DDPM SCORE MATCHING IS ASYMPTOTICALLY EFFICIENT

Sinho Chewi & Alkis Kalavasis & Anay Mehrotra & Omar Montasser
Yale University

ABSTRACT

The success of score-based generative models (SGMs), and particularly denoising diffusion probabilistic models (DDPMs), rests on the statistical technique of *score matching*, for which rigorous guarantees are nascent. In fact, recent work has shown that for estimation in parametric models, a variant of score matching known as implicit score matching is provably statistically inefficient for multimodal densities that are common in practice. In contrast, under mild conditions, we show that denoising score matching in DDPMs is asymptotically efficient, *i.e.*, the DDPM estimator is asymptotically normal with covariance matrix given by the inverse Fisher information. Our proof is based on a pointwise relationship between the empirical risks of DDPM and maximum likelihood estimation.

1 INTRODUCTION

Score-based generative models (SGMs), also known as diffusion models, have emerged as a popular approach to generate samples from complex data distributions. These models leverage learned score functions—that is, the logarithmic gradients of the probability density—to progressively transform white noise into samples from the target data distribution by following a stochastic differential equation (SDE). The remarkable empirical success of SGMs has not only led to impressive practical applications but has also spurred significant interest within the statistics community toward establishing rigorous theoretical foundations for SGMs.

A central component underlying SGMs is *score estimation* (Hyvärinen, 2005), which transforms the problem of learning the score function into a regression objective amenable to first-order optimization. Despite a number of recent works investigating its efficacy, a complete statistical understanding remains to be developed. In this work, we aim to address this gap by investigating this problem in the setting of point estimation in a finite-dimensional family $\mathscr{P} \coloneqq \{P_\theta \colon \theta \in \Theta \subseteq \mathbb{R}^p\}$.

Within this parametric framework, the prior work of Koehler et al. (2023) investigated a variant of score matching known as *implicit score matching* (ISM): Given i.i.d. samples $x^{(1)}, \dots, x^{(n)}$ drawn from $P_{\theta^\star}$, $\theta^\star \in \Theta$, the ISM estimator is

$$\widehat{\theta}_n^{\mathrm{ISM}} \coloneqq \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \left\{ \|\boldsymbol{\nabla} \log P_\theta(x^{(i)})\|^2 + 2\,\Delta \log P_\theta(x^{(i)}) \right\}.$$

Under appropriate regularity conditions, they prove asymptotic normality: $\sqrt{n}\,(\widehat{\theta}_n^{\mathrm{ISM}} - \theta^\star) \xrightarrow{\mathsf{d}} \mathcal{N}(0, \Sigma^{\mathrm{ISM}}(\theta^\star))$. Moreover, they bound the operator norm of $\Sigma^{\mathrm{ISM}}(\theta^\star)$ in terms of the asymptotic covariance of the maximum likelihood estimator (MLE)—*i.e.*, the inverse Fisher information matrix—and the so-called restricted Poincaré constant of the model. Follow-up work by Qin & Risteski (2024) generalized this asymptotic efficiency result to generalized (implicit) score matching estimators (Lyu, 2009) by establishing a connection between the *mixing time* of broad classes of Markov processes, and the statistical efficiency of an appropriately chosen *generalized score* matching loss (GISM). Under this framework, they managed to show that for Gaussian mixtures in d dimensions, the generalized score estimator is asymptotically normal with covariance matrix $\Sigma^{\mathrm{GISM}}(\theta^\star)$ which has an operator norm that is, roughly speaking, at most $\mathrm{poly}(d)$ times the (squared) operator norm of the inverse Fisher information (bypassing the lower bounds of Koehler et al. (2023)).

In short, both works (Koehler et al., 2023; Qin & Risteski, 2024) indicated strong statistical properties of (generalized) ISM. That said, they still cannot match the performance of MLE or come within a constant factor of it for general families $\mathscr{P}$, and they left open whether some diffusion-based estimator can achieve the statistical efficiency of MLE under mild assumptions on $\mathscr{P}$. This is particularly interesting as SGMs have seem immense success in sampling from multimodal densities.

