nu$, $\mathrm{Div}(p_{t_0}^\star\|\hat{p}_{t_0}) = 0$. On the other hand,

$$\int_{t_0}^T \|\mathbf{e}_s\|^2_{L^2(p_s^\star)}\,\mathrm{d}s = \int_{t_0}^T \|\mathbf{u}_s^\lambda\|^2_{L^2(p_s^\star)}\mathrm{d}s = \lambda^2\int_{t_0}^T \|\mathbf{u}_s^1\|^2_{L^2(p_s^\star)}\mathrm{d}s \underset{\lambda\to\infty}{\longrightarrow} +\infty$$

Since F is strictly increasing such that $F(0) = 0$, one can pick $\lambda > 0$ such that $F(\int_{t_0}^{T} \|\mathbf{u}_s^\lambda\|^2_{L^2(p_s^\star)} \mathrm{d}s) > 0$. But, applying (36) to the error field $\mathbf{e}$ yields

$$0 = \mathrm{Div}(p_{t_0}^\star \| \hat{p}_{t_0}) \;\geq\; F(\int_{t_0}^{T} \|\mathbf{u}_s^\lambda\|^2_{L^2(p_s^\star)} \mathrm{d}s) > 0.$$

This contradicts the previous inequality and proves that no such lower bound of the form (36) can hold uniformly over all score error fields $\mathbf{e}$. □

B.3 Proof of Theorem 3.2

We first state a useful lemma, which is similar to [14, Lemma 6], but we give a proof for completeness.

Lemma B.3. *Assume A1–A4. For $s \in [t_0, T]$ let $r_s := p_s^\star / \hat{p}_s$ denote the time-s ratio of marginals. Then, for a.e. $s \in [t_0, T]$,*

$$-\frac{\mathrm{d}}{\mathrm{d}s}\mathrm{KL}(p_s^\star \,\|\, \hat{p}_s) = -\frac{\sigma_s^2}{2}\,\mathcal{I}(p_s^\star \,\|\, \hat{p}_s) - \sigma_s^2 \int_{\mathbb{R}^d} (\mathbf{e}_s \cdot \nabla \log r_s)\, p_s^\star \,\mathrm{d}\mathbf{x}, \tag{37}$$

where $\mathcal{I}(p_s^\star \| \hat{p}_s) := \int_{\mathbb{R}^d} \|\nabla \log r_s\|^2 \, p_s^\star \,\mathrm{d}\mathbf{x}$ is the relative Fisher information of $p_s^\star$ with respect to $\hat{p}_s$. The left-hand side is the rate of change of the divergence along the reverse dynamics, i.e. as s decreases from T to t_0.

Proof. By A3, $s \mapsto \mathrm{KL}(p_s^\star \| \hat{p}_s)$ is absolutely continuous and may be differentiated under the integral sign. In the proof of this lemma, we reparametrize reverse time by $\tau = T - s$. Thus $\tau \in [0, T - t_0]$, and all quantities below are understood as

$$p_\tau^\star := p_{T-\tau}^\star, \qquad \hat{p}_\tau := \hat{p}_{T-\tau}, \qquad \mathbf{e}_\tau := \mathbf{e}_{T-\tau}, \qquad \boldsymbol{f}_\tau := \boldsymbol{f}_{T-\tau}, \qquad \sigma_\tau := \sigma_{T-\tau}.$$

We denote $r_\tau := \frac{p_\tau^\star}{\hat{p}_\tau}$ and the common part of the reverse drift as

$$\mathbf{b}_\tau(\mathbf{x}) := -\boldsymbol{f}_\tau(\mathbf{x}) + \sigma_\tau^2 \nabla \log p_\tau^\star(\mathbf{x}).$$

Then the true and learned backward Fokker-Planck equations read:

$$\partial_\tau p_\tau^\star = -\nabla \cdot (\mathbf{b}_\tau p_\tau^\star) + \frac{\sigma_\tau^2}{2} \Delta p_\tau^\star, \tag{38}$$

$$\partial_\tau \hat{p}_\tau = -\nabla \cdot (\mathbf{b}_\tau \hat{p}_\tau) + \frac{\sigma_\tau^2}{2} \Delta \hat{p}_\tau - \sigma_\tau^2 \nabla \cdot (\hat{p}_\tau \mathbf{e}_\tau). \tag{39}$$

The relative entropy between $\hat{p}_\tau$ and $p_\tau^\star$ (which are assumed absolutely continuous relative to each other) is defined as:

$$\mathrm{KL}(p_\tau^\star \,\|\, \hat{p}_\tau) = \int_{\mathbb{R}^d} p_\tau^\star\,(\mathbf{x}) \log \frac{p_\tau^\star(\mathbf{x})}{\hat{p}_\tau(\mathbf{x})} \,\mathrm{d}\mathbf{x} = \int_{\mathbb{R}^d} p_\tau^\star \,\log r_\tau \,\mathrm{d}\mathbf{x}.$$

Step 1: Differentiation of the relative entropy.

Using the product rule,

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mathrm{KL}(p_\tau^\star \,\|\, \hat{p}_\tau) = \int_{\mathbb{R}^d} \partial_\tau p_\tau^\star \,\log r_\tau \,\mathrm{d}\mathbf{x} + \int_{\mathbb{R}^d} p_\tau^\star \,\partial_\tau(\log r_\tau) \,\mathrm{d}\mathbf{x}.$$

Since $\partial_\tau(\log r_\tau) = \partial_\tau p_\tau^\star / p_\tau^\star - \partial_\tau \hat{p}_\tau / \hat{p}_\tau$, we obtain

$$\int_{\mathbb{R}^d} p_\tau^\star \,\partial_\tau(\log r_\tau) \,\mathrm{d}\mathbf{x} = \int_{\mathbb{R}^d} \partial_\tau p_\tau^\star \,\mathrm{d}\mathbf{x} - \int_{\mathbb{R}^d} p_\tau^\star \,\frac{\partial_\tau \hat{p}_\tau}{\hat{p}_\tau} \,\mathrm{d}\mathbf{x}.$$

Because $p_\tau^\star$ is a probability density for every τ, $\int \partial_\tau p_\tau^\star \,\mathrm{d}\mathbf{x} = 0$, hence

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}\mathrm{KL}(p_\tau^\star \,\|\, \hat{p}_\tau) &= \int_{\mathbb{R}^d} \partial_\tau p_\tau^\star \,\log r_\tau \,\mathrm{d}\mathbf{x} - \int_{\mathbb{R}^d} p_\tau^\star \,\frac{\partial_\tau \hat{p}_\tau}{\hat{p}_\tau} \,\mathrm{d}\mathbf{x} \\ &= \int_{\mathbb{R}^d} \partial_\tau p_\tau^\star \,\log r_\tau \,\mathrm{d}\mathbf{x} - \int_{\mathbb{R}^d} r_\tau \,\partial_\tau \hat{p}_\tau \,\mathrm{d}\mathbf{x}. \end{aligned} \tag{40}$$

Step 2: Justifying integration by parts.

Let $\mathbf{F}_\tau^\star := p_\tau^\star\big(\mathbf{b}_\tau - \frac{\sigma_\tau^2}{2}\nabla \log p_\tau^\star\big)$ and $\hat{\mathbf{F}}_\tau := \hat{p}_\tau\big(\mathbf{b}_\tau - \frac{\sigma_\tau^2}{2}\nabla \log \hat{p}_\tau + \sigma_\tau^2 \mathbf{e}_\tau\big)$, so that (38)–(39) read $\partial_\tau p_\tau^\star = -\nabla\cdot \mathbf{F}_\tau^\star$ and $\partial_\tau \hat{p}_\tau = -\nabla\cdot \hat{\mathbf{F}}_\tau$. Both are of the form $p\,\mathbf{v}$ with $\mathbf{v}$ locally Lipschitz of linear growth, by A2, A3 and A4. Applying Proposition B.2(i) with $(p,g) = (p_\tau^\star, \log r_\tau)$ and $(p,g) = (\hat{p}_\tau, r_\tau)$ in (40) gives, for a.e. τ,

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mathrm{KL}(p_\tau^\star \,\|\, \hat{p}_\tau) = \underbrace{\int \mathbf{b}_\tau p_\tau^\star \cdot \nabla \log r_\tau - \int \mathbf{b}_\tau \hat{p}_\tau \cdot \nabla r_\tau + \frac{\sigma_\tau^2}{2}\Big(-\int \nabla p_\tau^\star \cdot \nabla \log r_\tau + \int \nabla \hat{p}_\tau \cdot \nabla r_\tau\Big)}_{=:A_\tau}$$

$$\underbrace{-\sigma_\tau^2 \int \hat{p}_\tau\, \mathbf{e}_\tau \cdot \nabla r_\tau}_{=:B_\tau}. \tag{41}$$

Splitting is legitimate: each of the five integrands is of the form $(p\,\mathbf{v})\cdot\nabla g$ with $\mathbf{v} \in \{\mathbf{b}_\tau, \nabla \log p, \mathbf{e}_\tau\}$, hence integrable by Lemma B.1. We now compute A_τ and B_τ separately.

Step 3: Contribution of A_τ. Since $\nabla r_\tau = r_\tau \nabla \log r_\tau$ and $r_\tau \hat{p}_\tau = p_\tau^\star$, we have $\mathbf{b}_\tau \hat{p}_\tau \cdot \nabla r_\tau = \mathbf{b}_\tau p_\tau^\star \cdot \nabla \log r_\tau$, so the two transport integrals in A_τ cancel exactly, and

$$A_\tau = \frac{\sigma_\tau^2}{2}\left(-\int \nabla p_\tau^\star \cdot \nabla \log r_\tau \,\mathrm{d}\mathbf{x} + \int \nabla r_\tau \cdot \nabla \hat{p}_\tau \,\mathrm{d}\mathbf{x}\right).$$

Using $p_\tau^\star = r_\tau \hat{p}_\tau$, we have $\nabla p_\tau^\star = r_\tau \nabla \hat{p}_\tau + \hat{p}_\tau \nabla r_\tau$, and

$$\begin{aligned}
-\nabla p_\tau^\star \cdot \nabla \log r_\tau &= -r_\tau \nabla \hat{p}_\tau \cdot \nabla \log r_\tau - \hat{p}_\tau \nabla r_\tau \cdot \nabla \log r_\tau \\
&= -\nabla \hat{p}_\tau \cdot \nabla r_\tau - \hat{p}_\tau \nabla r_\tau \cdot \nabla \log r_\tau \\
&= -\nabla \hat{p}_\tau \cdot \nabla r_\tau - p_\tau^\star \nabla \log r_\tau \cdot \nabla \log r_\tau \\
&= -\nabla \hat{p}_\tau \cdot \nabla r_\tau - p_\tau^\star \|\nabla \log r_\tau\|^2,
\end{aligned}$$

where we used $\nabla \log r_\tau = \nabla r_\tau / r_\tau$ in the second step and $\hat{p}_\tau \nabla r_\tau = p_\tau^\star \nabla \log r_\tau$ in the third. The first term cancels the second integral in A_τ, leaving

$$A_\tau = -\frac{\sigma_\tau^2}{2}\int_{\mathbb{R}^d} p_\tau^\star(\mathbf{x})\, \|\nabla \log r_\tau(\mathbf{x})\|^2 \,\mathrm{d}\mathbf{x} = -\frac{\sigma_\tau^2}{2}\,\mathcal{I}(p_\tau^\star \,\|\, \hat{p}_\tau),$$

where $\mathcal{I}(p_\tau^\star \| \hat{p}_\tau) := \int_{\mathbb{R}^d} \|\nabla \log r_\tau\|^2 \, p_\tau^\star \,\mathrm{d}\mathbf{x}$ is the relative Fisher information of $p_\tau^\star$ with respect to $\hat{p}_\tau$.

Step 4: Contribution of B_τ. Using $\hat{p}_\tau \nabla r_\tau = p_\tau^\star \nabla \log r_\tau$ in the definition of B_τ in (41),

$$B_\tau = -\sigma_\tau^2 \int_{\mathbb{R}^d} p_\tau^\star(\mathbf{x})\, \mathbf{e}_\tau(\mathbf{x}) \cdot \nabla \log r_\tau(\mathbf{x}) \,\mathrm{d}\mathbf{x}.$$

Combining A_τ and B_τ in (41),

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\mathrm{KL}(p_\tau^\star \,\|\, \hat{p}_\tau) = -\frac{\sigma_\tau^2}{2}\,\mathcal{I}(p_\tau^\star \,\|\, \hat{p}_\tau) - \sigma_\tau^2 \int_{\mathbb{R}^d} (\mathbf{e}_\tau \cdot \nabla \log r_\tau)\, p_\tau^\star \,\mathrm{d}\mathbf{x}. \tag{42}$$

Equation (42) is written in forward time for $\tau \in [0, T - t_0]$, with $\mathrm{d}\tau > 0$. Since $s = T - \tau$, we have $\frac{\mathrm{d}}{\mathrm{d}\tau} = -\frac{\mathrm{d}}{\mathrm{d}s}$, which yields the statement of the lemma. □

Finally, here is the proof of Theorem 3.2.

Proof. By Lemma B.3, for a.e. $s \in [t_0, T]$ in the reverse-time diffusion,

$$-\frac{\mathrm{d}}{\mathrm{d}s}\mathrm{KL}(p_s^\star \,\|\, \hat{p}_s) = -\frac{\sigma_s^2}{2}\,\mathcal{I}(p_s^\star \,\|\, \hat{p}_s) - \sigma_s^2 \int_{\mathbb{R}^d} (\mathbf{e}_s \cdot \nabla \log r_s)\, p_s^\star \,\mathrm{d}\mathbf{x}, \tag{43}$$

where $r_s = p_s^\star / \hat{p}_s$. By Proposition B.2(ii), $\nabla \log r_s$ belongs to the gradient subspace $\mathcal{G}_s$ (defined in (10)). Using the Helmholtz-Hodge decomposition of the errors in (11):

$$\mathbf{e}_s = \Pi_{\mathcal{G}_s}\mathbf{e}_s + \Pi_{\mathcal{G}_s^\perp}\mathbf{e}_s \;\in L^2(p_s^\star; \mathbb{R}^d),$$

we obtain, since $\Pi_{\mathcal{G}_s^\perp}\mathbf{e}_s \in \mathcal{G}_s^\perp$ and $\nabla \log r_s \in \mathcal{G}_s$, that $\int_{\mathbb{R}^d}(\Pi_{\mathcal{G}_s^\perp}\mathbf{e}_s \cdot \nabla \log r_s)\, p_s^\star \mathrm{d}\mathbf{x} = 0$. Thus,

$$\int_{\mathbb{R}^d}(\mathbf{e}_s \cdot \nabla \log r_s)\, p_s^\star \,\mathrm{d}\mathbf{x} = \int_{\mathbb{R}^d}(\Pi_{\mathcal{G}_s}\mathbf{e}_s \cdot \nabla \log r_s)\, p_s^\star \,\mathrm{d}\mathbf{x}.$$

Then, by Cauchy–Schwarz,

$$-\int_{\mathbb{R}^d}(\mathbf{e}_s \cdot \nabla \log r_s)\, p_s^\star \,\mathrm{d}\mathbf{x} = -\int_{\mathbb{R}^d}(\Pi_{\mathcal{G}_s}\mathbf{e}_s \cdot \nabla \log r_s)\, p_s^\star \,\mathrm{d}\mathbf{x} \le \|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|_{L^2(p_s^\star)}\sqrt{\mathcal{I}(p_s^\star \,\|\, \hat{p}_s)}.$$

Substituting into (43), this yields

$$-\frac{\mathrm{d}}{\mathrm{d}s}\mathrm{KL}(p_s^\star \,\|\, \hat{p}_s) \le -\frac{\sigma_s^2}{2}\,\mathcal{I}(p_s^\star \,\|\, \hat{p}_s) + \sigma_s^2\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|_{L^2(p_s^\star)}\sqrt{\mathcal{I}(p_s^\star \,\|\, \hat{p}_s)}.$$

By Young's inequality $ab \le \frac{1}{2}a^2 + \frac{1}{2}b^2$ applied with $a = \sqrt{\mathcal{I}(p_s^\star \,\|\, \hat{p}_s)}$ and $b = \|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|_{L^2(p_s^\star)}$, one gets,

$$\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|_{L^2(p_s^\star)}\sqrt{\mathcal{I}(p_s^\star \,\|\, \hat{p}_s)} \le \frac{1}{2}\mathcal{I}(p_s^\star \,\|\, \hat{p}_s) + \frac{1}{2}\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p_s^\star)}.$$

Hence

$$-\frac{\mathrm{d}}{\mathrm{d}s}\mathrm{KL}(p_s^\star \,\|\, \hat{p}_s) \le \frac{\sigma_s^2}{2}\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p_s^\star)}. \tag{44}$$

Since $p_T^\star = \hat{p}_T$, we have $\mathrm{KL}(p_T^\star \| \hat{p}_T) = 0$, so that $\mathrm{KL}(p_{t_0}^\star \| \hat{p}_{t_0}) = -\int_{t_0}^T \frac{\mathrm{d}}{\mathrm{d}s}\mathrm{KL}(p_s^\star \| \hat{p}_s)\,\mathrm{d}s$ (absolute continuity by A3). Integrating (44) over $[t_0, T]$ gives

$$\mathrm{KL}(p_{t_0}^\star \| \hat{p}_{t_0}) \le \frac{1}{2}\int_{t_0}^T \sigma_s^2\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p_s^\star)}\,\mathrm{d}s,$$

which is exactly (15). □

C Connection with Girsanov's Theory

In this section, we formally develop the link between our Helmholtz-Hodge approach and its implications for the standard Girsanov-based approach to bound $\mathrm{KL}(p_{t_0}^\star \| \hat{p}_{t_0})$. We prove that the upper bound in Theorem 3.2 can also be established (with an additional assumption) using Girsanov's Theorem [11, 28, 41] by applying the Helmholtz-Hodge decomposition to the stochastic integral of errors and its quadratic variation.