Remark. The full-version of this paper more thoroughly explores the relation between DDPM score matching and different notions of distribution learning; it is available at `https://arxiv.org/abs/2504.05161`

In contrast to implicit score matching based estimators above, denoising diffusion probabilistic models (DDPMs)—arguably the most popular variant used in practice—do not rely on implicit score matching. Instead, DDPMs employ an alternative known as *denoising score matching* (Hyvärinen, 2008; Vincent, 2011) and extend the idea by applying score matching at many different noise levels (Song & Ermon, 2019; Ho et al., 2020; Yang et al., 2023). This raises a natural and fundamental question:

What is the statistical efficiency of DDPM score matching?

Toward answering this question, we consider the following idealized estimator. Below, $P_{\theta,t}$ denotes the law of $x_t \coloneqq \exp(-t)\, x_0 + \sqrt{1-\exp(-2t)}\, z_t$, where $x_0 \sim P_\theta$ and $z_t \sim \mathcal{N}(0, I)$ are independent. We provide relevant background on the DDPM objective in Appendix B.

Definition 1 (DDPM estimator)**.** *Fix a terminal time $T > 0$. Given samples $x_0^{(1)}, \dots, x_0^{(n)}$ and a family $\{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$, the DDPM estimator is $\widehat{\theta}_n^{\mathrm{DDPM}} \coloneqq \arg\min_{\theta\in\Theta} \widehat{\mathcal{R}}_n^{\mathrm{DDPM}}(\theta)$, where*

$$\widehat{\mathcal{R}}_n^{\mathrm{DDPM}}(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \int_0^T \mathbb{E}\Big[\|\nabla \log P_{\theta,t}(x_t^{(i)})\|^2 + \big\langle \nabla \log P_{\theta,t}(x_t^{(i)}), \frac{2z_t^{(i)}}{\sqrt{1-\exp(-2t)}}\big\rangle \;\Big|\; x_0^{(i)}\Big]\,\mathrm{d}t$$

and for each $i \in [n]$ and $t \in [0, T]$, we draw $z_t^{(i)} \sim \mathcal{N}(0, I)$ independently from $x_0^{(i)}$ and define the noised sample $x_t^{(i)} \coloneqq \exp(-t)\, x_0^{(i)} + \sqrt{1-\exp(-2t)}\, z_t^{(i)}$.

The expectation above is often replaced with $\mathbb{E}[\|\nabla \log P_{\theta,t}(x_t^{(i)}) + {}^{z_t^{(i)}}\!/\sqrt{1-\exp(-2t)}\|^2 \mid x_0^{(i)}]$, which formally defines an equivalent objective by completing the square. However, we prefer to write the objective as above since $\int_0^T \mathbb{E}[\|{}^{z_t^{(i)}}\!/\sqrt{1-\exp(-2t)}\|^2]\,\mathrm{d}t = \infty$.

1.1 Our contribution: DDPM is asymptotically efficient

Our main result is that, under mild regularity assumptions on the distribution family $\mathcal{P}$ (essentially the same conditions needed for the asymptotic normality of the MLE, see Assumption 1) and by choosing the terminal time $T = T_n$ to grow sufficiently rapidly with the number of samples n (namely, $T_n - \frac{1}{2}\log n \to \infty$), the DDPM estimator $\widehat{\theta}_n^{\mathrm{DDPM}}$ converges in distribution to a Gaussian centered at $\theta^\star$ with covariance *exactly* equal to the inverse Fisher information. Recall that the inverse Fisher information is also the asymptotic covariance of the MLE and is the best possible for any unbiased estimator (by the Cramér–Rao or information inequality), so this statement can be interpreted as a form of asymptotic optimality; furthermore, by comparison of experiments, the MLE can be shown to be locally asymptotically minimax.

To state the result formally, let $\widehat{\theta}_n^{\mathrm{DDPM}}$ denote the DDPM estimator as defined in Definition 1 on n i.i.d. samples $x_0^{(1)}, \dots, x_n^{(n)} \sim P_{\theta^\star}$.

Informal Theorem 1 (DDPM is asymptotically efficient)**.** *Under standard assumptions (see Theorem 1 for more precise statement),*

$$\sqrt{n}\,\big(\widehat{\theta}_n^{\mathrm{DDPM}} - \theta^\star\big) \xrightarrow{\mathsf{d}} \mathcal{N}(0, I(\theta^\star)^{-1}) \qquad \textit{as} \qquad n \to \infty\,,$$

where $I(\theta^\star)$ denotes the Fisher information matrix at $\theta^\star$.

Informal Theorem 1 provides a principled explanation for the statistical power of the DDPM estimator in the asymptotic regime. This finding not only advances our theoretical understanding of DDPM but also sheds some light on its empirical success in generating highly multimodal densities that pose significant challenges for standard score matching approaches (as proved by Koehler et al., 2023). In terms of the techniques used to prove Informal Theorem 1, our main tool is a strong *pointwise* relationship between the DDPM and MLE objectives.

Proposition 1 (Tight connection between DDPM and MLE)**.** *The DDPM objective $\widehat{\mathcal{R}}_n^{\mathrm{DDPM}}$ and the maximum likelihood objective $\widehat{\mathcal{R}}_n^{\mathrm{MLE}}$ satisfy:*

$$\widehat{\mathcal{R}}_n^{\mathrm{MLE}}(\theta) = \widehat{\mathcal{R}}_n^{\mathrm{DDPM}}(\theta) + C_{d,T} + \frac{1}{n}\sum\nolimits_{i=1}^{n} \mathsf{KL}\big(Q_{T|0}(\cdot \parallel X_0^{(i)}) \,\big\|\, P_{\theta,T}\big)$$

where $\widehat{\mathcal{R}}_n^{\mathrm{MLE}}(\theta) \coloneqq -\frac{1}{n}\sum_{i=1}^n \log P_\theta(X_0^{(i)})$, $\widehat{\mathcal{R}}_n^{\mathrm{DDPM}}$ is given in Definition 1, and $C_{d,T} = d\,(T + \frac{1}{2}\log(2\pi e\,(1-e^{-2T})))$ is a fixed constant.

Related work. Closest to our paper is the work of Koehler et al. (2023) that studies the implicit score matching estimator, as discussed in the introduction. To the best of our knowledge, the first appearance of an objective such as Definition 1 for the purpose of point estimation is Shah et al.

(2023), in the context of Gaussian mixture models, which also shows how to algorithmically minimize the DDPM objective (at carefully selected noise levels). We are not aware of general statistical theory for $\widehat{\theta}_n^{\mathrm{DDPM}}$. Most works studying score estimation in DDPM instead consider estimating the score functions at different times separately (as opposed to $\widehat{\theta}_n^{\mathrm{DDPM}}$, which finds the value of the parameter that optimizes an objective using all of the scores). In particular, a line of work shows that score estimation can achieve minimax rates for density estimation. We review these and other related works in Appendix C. Finally, we note that variants of Proposition 1 have appeared in the literature, *e.g.*, Song et al. (2021b) shows that the DDPM loss can be pointwise lower bounded in terms of the MLE loss, and Chen et al. (2022) proves an analogous result for the Schrödinger bridge. A similar formula was also put forth in the work of Song et al. (2021b), where it was presented as a variational lower bound on the log-likelihood, although the connection likely dates back even earlier to ideas by Jarzynski (Jarzynski, 1997).

We now proceed with the proof of Informal Theorem 1 regarding the asymptotic efficiency of DDPM score matching. Given a probability measure P over $\mathbb{R}^d$, we let P_t denote the law of the Ornstein–Uhlenbeck (OU) process started at time $t \in [0, T]$ (abstracting the notation of the previous section). We further denote by $Q_{t|0}$ the transition density of the OU process.[1]

Step 1 (An identity). First, we prove Lemma 1, which, as we will see is similar to Proposition 1.

Lemma 1 (Likelihood identity)**.** *Let P be a continuous density over $\mathbb{R}^d$ with finite second moment. Then, for all $x_0 \in \mathbb{R}^d$,*

$$\int \log P_T \, \mathrm{d}Q_{T|0}(\cdot \mid x_0) - \underbrace{\log P(x_0)}_{\text{log-density at } x_0}$$
$$= \underbrace{\int_0^T \int \left\{ \|\nabla \log P_t\|^2 - 2 \left\langle \nabla \log P_t, \nabla \log Q_{t|0}(\cdot \mid x_0) \right\rangle \right\} \mathrm{d}Q_{t|0}(\cdot \mid x_0) \, \mathrm{d}t}_{\text{integrated DDPM score matching objective at } x_0} + \underbrace{d \cdot T}_{\text{known constant}} .$$