C.1 Technical Background

We work on a complete filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[t_0,T]}, \mathbb{P})$ equipped with a $\mathbb{P}$-Brownian motion $(\mathbf{B}_t)_{t\in[t_0,T]}$. We denote by $(\bar{\mathbf{B}}_t)_{t\in[t_0,T]}$ a Brownian motion under the reverse filtration and recall the true reverse SDE,

$$\mathrm{d}\mathbf{Y}_s = [\boldsymbol{f}_s(\mathbf{Y}_s) - \sigma_s^2\,\nabla_{\mathbf{x}}\log p_s^\star(\mathbf{Y}_s)]\,\mathrm{d}s + \sigma_s\,\mathrm{d}\bar{\mathbf{B}}_s, \qquad \mathbf{Y}_T \sim p_T^\star, \tag{45}$$

and the learned reverse SDE,

$$\mathrm{d}\hat{\mathbf{Y}}_s = [\boldsymbol{f}_s(\hat{\mathbf{Y}}_s) - \sigma_s^2\,\nabla_{\mathbf{x}}\log p_s^\star(\hat{\mathbf{Y}}_s) - \sigma_s^2\mathbf{e}_s(\hat{\mathbf{Y}}_s)]\,\mathrm{d}s + \sigma_s\,\mathrm{d}\bar{\mathbf{B}}_s, \qquad \hat{\mathbf{Y}}_T \sim p_T^\star, \tag{46}$$

where $\mathbf{e}_s(\mathbf{x}) := \mathbf{s}_\theta(\mathbf{x}, s) - \nabla_{\mathbf{x}}\log p_s^\star(\mathbf{x})$. Under mild (and assumed) regularity assumptions both SDEs are well-posed in law, and we denote $(\mathbf{Y})_s$ (resp. $(\hat{\mathbf{Y}}_s)_s$) the weak solution [41, 48] of (45) (resp. of (46)). In the original probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in[t_0,T]}, \mathbb{P})$, these are two distinct processes (functions of $\Omega \to C([t_0, T], \mathbb{R}^d)$) under a same Brownian motion $\bar{\mathbf{B}}$.

Based on these two weak solutions, we define $\mathbb{P}^\star = \mathrm{Law}((\mathbf{Y}_s)_s)$ and $\hat{\mathbb{P}} = \mathrm{Law}((\hat{\mathbf{Y}}_s)_s)$ to be the path laws of (45) and (46). The path-space distributions $\mathbb{P}^\star$ and $\hat{\mathbb{P}}$ live in the canonical space of paths $C([t_0, T], \mathbb{R}^d)$, and are two different measures on this space. We equip this space with the coordinate process $(\mathbf{X}_s = \omega(s))_{s\in[t_0,T]}$, which is the only process we define on $C([t_0, T], \mathbb{R}^d)$ and

which corresponds to the value of the path at time s. On this new path-space, there is one single process $(\mathbf{X}_s)_{s\in[t_0,T]}$ and two distributions such that, under $\mathbb{P}^\star$, $(\mathbf{X}_s)_{s\in[t_0,T]}$ satisfies,

$$\mathrm{d}\mathbf{X}_s = [\boldsymbol{f}_s(\mathbf{X}_s) - \sigma_s^2\,\nabla_{\mathbf{x}}\log p_s^\star(\mathbf{X}_s)]\,\mathrm{d}s + \sigma_s\,\mathrm{d}\bar{\mathbf{B}}_s^{\mathbb{P}^\star}, \qquad \mathbf{X}_T \sim p_T^\star, \tag{47}$$

and under $\hat{\mathbb{P}}$, $(\mathbf{X}_s)_{s\in[t_0,T]}$ satisfies,

$$\mathrm{d}\mathbf{X}_s = [\boldsymbol{f}_s(\mathbf{X}_s) - \sigma_s^2\,\nabla_{\mathbf{x}}\log p_s^\star(\mathbf{X}_s) - \sigma_s^2\mathbf{e}_s(\mathbf{X}_s)]\,\mathrm{d}s + \sigma_s\,\mathrm{d}\bar{\mathbf{B}}_s^{\hat{\mathbb{P}}}, \qquad \mathbf{X}_T \sim p_T^\star, \tag{48}$$

where $\bar{\mathbf{B}}_s^{\mathbb{P}^\star}$ (resp. $\bar{\mathbf{B}}_s^{\hat{\mathbb{P}}}$) is a Brownian motion under $\mathbb{P}^\star$ (resp. under $\hat{\mathbb{P}}$). As discussed below, under suitable assumptions we have $\mathbb{P}^\star \ll \hat{\mathbb{P}}$, with the Radon-Nikodym derivative $\frac{\mathrm{d}\mathbb{P}^\star}{\mathrm{d}\hat{\mathbb{P}}}$ given by Girsanov's theorem (see (51)). Furthermore, the processes $\bar{\mathbf{B}}_s^{\mathbb{P}^\star}$ and $\bar{\mathbf{B}}_s^{\hat{\mathbb{P}}}$ are related as

$$\bar{\mathbf{B}}_t^{\mathbb{P}^\star} = \bar{\mathbf{B}}_t^{\hat{\mathbb{P}}} - \int_T^t \sigma_s\,\mathbf{e}_s\,\mathrm{d}s. \tag{49}$$

Girsanov density. Beyond Assumptions A1–A4, Girsanov-based arguments further require the following assumption:

(A5) Define the stochastic integral and its quadratic variation (both random variables),

$$M_s := \int_{t_0}^s \sigma_u\,\mathbf{e}_u(\mathbf{X}_u)\cdot\mathrm{d}\bar{\mathbf{B}}_u^{\hat{\mathbb{P}}}, \qquad \langle M\rangle_s := \int_{t_0}^s \sigma_u^2\,\|\mathbf{e}_u(\mathbf{X}_u)\|_2^2\,\mathrm{d}u, \tag{50}$$

and the associated Doléans–Dade exponential

$$Z_s := \exp\!\big(M_s - \tfrac{1}{2}\langle M\rangle_s\big)\,.$$

We assume that $(Z_s)_{s\in[t_0,T]}$ is a true $\hat{\mathbb{P}}$-martingale on $[t_0,T]$.

Assumption A5 ensures that the Girsanov transformation yields a valid probability measure (i.e., $\mathbb{E}_{\hat{\mathbb{P}}}[Z_T]=1$). While Z_s is guaranteed to be a local martingale by construction, different sufficient conditions can be found in the literature to guarantee the full martingale status (Novikov's condition [59], Beneš's condition [6]).

Under Assumptions A1–A5, Girsanov's theorem [11, 28, 41] implies that the Radon–Nikodym derivative of $\mathbb{P}^\star$ with respect to $\hat{\mathbb{P}}$ is given by the Doléans–Dade exponential [20] Z_T,

$$\frac{\mathrm{d}\mathbb{P}^\star}{\mathrm{d}\hat{\mathbb{P}}} = Z_T = \exp\Big(\int_{t_0}^T \sigma_s\mathbf{e}_s\cdot\mathrm{d}\bar{\mathbf{B}}_s^{\hat{\mathbb{P}}} - \tfrac{1}{2}\int_{t_0}^T \sigma_s^2\|\mathbf{e}_s\|_2^2\,\mathrm{d}s\Big). \tag{51}$$

For a similar application of Girsanov's theorem to analyze the sampling quality of diffusion models, we refer the reader to [16].

C.2 Standard Girsanov-Based Upper Bound

To capture errors along the entire true (45) and learned (46) reverse trajectories, we define

$$\varepsilon_\star^2 := \int_{t_0}^T \sigma_s^2\|\mathbf{e}_s\|_{L^2(p_s^\star)}^2\,\mathrm{d}s, \qquad \hat{\varepsilon}^2 := \int_{t_0}^T \sigma_s^2\|\mathbf{e}_s\|_{L^2(\hat{p}_s)}^2\,\mathrm{d}s, \tag{52}$$

where $p_s^\star$ and $\hat{p}_s$ are the time-s marginals of $\mathbb{P}^\star$ and $\hat{\mathbb{P}}$, respectively. We note that Assumptions A1–A2 imply that,

$$\varepsilon_\star^2 = \mathbb{E}_{\mathbb{P}^\star}\left[\int_{t_0}^T \sigma_s^2\|\mathbf{e}_s(\mathbf{X}_s)\|_2^2\mathrm{d}s\right] < \infty.$$

We now formally state the standard upper bound, which Theorem 3.2 improves upon.