Proof of Lemma 1. Let $(B_t)_{t \geq 0}$ be standard Brownian motion and let $(X_t)_{t \geq 0}$ denote the OU process started at $X_0 = x_0$. By parabolic regularity (or direct computation with the OU semigroup), the mapping $(t, x) \mapsto P_t(x)$ is strictly positive and smooth on $\mathbb{R}_{>0} \times \mathbb{R}^d$, with $P_t \to P$ pointwise as $t \searrow 0$. Therefore, the Fokker–Planck equation implies

$$\partial_t \log P_t = \frac{\Delta P_t + \operatorname{div}(P_t \, x_t)}{P_t} = \Delta \log P_t + \|\nabla \log P_t\|^2 + d + \langle \nabla \log P_t, x_t \rangle .$$

By Itô's formula,

$$\begin{aligned} \mathrm{d} \log P_t(X_t) &= \left\{ \partial_t \log P_t(X_t) - \langle \nabla \log P_t(X_t), X_t \rangle + \Delta \log P_t(X_t) \right\} \mathrm{d}t + \sqrt{2} \langle \nabla \log P_t(X_t), \mathrm{d}B_t \rangle \\ &= \left\{ \|\nabla \log P_t(X_t)\|^2 + 2 \, \Delta \log P_t(X_t) + d \right\} \mathrm{d}t + \sqrt{2} \langle \nabla \log P_t(X_t), \mathrm{d}B_t \rangle . \end{aligned}$$

Integrating over time and taking expectations, for $\varepsilon > 0$,

$$\mathbb{E}\big[\log P_T(X_T) - \log P_\varepsilon(X_\varepsilon)\big] = d \, (T - \varepsilon) + \int_\varepsilon^T \mathbb{E}\big\{ \|\nabla \log P_t(X_t)\|^2 + 2 \, \Delta \log P_t(X_t) \big\} \, \mathrm{d}t ,$$

where we used the fact that $\{\int_\varepsilon^t \langle \nabla \log P_s(X_s), \mathrm{d}B_s \rangle\}_{t \in [\varepsilon, T]}$ is a martingale which, in turn, can be deduced because $\mathbb{E}[\|\nabla \log P_t(X_t)\|^2] = O(1/t^2)$ (cf. Otto & Villani, 2001). On the other hand, for any $t > 0$, we note that

$$\begin{aligned} \int \langle \nabla \log P_t(x_t), \nabla \log Q_{t|0}(x_t \mid x_0) \rangle \, Q_{t|0}(\mathrm{d}x_t \mid x_0) &= \int \langle \nabla \log P_t(x_t), \nabla Q_{t|0}(x_t \mid x_0) \rangle \, \mathrm{d}x_t \\ &= - \int \Delta \log P_t(x_t) \, Q_{t|0}(\mathrm{d}x_t \mid x_0) . \end{aligned}$$

Substituting this in and taking $\varepsilon \searrow 0$ completes the proof. □

[1] We overload the notation using the same symbols for probability distributions and their Lebesgue densities.

Step 2 (Relating MLE and DDPM). Therefore, if we consider the one-sample empirical risks (where Equation (1) is proved in Appendix B),

$$\widehat{\mathcal{R}}^{\mathrm{MLE}}(\theta) = -\log P_\theta(x_0)\,,$$

$$\widehat{\mathcal{R}}^{\mathrm{DDPM}}(\theta) = \int_0^T \int \{\|\boldsymbol{\nabla} \log P_{\theta,t}\|^2 - 2\,\langle \boldsymbol{\nabla} \log P_{\theta,t}, \boldsymbol{\nabla} \log Q_{t|0}(\cdot \mid x_0)\rangle\}\,\mathrm{d}Q_{t|0}(\cdot \mid x_0)\,\mathrm{d}t\,, \quad (1)$$