Proposition C.1 (Standard Girsanov upper bound)**.** *Assume A1–A5, then* $\mathbb{P}^\star \ll \hat{\mathbb{P}}$,

$$\mathrm{KL}(p_{t_0}^\star\|\hat{p}_{t_0}) \le \mathrm{KL}(\mathbb{P}^\star\,\|\,\hat{\mathbb{P}}) = \tfrac{1}{2}\,\varepsilon_\star^2. \tag{53}$$

We provide a short proof for completeness, although the result is standard [16, 7, 14, 42]. This proof also clearly indicates the critical points where modifications are needed to obtain our new bound (Theorem 3.2).

Proof. By Assumption A5, $(Z_s)_{s\in[t_0,T]} = (\exp\big(M_s - \frac{1}{2}\langle M\rangle_s\big))_{s\in[t_0,T]}$ is a $\hat{\mathbb{P}}$-martingale, so in particular $\mathbb{E}_{\hat{\mathbb{P}}}[Z_T] = 1$. Therefore, by Girsanov's Theorem, the Radon–Nikodym derivative of $\mathbb{P}^\star$ with respect to $\hat{\mathbb{P}}$ is given by (51).

Moreover, recall from Section C.1 that $\mathbb{P}^\star$ and $\hat{\mathbb{P}}$ induce two Brownian motions, $(\bar{\mathbf{B}}_s^{\mathbb{P}^\star})_s$ and $(\bar{\mathbf{B}}_s^{\hat{\mathbb{P}}})_s$, respectively, which are linked by the identity (49), which we can write as

$$\mathrm{d}\bar{\mathbf{B}}_s^{\hat{\mathbb{P}}} = \mathrm{d}\bar{\mathbf{B}}_s^{\mathbb{P}^\star} + \sigma_s \mathbf{e}_s \,\mathrm{d}s. \tag{54}$$

Recalling that $\mathrm{KL}(\mathbb{P}^\star \,\|\, \hat{\mathbb{P}}) = \mathbb{E}_{\mathbb{P}^\star}[\log(\mathrm{d}\mathbb{P}^\star/\mathrm{d}\hat{\mathbb{P}})]$, substituting (54) in the log-likelihood ratio (obtained from (51)) yields,

$$\begin{aligned}
\log \frac{\mathrm{d}\mathbb{P}^\star}{\mathrm{d}\hat{\mathbb{P}}} &= \int_{t_0}^{T} \sigma_s\, \mathbf{e}_s \cdot \mathrm{d}\bar{\mathbf{B}}_s^{\hat{\mathbb{P}}} - \tfrac{1}{2}\int_{t_0}^{T} \sigma_s^2 \,\|\mathbf{e}_s\|_2^2 \,\mathrm{d}s \\
&= \int_{t_0}^{T} \sigma_s\, \mathbf{e}_s \cdot \big(\mathrm{d}\bar{\mathbf{B}}_s^{\mathbb{P}^\star} + \sigma_s\, \mathbf{e}_s \,\mathrm{d}s\big) - \tfrac{1}{2}\int_{t_0}^{T} \sigma_s^2 \,\|\mathbf{e}_s\|_2^2 \,\mathrm{d}s \\
&= \int_{t_0}^{T} \sigma_s\, \mathbf{e}_s \cdot \mathrm{d}\bar{\mathbf{B}}_s^{\mathbb{P}^\star} + \tfrac{1}{2}\int_{t_0}^{T} \sigma_s^2 \,\|\mathbf{e}_s\|_2^2 \,\mathrm{d}s.
\end{aligned}$$

Taking expectation under $\mathbb{P}^\star$,

$$\mathbb{E}_{\mathbb{P}^\star}\left[\log \frac{\mathrm{d}\mathbb{P}^\star}{\mathrm{d}\hat{\mathbb{P}}}\right] = \mathbb{E}_{\mathbb{P}^\star}\left[\int_{t_0}^{T} \sigma_s\, \mathbf{e}_s \cdot \mathrm{d}\bar{\mathbf{B}}_s^{\mathbb{P}^\star}\right] + \tfrac{1}{2}\mathbb{E}_{\mathbb{P}^\star}\left[\int_{t_0}^{T} \sigma_s^2 \,\|\mathbf{e}_s\|_2^2 \,\mathrm{d}s\right]. \tag{55}$$

The first term is the expectation of a stochastic integral. By Assumptions A1–A2, the integrand $\sigma_s \mathbf{e}_s$ is square-integrable (since $\varepsilon_\star^2 = \mathbb{E}_{\mathbb{P}^\star}\left[\int_{t_0}^{T} \sigma_s^2 \,\|\mathbf{e}_s\|_2^2 \,\mathrm{d}s\right] < \infty$); as $\bar{\mathbf{B}}_s^{\mathbb{P}^\star}$ is a $\mathbb{P}^\star$-Brownian motion, the integral is a true $\mathbb{P}^\star$-martingale started at 0 and its expectation under $\mathbb{P}^\star$ vanishes so the first term in the right-hand term of (55) is zero. The second term is exactly $\frac{1}{2}\varepsilon_\star^2$, by the definition of $\varepsilon_\star^2$ in (52).

Data Processing Inequality Step. Let $\pi_{t_0} : C([t_0, T], \mathbb{R}^d) \to \mathbb{R}^d$ be the projection map $\omega \mapsto \omega(t_0)$. The endpoint marginals are push-forwards: $p_{t_0}^\star = (\pi_{t_0})_\# \mathbb{P}^\star$ and $\hat{p}_{t_0} = (\pi_{t_0})_\# \hat{\mathbb{P}}$. By the data-processing inequality (contraction of KL under push-forward),

$$\mathrm{KL}(p_{t_0}^\star \,\|\, \hat{p}_{t_0}) \le \mathrm{KL}(\mathbb{P}^\star \,\|\, \hat{\mathbb{P}}) = \tfrac{1}{2}\varepsilon_\star^2.$$

□

C.3 A Helmholtz–Hodge View of Girsanov's Theorem

In this section we explain how the projected bound of Theorem 3.2 can also be recovered from a Girsanov-type argument. The key point is that Girsanov's theorem is a path-space statement, whereas the Helmholtz–Hodge decomposition identified in Section 3 is intrinsic to marginal dynamics. Consequently, the projected bound is obtained not by applying Girsanov directly to the original learned process, but by first replacing the learned process by a marginally equivalent representative whose drift error is the observable component of the score error.

Recall that the true reverse process has path law $\mathbb{P}^\star$ and satisfies

$$\mathrm{d}\mathbf{Y}_s = \big[\boldsymbol{f}_s(\mathbf{Y}_s) - \sigma_s^2 \nabla \log p_s^\star(\mathbf{Y}_s)\big]\mathrm{d}s + \sigma_s \,\mathrm{d}\bar{\mathbf{B}}_s, \qquad \mathbf{Y}_T \sim p_T^\star, \qquad s : T \to t_0. \tag{56}$$

The learned reverse process has path law $\hat{\mathbb{P}}$ and satisfies

$$\mathrm{d}\hat{\mathbf{Y}}_s = \big[\boldsymbol{f}_s(\hat{\mathbf{Y}}_s) - \sigma_s^2 \mathbf{s}_\theta(\hat{\mathbf{Y}}_s, s)\big]\mathrm{d}s + \sigma_s \,\mathrm{d}\bar{\mathbf{B}}_s, \qquad \hat{\mathbf{Y}}_T \sim p_T^\star, \qquad s : T \to t_0. \tag{57}$$

Moreover, plugging $\nabla \log p_s^\star = \mathbf{s}_\theta(\cdot, s) - \mathbf{e}_s$ in (56) yields,

$$\mathrm{d}\mathbf{Y}_s = \big[\boldsymbol{f}_s(\mathbf{Y}_s) - \sigma_s^2 \mathbf{s}_\theta(\mathbf{Y}_s, s) + \sigma_s^2 \mathbf{e}_s(\mathbf{Y}_s)\big]\mathrm{d}s + \sigma_s \,\mathrm{d}\bar{\mathbf{B}}_s, \qquad \mathbf{Y}_T \sim p_T^\star, \qquad s : T \to t_0. \tag{58}$$

The drift discrepancy between (57) and (56) is given by weighted score estimation errors $\sigma_s^2 \mathbf{e}_s$. Recall from the proof of Proposition C.1 that

$$\mathrm{KL}(p_{t_0}^\star \,\|\, \hat{p}_{t_0}) \le \mathrm{KL}(\mathbb{P}^\star \,\|\, \hat{\mathbb{P}}) = \frac{1}{2}\int_{t_0}^{T} \sigma_s^2 \|\mathbf{e}_s\|^2_{L^2(p_s^\star)}\,\mathrm{d}s. \tag{59}$$

The looseness of this bound comes from the fact that it considers all $\|\mathbf{e}_s\|^2_{L^2(p_s^\star)}$ errors, even if not every component of $\mathbf{e}_s$ actually affects marginals.

Bridge between Girsanov's approach and ours. To bridge this gap, we decompose the error into its observable gradient component and its invisible solenoidal component, as explained in Section 3. Formally, at each time s,

$$\mathbf{e}_s = \Pi_{\mathcal{G}_s}\mathbf{e}_s + \Pi_{\mathcal{G}_s^\perp}\mathbf{e}_s,$$

where we recall that $\mathcal{G}_s := \mathrm{cl}\{\nabla\varphi : \varphi \in C_c^\infty(\mathbb{R}^d)\}$, with the closure taken with respect to $L^2(p_s^\star;\mathbb{R}^d)$. We now construct a marginally equivalent representative in which these invisible directions are removed. As discussed in Section 3, this decomposition implies $\nabla\cdot(p_s^\star \mathbf{e}_s) = \nabla\cdot\big(p_s^\star \Pi_{\mathcal{G}_s}\mathbf{e}_s\big)$, which allows us to rewrite the true marginal dynamics associated with the reverse SDE (58) as in (12), which we restate for convenience:

$$\partial_s p_s^\star = -\nabla\cdot\Big([\boldsymbol{f}_s - \sigma_s^2 \mathbf{s}_\theta(\cdot,s)]p_s^\star\Big) - \frac{\sigma_s^2}{2}\Delta p_s^\star - \sigma_s^2\nabla\cdot\big(p_s^\star \Pi_{\mathcal{G}_s}\mathbf{e}_s\big). \tag{60}$$

Now introduce a process $\tilde{\mathbf{Y}} = (\tilde{\mathbf{Y}}_s)_{s\in[t_0,T]}$ that is the solution to the following SDE

$$\mathrm{d}\tilde{\mathbf{Y}}_s = \big[\boldsymbol{f}_s(\tilde{\mathbf{Y}}_s) - \sigma_s^2\mathbf{s}_\theta(\tilde{\mathbf{Y}}_s,s) + \sigma_s^2\Pi_{\mathcal{G}_s}\mathbf{e}_s(\tilde{\mathbf{Y}}_s)\big]\mathrm{d}s + \sigma_s\,\mathrm{d}\bar{\mathbf{B}}_s, \qquad \tilde{\mathbf{Y}}_T \sim p_T^\star. \tag{61}$$

Then the marginals $(\tilde{p}_s)_{s\in[t_0,T]}$, where $\tilde{p}_s := \mathrm{Law}(\tilde{\mathbf{Y}}_s)$, solve the Fokker-Planck equation (60). We denote by $\tilde{\mathbb{P}}$ the path-space law of (61).

On the other hand, since the true marginals $p_s^\star$ solve the Fokker-Planck associated with the reverse SDE (58), we have

$$\partial_s p_s^\star = -\nabla\cdot\Big([\boldsymbol{f}_s - \sigma_s^2\mathbf{s}_\theta(\cdot,s)]p_s^\star\Big) - \frac{\sigma_s^2}{2}\Delta p_s^\star - \sigma_s^2\nabla\cdot(p_s^\star\mathbf{e}_s). \tag{62}$$

But, since $\nabla\cdot(p_s^\star\mathbf{e}_s) = \nabla\cdot\big(p_s^\star\Pi_{\mathcal{G}_s}\mathbf{e}_s\big)$ and terminal conditions match, the marginal dynamics (62) and (60) exactly match, i.e., $\tilde{p}_s = p_s^\star$ for all $s\in[t_0,T]$.

In particular, endpoint marginals match $\tilde{p}_{t_0} = p_{t_0}^\star$.

We may now apply Girsanov's theorem between $\hat{\mathbb{P}}$ and the observable representative $\tilde{\mathbb{P}}$ exactly as is done in Proposition C.1, replacing $\mathbb{P}^\star$ by $\tilde{\mathbb{P}}$. The SDEs (57) and (61) have the same diffusion coefficient and the same terminal law $p_T^\star$; their drift discrepancy is $\sigma_s^2\Pi_{\mathcal{G}_s}\mathbf{e}_s$. Assuming the analogue of A5 for $\Pi_{\mathcal{G}_s}\mathbf{e}_s$, the corresponding Doléans–Dade exponential (obtained from (50) by replacing $\mathbf{e}_s$ with $\Pi_{\mathcal{G}_s}\mathbf{e}_s$) is a true martingale, Girsanov's formula gives

$$\mathrm{KL}(\tilde{\mathbb{P}}\,\|\,\hat{\mathbb{P}}) = \frac{1}{2}\int_{t_0}^{T}\sigma_s^2\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(\tilde{p}_s)}\,\mathrm{d}s. \tag{63}$$

Using (63), $\tilde{p}_{t_0} = p_{t_0}^\star$ and data-processing, this becomes,

$$\mathrm{KL}(p_{t_0}^\star \,\|\, \hat{p}_{t_0}) = \mathrm{KL}(\tilde{p}_{t_0} \,\|\, \hat{p}_{t_0}) \le \mathrm{KL}(\tilde{\mathbb{P}} \,\|\, \hat{\mathbb{P}}) = \frac{1}{2}\int_{t_0}^{T}\sigma_s^2\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p_s^\star)}\,\mathrm{d}s, \tag{64}$$

which is exactly the bound of Theorem 3.2.

D Optimal Transport Interpretation

The variational characterization to compute the gradient component of errors, given in Proposition 4.1 by,

$$\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p_s^\star)} = \inf_{\mathbf{v}_s:\,\nabla\cdot(p_s^\star\mathbf{v}_s)=\nabla\cdot(p_s^\star\mathbf{e}_s)} \int_{\mathbb{R}^d} \|\mathbf{v}_s(\mathbf{x})\|_2^2\, p_s^\star(\mathbf{x})\,\mathrm{d}\mathbf{x}. \tag{65}$$

has an interpretation in the dynamic formulation of optimal transport. Recall $\mathcal{P}_2(\mathbb{R}^d)$ denotes the space of probability densities on $\mathbb{R}^d$ with finite second moment. Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ be a smooth positive density, and consider a differentiable curve of probability densities $(\mu_\tau)_{\tau\in(-\epsilon,\epsilon)}$ such that $\mu_0 = \mu$. Its tangent vector at $\tau = 0$ is the signed density

$$\zeta := \partial_\tau \mu_\tau|_{\tau=0}, \qquad \int_{\mathbb{R}^d} \zeta(\mathbf{x})\,\mathrm{d}\mathbf{x} = 0.$$

In the Benamou–Brenier geometry [5], this tangent perturbation is represented by a velocity field $\mathbf{v} \in L^2(\mu;\mathbb{R}^d)$ through the continuity equation [1, 67, 17]

$$\zeta = -\nabla\cdot(\mu\mathbf{v}).$$

This representation is not unique: if $\mathbf{w}$ satisfies $\nabla\cdot(\mu\mathbf{w}) = 0$, then $\mathbf{v}+\mathbf{w}$ represents the same tangent vector ζ. Wasserstein geometry removes this ambiguity by selecting the minimum-kinetic-energy representative,

$$\inf_{\mathbf{v}:\,-\nabla\cdot(\mu\mathbf{v})=\zeta} \int_{\mathbb{R}^d} \|\mathbf{v}(\mathbf{x})\|_2^2\,\mu(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

Equivalently, this minimum-energy representative belongs to the closure of gradient fields [1, 77, 17],

$$T_\mu\mathcal{P}_2 = \mathrm{cl}(\{\nabla\varphi : \varphi \in C_c^\infty(\mathbb{R}^d)\}),$$

where the closure cl is taken in $L^2(\mu;\mathbb{R}^d)$. The Otto metric [61] equips this tangent space with the norm obtained by pulling back the $L^2(\mu)$ geometry through the minimal kinetic energy gradient representative,

$$\|\zeta\|^2_{T_\mu\mathcal{P}_2} = \inf_{\mathbf{v}:\,-\nabla\cdot(\mu\mathbf{v})=\zeta} \int_{\mathbb{R}^d} \|\mathbf{v}\|_2^2\,\mu\,\mathrm{d}\mathbf{x}.$$