we can re-write the identity of Lemma 1 as

$$\begin{aligned}\widehat{\mathcal{R}}^{\mathrm{MLE}}(\theta) &= \widehat{\mathcal{R}}^{\mathrm{DDPM}}(\theta) + dT - \int \log P_{\theta,T}\,\mathrm{d}Q_{T|0}(\cdot \mid x_0) \\ &= \widehat{\mathcal{R}}^{\mathrm{DDPM}}(\theta) + d\left(T + \frac{1}{2}\log(2\pi e\,(1-\exp(-2T)))\right) + \mathsf{KL}(Q_{T|0}(\cdot \mid x_0) \parallel P_{\theta,T})\,, \quad (2)\end{aligned}$$

where the last line follows by adding and subtracting $\int \log Q_{T|0}(\cdot \mid x_0)\,\mathrm{d}Q_{T|0}(\cdot \mid x_0)$ and using the formula for the differential entropy of a Gaussian. Equation (2) is a slightly different version of Proposition 1 that we will use to complete the proof.

Step 3 (Exponential decay of KL). The last term in (2) is controlled by the following lemma.

Lemma 2. *For any probability measure P with finite second moment and any $x_0 \in \mathbb{R}^d$,*

$$\mathsf{KL}(Q_{T|0}(\cdot \mid x_0) \parallel P_T) \le \frac{1}{\exp(2T)-1}\left(\|x_0\|^2 + \int \|x\|^2\,P(\mathrm{d}x)\right).$$

Proof of Lemma 2. By the dimension-free log-Harnack inequality (see, *e.g.*, Bobkov et al., 2001; Wang, 2006; Altschuler & Chewi, 2024), $\mathsf{KL}(Q_{T|0}(\cdot \mid x_0), P_T) \le W_2^2(\delta_{x_0}, P)/\{2\,(\exp(2T)-1)\}$. The result follows from the triangle inequality for W_2. □

Statement and Proof of Informal Theorem 1. To state our result , we build on the following standard conditions for asymptotic normality of the MLE. Note that it is implicitly assumed that the MLE exists for sufficiently large n. (This assumption could also be relaxed.)

Assumption 1 (Conditions for asymptotic normality of MLE (van der Vaart, 1998))**.** *The family $\{P_\theta\}_{\theta\in\Theta}$ is differentiable in quadratic mean (DQM) at an interior point $\theta^\star \in \Theta \subseteq \mathbb{R}^p$. Furthermore, there exists a function L such that for all θ, θ' in a neighborhood of Θ, $|\log P_\theta - \log P_{\theta'}| \le L\,\|\theta-\theta'\|$ with $\int L^2\,\mathrm{d}P_{\theta^\star} < \infty$. The Fisher information matrix $I_{\theta^\star}$ is positive definite. Finally, the MLE $\widehat{\theta}_n^{\mathrm{MLE}}$ is consistent: $\widehat{\theta}_n^{\mathrm{MLE}} \to \theta^\star$ in probability as $n \to \infty$.*

Here, the DQM condition weakens the classical assumptions for asymptotic normality of the MLE, which require the existence of a third derivative of $\theta \mapsto \log P_\theta$, and instead asks for the existence of a derivative of $\theta \mapsto \sqrt{P_\theta}$ at $\theta^\star$ in $L^2(P_{\theta^\star})$. This covers non-differentiable examples such as the two-sided exponential location family. Under Assumption 1, it is shown in van der Vaart (1998, Theorem 5.39) that $\sqrt{n}\,(\widehat{\theta}_n^{\mathrm{MLE}} - \theta^\star) \xrightarrow{\mathsf{d}} \mathcal{N}(0, I(\theta^\star)^{-1})$. We prove the following result.

Theorem 1 (Asymptotic normality of the DDPM estimator)**.** *Adopt Assumption 1. Consider the DDPM estimator $\widehat{\theta}_n^{\mathrm{DDPM}}$ where the time T_n of the diffusion satisfies $T_n - \frac{1}{2}\log n \to \infty$. Assume that for some neighborhood Θ' of $\theta^\star$, $\sup_{\theta\in\Theta'} \int \|x\|^2\,P_\theta(\mathrm{d}x) < \infty$, and that the DDPM estimator is consistent. Then, the DDPM estimator is asymptotically efficient: $\sqrt{n}\,(\widehat{\theta}_n^{\mathrm{DDPM}} - \theta^\star) \xrightarrow{\mathsf{d}} \mathcal{N}(0, I(\theta_\star)^{-1})$.*