This is the $H^{-1}(\mu)$-norm [10] of the signed density ζ. To recover the setting of Section 4, fix a diffusion time $s \in [t_0, T]$, take $\mu = p_s^\star$, and let $\zeta_s := -\nabla\cdot(p_s^\star \mathbf{e}_s)$ be the tangent vector generated by the score error (in the Fokker–Planck equation (62) the error contributes $-\sigma_s^2\,\nabla\cdot(p_s^\star\mathbf{e}_s)$ to $\partial_s p_s^\star$). Then, by the definition above together with Proposition 4.1,

$$\|\zeta_s\|^2_{T_{p_s^\star}\mathcal{P}_2} = \inf_{\mathbf{v}:\,\nabla\cdot(p_s^\star\mathbf{v})=\nabla\cdot(p_s^\star\mathbf{e}_s)} \int_{\mathbb{R}^d} \|\mathbf{v}\|_2^2\,p_s^\star\,\mathrm{d}\mathbf{x} = \|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p_s^\star)}.$$

Equivalently, this tangent norm is the $H^{-1}(p_s^\star)$ norm $\|\nabla\cdot(p_s^\star\mathbf{e}_s)\|_{H^{-1}(p_s^\star)}$ of (17): the Otto metric tensor and the $H^{-1}(p_s^\star)$ norm coincide. Thus, the target-weighted observable error is the squared Wasserstein tangent norm of the infinitesimal curve of marginals generated by the score error by the velocity field given by the score error at $p_s^\star$. The solenoidal component of $\mathbf{e}_s$ is invisible because it changes the velocity representation without changing the tangent vector ζ_s, i.e. without changing the infinitesimal motion of probability mass.

E Experimental Setting

E.1 Experimental Setup Fashion-MNIST

We detail the experimental setup used to produce the Fashion-MNIST error-decomposition curves, KL upper-bound estimates, and Spearman correlations of Section 3. All image experiments (on CIFAR-10 and Fashion-MNIST) were run on a single NVIDIA RTX 6000 Blackwell GPU with 96GB memory; a full CIFAR-10 score network training run required approximately 10 hours of wall-clock time.

Training setup. We train standard score-based diffusion models on Fashion-MNIST [82] using a variance-preserving SDE with linear schedule $\beta(t) = \beta_{\min} + t(\beta_{\max} - \beta_{\min})$, $\beta_{\min} = 0.1$, $\beta_{\max} = 20$, time horizon $T = 4$ and truncation $t_0 = 5\times10^{-2}$. The score network $\mathbf{s}_\theta$ is a three-resolution-level U-Net with self-attention [65, 76] parametrized by a base channel multiplier c and a per-resolution depth n_{res}. We sweep three capacity tiers, *tiny* (c=32, n_{res}=1, 1.76M parameters), *small* (c=48, n_{res}=2, 3.92M) and *full* (c=64, n_{res}=2, 6.27M), and train each for 400 epochs with AdamW [55] (learning rate 5×10^{-4} with linear warmup over 2,000 steps and cosine decay, weight decay 10^{-4}, batch size 512, gradient clipping at 1.0). We use a log-uniform time-sampling distribution over $[t_0, T]$, an exponential moving average of the score-network weights with decay 0.9999 used for all evaluations, and five independent seeds per capacity. Permanent checkpoints are saved every 20 epochs, yielding 20 checkpoints per run.

Estimating the error decomposition. At each saved checkpoint we evaluate the gradient and solenoidal components of the score error on the EMA score network via the computational procedure of Section 4.2. The full L^2 error is computed directly from the unbiased residual $\mathbf{r}_\theta(\mathbf{X}_s, s) = \hat{\mathbf{s}}_\theta(\mathbf{X}_s, s) - \mathbf{s}_{\text{target}}(\mathbf{X}_s, s)$ defined in equation (22), by Monte Carlo averaging over a fixed pool of $6{,}400$ noised samples drawn from the Fashion-MNIST training set with t sampled from the same log-uniform distribution used during training (the same $(\mathbf{X}_0, t, \boldsymbol{\varepsilon}, \mathbf{X}_t)$ tuples are reused across all checkpoints of a run to remove sampling noise from the curves). The gradient component is estimated via the dual variational identity of equation (20): at each evaluation, we instantiate a freshly initialized critic potential φ_ψ and train it for $1{,}500$ steps to maximize the empirical objective $\widehat{\mathcal{J}}(\psi)$ of equation (23), using Adam [43] (learning rate 10^{-3} with cosine decay to 10^{-5}, gradient clipping at 1.0, batch size 256). The critic architecture mirrors the score network's encoder–decoder structure but with reduced width (base channel multiplier 24, scalar output, single attention block at the bottleneck), and its gradient $\nabla_{\mathbf{x}}\varphi_\psi$ is taken via automatic differentiation. The solenoidal component is obtained as $\|\Pi_{\mathcal{G}_s^\perp}\mathbf{e}_s\|^2 = \max(\|\mathbf{e}_s\|^2 - \|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2, 0)$, with the floor at zero handling the rare cases where finite-sample noise makes the dual lower bound exceed the empirical full norm. For the final-checkpoint tight bound used in Theorem 3.2, we re-estimate the gradient component with a $2\times$ wider critic (base channel multiplier 48) trained for $5{,}000$ steps with three independent restarts, reporting the across-restart mean and standard deviation.

KL upper-bound estimates. The right panel of Figure 5 reports both Girsanov-style and Helmholtz–Hodge bounds on $\mathrm{KL}(p^\star_{t_0} \| \hat{p}_{t_0})$. The two bounds, equations (14) and (15), are respectively the time integrals $\frac{1}{2}\int_{t_0}^{T} \sigma_s^2 \, \|\mathbf{e}_s\|^2_{L^2(p^\star_s)} \, \mathrm{d}s$ and $\frac{1}{2}\int_{t_0}^{T} \sigma_s^2 \, \|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p^\star_s)} \, \mathrm{d}s$. As in the CIFAR-10 setup (described in the following section), we use a common multiplicative constant $K = (T - t_0)\bar{\sigma}^2/2$ where $\bar{\sigma}^2$ is the schedule-averaged σ_s^2 over $[t_0, T]$. The constant K is identical for both curves and does not depend on the score network, epoch, or seed, so the ratio of the two curves — and therefore the inobservability gap — is faithful to theoretical predictions independently of K.

Sample-quality metric. Sample quality is measured by feature-FID computed in the 128-dimensional penultimate-layer feature space of a small CNN classifier [49] ($> 92\%$ test accuracy on Fashion-MNIST), evaluated on $5{,}000$ samples generated via a 200-step predictor–corrector reverse SDE with the EMA score network. Because the diagnostic and FID are computed from independent quantities (training residual vs. generated samples), no leakage is possible between the two columns of Figure 2.

Spearman rank correlation. Per-seed Spearman correlations are computed across the 20 checkpoints of a run between feature-FID and each of $\|\mathbf{e}\|^2$, $\|\Pi_{\mathcal{G}}\mathbf{e}\|^2$. Pooled correlations are computed across the resulting 60 (seed, epoch) pairs at each capacity. We report Spearman rather than Pearson because both error norms and FID range over multiple orders of magnitude during training, so a rank-based statistic is more meaningful than a linear one. Across all nine (capacity, seed) combinations the gradient-component correlation lies in $[0.93, 0.99]$, whereas the full-error correlation ranges as low as 0.65, indicating that $\|\Pi_{\mathcal{G}}\mathbf{e}\|^2$ is not only a better but also a substantially more reliable proxy for sample quality than the full L^2 score error.

Critic-suboptimality control. The dual estimator of $\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2$ is, at finite critic capacity, a lower bound. To verify that the apparent solenoidal plateau is not an artifact of critic underfitting, we ran an ablation on the *small* capacity tier in which the eval-time critic was trained for $4{,}500$ steps ($3\times$ longer) and with a $2\times$ wider architecture (base channel multiplier 48). The resulting gradient-component curve was within the inter-quartile band of the reported one, and the solenoidal plateau was unchanged at the displayed log-scale resolution. We therefore interpret the plateau as a genuine property of the score network rather than as a critic-capacity artifact.

E.2 Additional Results Fashion-MNIST

The same qualitative behavior holds on Fashion-MNIST across three capacity tiers (*tiny*, *small*, *full*): the gradient component of the score error decreases by more than three orders of magnitude during training while the solenoidal component remains almost flat and dominates the total error throughout (Figure 6, Appendix E.1); the standard Girsanov bound therefore stays effectively constant, and is a loose proxy of learning quality while our bound from Theorem 3.2 tracks sample quality and

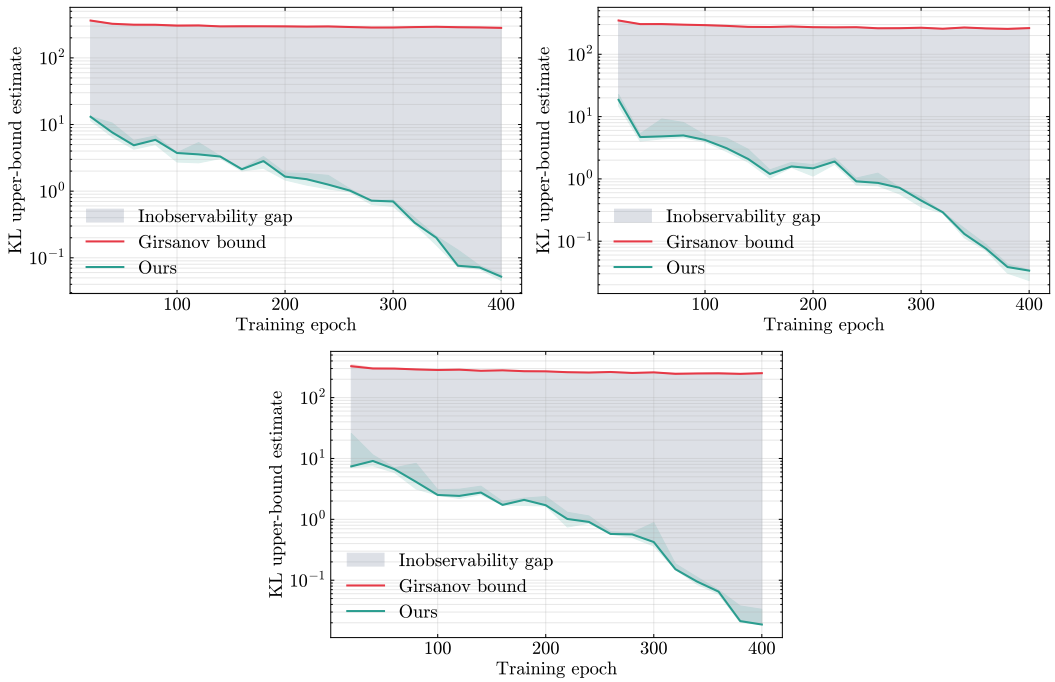

Figure 5: **The Girsanov bound saturates while ours keeps decreasing, across all model capacities.** KL upper-bound estimates on Fashion-MNIST for *tiny*-sized (top left) *small*-sized (top right) and *normal*-sized (bottom) score networks. The Girsanov bound (14) stays effectively constant at the solenoidal floor, whereas our bound from Theorem 3.2 tracks the gradient component and decreases by roughly three orders of magnitude. The shaded *inobservability gap* is the slack the standard analysis incurs by penalizing error components that, by Theorem 3.1, do not affect generated marginals. Median (solid) and inter-quartile range (band) over three seeds. Setup, sampling protocol, and the choice of bound prefactor in Appendix E.1.

continues to decrease (Figure 5). This matches the Spearman correlation results between the FID and the two error components in Figure 2. We deliberately let the model size as well as the dataset vary (see results on CIFAR-10 in Figure 4) to illustrate that the asymmetry between observable and invisible components is not an artifact of architecture, dataset, or model capacity, but a structural property of standard score matching predicted by Theorem 3.1.

The same separation between observable and invisible error components, and the same gap between the standard and observable KL bounds, hold on Fashion-MNIST across three capacity tiers (Appendix E.1, Figures 6–5).

E.3 Experimental Setup CIFAR-10

We detail the experimental setting used to produce the error decomposition and KL upper-bound curves of Figure 4.

Training setup. We train a standard score-based diffusion model on CIFAR-10 with the variance-preserving SDE of [73]. With our notations in Section 2, we choose $\boldsymbol{f}_t(\mathbf{x}) = -\frac{1}{2}\beta(t)\mathbf{x}$ and $\sigma_t = \sqrt{\beta(t)}$, with a linear schedule $\beta(t) = \beta_{\min} + t(\beta_{\max} - \beta_{\min})$, $\beta_{\min} = 0.1$, $\beta_{\max} = 20$, and time horizon $T = 1$ with truncation at $t_0 = 0.01$. We parametrize the score network $\mathbf{s}_\theta$ via a three-resolution level U-Net [65] with self-attention [76] (approximately 12.6M trainable parameters). Models are trained for 350 epochs with AdamW [55] (learning rate 2×10^{-4} with linear warm-up over 3000 steps and cosine decay to 10^{-5}, weight decay 0.01, batch size 128, gradient clipping at 1.0). We use Min-SNR-γ loss weighting [31] with $\gamma = 5$, and maintain an exponential moving average of the score-network weights with decay 0.9999 that is used for all evaluations. Curves are reported as median and inter-quartile range over four independent seeds.

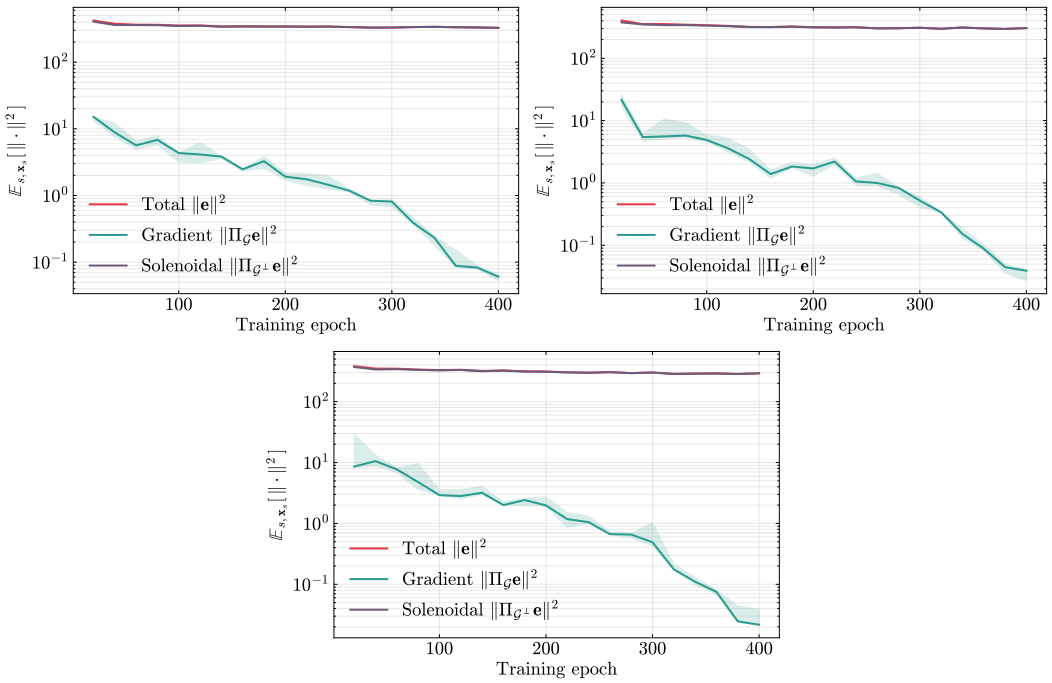

Figure 6: **DSM does not reduce the invisible component of the score error, across all model capacities.** Decomposition of the score estimation error $\mathbf{e}_s$ into its observable gradient component $\Pi_{\mathcal{G}_s}\mathbf{e}_s$ and its invisible solenoidal component $\Pi_{\mathcal{G}_s^\perp}\mathbf{e}_s$ on Fashion-MNIST, for *tiny*-sized (top left) *small*-sized (top right) and *normal*-sized (bottom) score networks. Across all capacities, the gradient component falls by more than three orders of magnitude during training, while the solenoidal component remains essentially flat and accounts for nearly all of the residual error throughout. Median (solid) and interquartile range (band) over five seeds. Setup in Appendix E.1.

Estimating the error decomposition. Every 10 epochs, we evaluate the gradient and solenoidal components of the score error on the EMA score network via the computational procedure described in Section 4.2. The full L^2 error is computed directly from the unbiased residual $\mathbf{r}_\theta(\mathbf{X}_s, s) = \hat{\mathbf{s}}_\theta(\mathbf{X}_s, s) - \mathbf{s}_{\text{target}}(\mathbf{X}_s, s)$ defined in equation (22), by Monte Carlo averaging over a fixed pool of $\sim$ 6,400 noised samples drawn from the CIFAR-10 training set with $t \sim \mathcal{U}[t_0, T]$ (the same $(\mathbf{X}_0, t, \boldsymbol{\varepsilon}, \mathbf{X}_t)$ tuples are reused across all evaluations to remove sampling noise from the curves). The gradient component is estimated via the dual variational identity of equation (20): at each evaluation, we instantiate a freshly initialized critic potential φ_ψ and train it for 4,000 steps to maximize the empirical objective $\widehat{\mathcal{J}}(\psi)$ in equation (23), using Adam [43] (learning rate 10^{-3} with cosine decay to 10^{-5}, gradient clipping at 1.0). The critic architecture mirrors the score network's encoder-decoder structure but with reduced width (base channel multiplier 96, scalar output), and its gradient $\nabla_{\mathbf{x}}\varphi_\psi$ is taken via automatic differentiation. To stabilize the estimator we average over three independent critic restarts per evaluation; the standard deviation of the gradient estimate across restarts is reported alongside the mean. The solenoidal component is then obtained as $\|\Pi_{\mathcal{G}_s^\perp}\mathbf{e}_s\|^2 = \max(\|\mathbf{e}_s\|^2 - \|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2, 0)$, with the floor at zero handling the rare cases where finite-sample noise makes the dual lower bound exceed the empirical full norm.