Proof of Theorem 1. We modify the proof of van der Vaart (1998, Theorem 5.39), which relies on Theorem 5.23 therein. For $\theta \in \Theta$, let $m_\theta := \log p_\theta$ and $\mathbb{P}_n := (1/n)\sum_{i=1}^n \delta_{X_i}$. In order to invoke Theorem 5.23, it suffices to show that $\mathbb{P}_n m_{\widehat{\theta}_n^{\mathrm{DDPM}}} \ge \mathbb{P}_n m_{\widehat{\theta}_n^{\mathrm{MLE}}} - o_{P_{\theta^\star}}(n^{-1})$. By (2),

$$\begin{aligned}-\mathbb{P}_n m_{\widehat{\theta}_n^{\mathrm{DDPM}}} &= \widehat{\mathcal{R}}_n^{\mathrm{DDPM}}(\widehat{\theta}_n^{\mathrm{DDPM}}) + c_{d,T} + \mathbb{P}_n \mathsf{err}(\widehat{\theta}_n^{\mathrm{DDPM}}) \\ &\le \widehat{\mathcal{R}}_n^{\mathrm{DDPM}}(\widehat{\theta}_n^{\mathrm{MLE}}) + c_{d,T} + \mathbb{P}_n \mathsf{err}(\widehat{\theta}_n^{\mathrm{DDPM}}) \\ &= -\mathbb{P}_n m_{\widehat{\theta}_n^{\mathrm{MLE}}} + \mathbb{P}_n[\mathsf{err}(\widehat{\theta}_n^{\mathrm{DDPM}}) - \mathsf{err}(\widehat{\theta}_n^{\mathrm{MLE}})]\,,\end{aligned}$$

where $c_{d,T}$ is a constant and $\mathsf{err}(\theta, x) := \mathsf{KL}(Q_{T|0}(\cdot \mid x) \parallel P_{\theta,T})$. Since err is non-negative, it yields $\mathbb{P}_n m_{\widehat{\theta}_n^{\mathrm{DDPM}}} \ge \mathbb{P}_n m_{\widehat{\theta}_n^{\mathrm{MLE}}} - \mathbb{P}_n \mathsf{err}(\widehat{\theta}_n^{\mathrm{DDPM}})$. Since the DDPM estimator is consistent, Lemma 2 and our assumptions imply $P_{\theta_\star}\mathsf{err}(\widehat{\theta}_n^{\mathrm{DDPM}}) \le 2\,(\exp(2T)-1)^{-1} \sup_{\theta'\in\Theta} \int \|x\|^2\,P_\theta(\mathrm{d}x) = o(1/n)$. By Markov's inequality, $\mathbb{P}_n \mathsf{err}(\widehat{\theta}_n^{\mathrm{DDPM}}) = o_{P_{\theta^\star}}(1/n)$. The rest of the proof is unchanged. □

The assumption of consistency for the MLE and the DDPM estimators is typically mild and can be handled by standard tools, *e.g.*, van der Vaart (1998, §5.2).

REFERENCES

Jason M. Altschuler and Sinho Chewi. Shifted composition I: Harnack and reverse transport inequalities. *IEEE Transactions on Information Theory*, pp. 1–1, 2024.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Cham, 2014.

Sergey G. Bobkov, Ivan Gentil, and Michel Ledoux. Hypercontractivity of Hamilton–Jacobi equations. *J. Math. Pures Appl. (9)*, 80(7):669–696, 2001.

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

Sitan Chen, Vasilis Kontonis, and Kulin Shah. Learning general Gaussian mixtures with efficient score matching. *arXiv preprint 2404.18893*, 2024.

Tianrong Chen, Guan-Horng Liu, and Evangelos A. Theodorou. Likelihood training of Schrödinger bridge using forward-backward SDEs theory. In *The Tenth International Conference on Learning Representations, ICLR*, 2022.

Hugo Cui, Florent Krzakala, Eric Vanden-Eijnden, and Lenka Zdeborova. Analysis of learning a flow-based generative model from limited sample complexity. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

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

Aapo Hyvärinen. Optimal approximation of signal priors. *Neural Computation*, 20(12):3087–3110, 12 2008.

Christopher Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690, 1997.

Frederic Koehler and Thuy-Duong Vuong. Sampling multimodal distributions with the vanilla score: benefits of data-based initialization. In *The Twelfth International Conference on Learning Representations*, 2024.

Frederic Koehler, Alexander Heckett, and Andrej Risteski. Statistical efficiency of score matching: the view from isoperimetry. In *The Eleventh International Conference on Learning Representations*, 2023.

Siwei Lyu. Interpretation and generalization of score matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pp. 359—366, Arlington, Virginia, USA, 2009. AUAI Press.

Song Mei and Yuchen Wu. Deep networks as denoising algorithms: sample-efficient learning of diffusion models in high-dimensional graphical models. *IEEE Trans. Inform. Theory*, 71(4): 2930–2954, 2025.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26517–26582. PMLR, 7 2023.

Felix Otto and Cédric Villani. Comment on: "Hypercontractivity of Hamilton–Jacobi equations" [J. Math. Pures Appl. (9) **80** (2001), no. 7, 669–696] by S. G. Bobkov, I. Gentil and M. Ledoux. *J. Math. Pures Appl. (9)*, 80(7):697–700, 2001.

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

Kulin Shah, Sitan Chen, and Adam Klivans. Learning mixtures of Gaussians using the DDPM objective. *Advances in Neural Information Processing Systems*, 36:19636–19649, 2023.

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

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

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1415–1428. Curran Associates, Inc., 2021b.

Aad W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

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

Feng-Yu Wang. Dimension-free Harnack inequality and its applications. *Front. Math. China*, 1(1): 53–72, 2006.

Andre Wibisono, Yihong Wu, and Kaylee Y. Yang. Optimal score estimation via empirical Bayes smoothing. In Shipra Agrawal and Aaron Roth (eds.), *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*. PMLR, 6 2024.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: a comprehensive survey of methods and applications. *ACM Comput. Surv.*, 56(4), 11 2023.

Supplementary Material For "DDPM Score Matching Is Asymptotically Efficient"

CONTENTS

A NOTATION

In this section, we briefly present the notation used in the main body. We focus on continuous distributions over $\mathbb{R}^d$ that are absolutely continuous with respect to the Lebesgue measure. Given a distribution P, for each point $x \in \mathbb{R}^d$, we abuse notation by using $P(x)$ to denote its Lebesgue density evaluated at x. We use standard definitions of distances and divergences between distributions. Namely, for two distributions P and Q over $\mathbb{R}^d$, the KL divergence of P with respect to Q is $\mathsf{KL}(P \parallel Q) \coloneqq \int \log \frac{\mathrm{d}P}{\mathrm{d}Q} \, \mathrm{d}P$ (provided $P \ll Q$), and the 2-Wasserstein distance between P and Q is $W_2(P, Q) = \inf_{\gamma \in \mathcal{C}(\mu,\nu)} \left(\int \|x - y\|^2 \, \gamma(\mathrm{d}x, \mathrm{d}y) \right)^{1/2}$, where the infimum is over the set $\mathcal{C}(\mu, \nu)$ of all couplings of P and Q.

B BACKGROUND ON DENOISING DIFFUSION PROBABILISTIC MODELING

In this section, we provide standard background on denoising diffusion probabilistic models (DDPMs); see Chen et al. (2023) for further details on sampling.

DDPM employs two Markov chains. The first Markov chain iteratively adds noise to the data and the second Markov chain reverses this process, converting noise back to the original data. The first Markov chain is usually handcrafted, the most prevalent choice being the addition of standard Gaussian noise. The second Markov chain, *i.e.*, the reverse process, is parameterized by learned neural networks. Below, we present the continuous time extension of DDPM with standard Gaussian noise, which corresponds to the Ornstein–Uhlenbeck (OU) process in continuous time.

Forward process arising from OU. The forward process arising from the OU process is the following stochastic differential equation (SDE):

$$\mathrm{d}X_t = -X_t \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}B_t \,, \qquad X_0 \sim P_{\theta^\star} \,, \tag{3}$$

where $(B_t)_{t \geq 0}$ is a standard Brownian motion in $\mathbb{R}^d$. The forward process transforms samples from the data distribution $P_{\theta^\star}$ into standard Gaussian noise. We denote by $P_{\theta^\star, t}$ the law of X_t. It holds that $\lim_{t \to \infty} P_{\theta^\star, t} = \gamma$. In fact, the convergence is exponentially fast in many metrics and divergences (Bakry et al., 2014).