KL upper-bound estimates. The right panel of Figure 4 reports both Girsanov-style and Helmholtz–Hodge bounds on $\mathrm{KL}(p^\star_{t_0}\|\hat{p}_{t_0})$. The two bounds, equations (14) and (15), are respectively the time integrals $\frac{1}{2}\int_{t_0}^T \sigma_s^2 \|\mathbf{e}_s\|^2_{L^2(p^\star_s)}\,\mathrm{d}s$ and $\frac{1}{2}\int_{t_0}^T \sigma_s^2 \|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p^\star_s)}\,\mathrm{d}s$. For simplicity, we chose the same constant for these two bounds, corresponding to $K = (T - t_0)\bar{\sigma}^2/2$ where $\bar{\sigma}^2$ is the schedule-averaged σ_s^2 over $[t_0, T]$. The constant K is identical for both curves and does not depend on the score network, the epoch, or the seed, making the ratio of the two curves at any epoch (and therefore the inobservability gap) faithful to theoretical predictions.

Algorithm 1 Observable-error critic estimation

Require: Frozen score network $\hat{\mathbf{s}}_\theta$; critic network φ_ψ; training dataset $\mathcal{D} \sim p^\star$; time sampler ν over $[t_0, T]$; noising schedules $\alpha(s), \sigma(s)$; batch size B; critic steps N_{crit}; learning rate η.

Ensure: Estimate $\widehat{\mathcal{L}}_{\mathrm{obs}}$ of $\mathbb{E}_{s\sim\nu}\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2_{L^2(p^\star_s)}$.

1: Initialize critic parameters ψ.
2: **for** $k = 1, \dots, N_{\mathrm{crit}}$ **do**
3: $\quad \{\mathbf{x}_0^{(i)}\}_{i=1}^B \sim \mathcal{D}$; $\{s^{(i)}\}_{i=1}^B \sim \nu$; $\{\boldsymbol{\varepsilon}^{(i)}\}_{i=1}^B \sim \mathcal{N}(0, \mathbf{I}_d)$
4: $\quad \mathbf{x}_s^{(i)} \leftarrow \alpha(s^{(i)})\mathbf{x}_0^{(i)} + \sigma(s^{(i)})\boldsymbol{\varepsilon}^{(i)}$
5: $\quad \mathbf{s}_{\mathrm{target}}^{(i)} \leftarrow -\dfrac{\mathbf{x}_s^{(i)} - \alpha(s^{(i)})\mathbf{x}_0^{(i)}}{\sigma(s^{(i)})^2}$
6: $\quad \mathbf{r}_\theta^{(i)} \leftarrow \mathbf{s}_\theta(\mathbf{x}_s^{(i)}, s^{(i)}) - \mathbf{s}_{\mathrm{target}}^{(i)}$
7: $\quad \mathbf{g}_\psi^{(i)} \leftarrow \nabla_{\mathbf{x}}\varphi_\psi(\mathbf{x}_s^{(i)}, s^{(i)})$
8: $\quad \widehat{\mathcal{J}}_B(\psi) \leftarrow \dfrac{1}{B}\sum_{i=1}^B \left[2\,\mathbf{g}_\psi^{(i)} \cdot \mathbf{r}_\theta^{(i)} - \|\mathbf{g}_\psi^{(i)}\|_2^2\right]$
9: $\quad \psi \leftarrow \psi + \eta\,\nabla_\psi \widehat{\mathcal{J}}_B(\psi)$
10: **end for**
11: Draw a fresh validation batch and compute $\widehat{\mathcal{J}}_{\mathrm{val}}(\psi)$ by repeating lines 3–8 (no update on ψ).
12: $\widehat{\mathcal{L}}_{\mathrm{obs}} \leftarrow \widehat{\mathcal{J}}_{\mathrm{val}}(\psi)$
13: **return** $\widehat{\mathcal{L}}_{\mathrm{obs}}$

Critic-suboptimality control. The dual estimator of $\|\Pi_{\mathcal{G}_s}\mathbf{e}_s\|^2$ is, at finite critic capacity, a lower bound. To verify that the apparent solenoidal plateau is not an artifact of critic underfitting, we ran an ablation in which the eval-time critic was trained for $12{,}000$ steps ($3\times$ longer) and with a wider architecture (base channel multiplier 192 instead of 96). The resulting gradient-component curve was within the interquartile band of the reported one, and the solenoidal plateau was unchanged at the displayed log-scale resolution. We therefore interpret the plateau as a genuine property of the score network rather than as a critic-capacity artifact.

E.4 Low-Dimensional Experiment with Exact Decompositions

The target is a Gaussian mixture with 8 modes on a circle of radius 2 with uneven weights so that forward marginals and scores are available in closed form. The forward SDE is set to a variance-preserving schedule with $\beta_{\min} = 0.1$, $\beta_{\max} = 20$, $T = 1$, and t_0 chosen so that the marginal noise level at t_0 is 0.05 ($t_0 \approx 0.0116$). The perturbed mixtures $\tilde{p}$ are obtained by adding $\mathcal{N}(0, 0.6^2)$ noise to the log-weights and $\mathcal{N}(0, 0.15^2)$ noise to the means (four draws). We add both purely solenoidal and purely gradient error fields. Both the standard Girsanov bound and ours are evaluated by Gauss–Legendre quadrature in $\log s$. The learned endpoint $\hat{p}_{t_0}$ is obtained by solving the reverse Fokker–Planck equation (8) from $\hat{p}_T = p^\star_T$ on a 224^2 grid over $[-4.5, 4.5]^2$, with 200 steps. KL is computed against a reference solve with the exact score on the same grid and time steps, so that discretisation errors largely cancel. The reference solve differs from the analytic $p^\star_{t_0}$ by 10^{-5}.

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 state the paper's main theoretical contributions. These claims are matched to formal results in Sections 3 and 4, Appendix C, and to empirical validation described in the main text and in Appendix E.

   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: Limitations are discussed in Section 5 (Discussion and Future Work).

   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 explicitly stated (e.g. A1, . . .), discussed, and referenced in order depending on the result for which they are needed. Theorems are stated in the main text, and their proofs are systematically included in the appendix and appropriately referenced in the main text.

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: We include an Experimental Setting section in Appendix E in which we describe details of both the synthetic and real-data settings (dataset, architecture, optimization details etc.). These details are sufficient to understand and reproduce the reported empirical results supporting the theoretical claims.

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: [Yes]

   Justification: Code for all experiments, with instructions to reproduce every figure, is available at GitHub. CIFAR-10 and FashionMNIST datasets are publicly available.

   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: Appendix E specifies the main experimental settings, including the Fashion-MNIST and CIFAR-10 diffusion setup, U-Net architecture, optimizer, learning rate, batch size, training epochs, sampling method, number of generations, number of seeds, and FID evaluation protocol.

   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: Experiments include statistical significance results via error bars on bar plots (e.g. Figure 2) or shaded area on line plots (e.g. Figures 4, 5, 6). The meaning of these statistical significance results are detailed in the experimental section in Appendix E.

   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: [Yes]

   Justification: These details can be found in Appendix E.

   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: The research is mainly theoretical, on the properties of diffusion models, and the empirical validation uses standard public image datasets. It does not involve human-subject experiments, private data collection, deceptive practices, or the release of high-risk models.

   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: [No]

Justification: The paper is primarily foundational. We do not believe that there are potential societal consequences that need to be highlighted, so the paper does not currently include a societal impact discussion.

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 does not introduce or release a high-risk pretrained generative model, scraped dataset, or deployment-ready system. The experiments use synthetic distributions and standard public benchmark data for scientific validation.

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: [No]

Justification: The paper cites the existing datasets, models, and methods used in the experiments, including CIFAR-10 and Fashion-MNIST and standard diffusion/sampling components, but the current version does not explicitly list licenses or terms of use for all external assets.

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 introduce a new dataset, benchmark, pretrained model, or other reusable asset. The main contributions are theoretical results and empirical validation.

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: This theoretical work did 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: This theoretical work did 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: The core methodology does not use LLMs as an important, original, or non-standard component. Any use of LLMs, if any, was limited to writing, editing, or formatting and did not affect the scientific content or methodology.

    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.