Reversing the OU process. The ultimate goal of generative modeling is to generate samples from $P_{\theta^\star}$. To this end, we must reverse the process (3) in time, which yields the reverse process that transforms pure noise back to samples from the target distribution.

Fix a terminal time $T > 0$. Denote

$$X_t^{\leftarrow} = X_{T-t} \,, \qquad t \in [0, T] \,.$$

It turns out that the time reversal of (3) is

$$\mathrm{d}X_t^{\leftarrow} = \{X_t^{\leftarrow} + 2 \, \boldsymbol{\nabla} \log P_{\theta^\star, T-t}(X_t^{\leftarrow})\} \, \mathrm{d}t + \sqrt{2} \, \mathrm{d}B_t \,, \qquad X_0^{\leftarrow} \sim P_{\theta^\star, T} \,, \tag{4}$$

where now $(B_t)_{t \geq 0}$ is the reversed Brownian motion.

DENOISING SCORE MATCHING WITH DDPM

To implement the reverse process, one needs to learn the unknown score functions $\boldsymbol{\nabla} \log P_{\theta^\star, T-t}$ for $t \in [0, T]$. This is where denoising score matching (DSM) is utilized in the DSM–DDPM estimator Ho et al. (2020); Song et al. (2021a). We start with the objective

$$\operatorname*{arg\,min}_{\theta \in \Theta} \int_0^T \| \boldsymbol{\nabla} \log P_{\theta, t} - \boldsymbol{\nabla} \log P_{\theta^\star, t} \|^2_{L^2(P_{\theta^\star, t})} \, \mathrm{d}t \,. \tag{5}$$

This objective is not amenable to empirical risk minimization since it involves the unknown score function $\boldsymbol{\nabla} \log P_{\theta^\star, t}$. Instead, we note that for fixed $t \in (0, T]$,

$$\begin{aligned} &\int \| \boldsymbol{\nabla} \log P_{\theta, t} - \boldsymbol{\nabla} \log P_{\theta^\star, t} \|^2 \, \mathrm{d}P_{\theta^\star, t} \\ &\qquad = \int \| \boldsymbol{\nabla} \log P_{\theta, t} \|^2 \, \mathrm{d}P_{\theta^\star, t} - 2 \int \langle \boldsymbol{\nabla} \log P_{\theta, t}, \boldsymbol{\nabla} \log P_{\theta^\star, t} \rangle \, \mathrm{d}P_{\theta^\star, t} + \text{CONST.} \end{aligned}$$

In this derivation, CONST. refers to any term that does not depend on the optimization variable θ. Continuing, the second term above can be written

$$\begin{aligned}
&-2\int \langle \nabla \log P_{\theta,t}, \nabla \log P_{\theta^\star,t}\rangle \,\mathrm{d}P_{\theta^\star,t} = -2\int \langle \nabla \log P_{\theta,t}(x_t), \nabla P_{\theta^\star,t}(x_t)\rangle \,\mathrm{d}x_t \\
&\quad = -2\int \Big\langle \nabla \log P_{\theta,t}(x_t), \nabla_{x_t} \int Q_{t|0}(x_t \mid x_0)\, P_{\theta^\star}(\mathrm{d}x_0)\Big\rangle \,\mathrm{d}x_t \\
&\quad = -2\iint \langle \nabla \log P_{\theta,t}(x_t), \nabla Q_{t|0}(x_t \mid x_0)\rangle \, P_{\theta_\star}(\mathrm{d}x_0)\,\mathrm{d}x_t \\
&\quad = -2\iint \langle \nabla \log P_{\theta,t}(x_t), \nabla \log Q_{t|0}(x_t \mid x_0)\rangle \, Q_{t|0}(\mathrm{d}x_t \mid x_0)\, P_{\theta_\star}(\mathrm{d}x_0)\,.
\end{aligned}$$

We deduce that the original problem (5) is equivalent to

$$\arg\min_{\theta\in\Theta} \int \ell^{\mathrm{DDPM}}(\theta; x_0)\, P_{\theta^\star}(\mathrm{d}x_0)\,,$$

where

$$\ell^{\mathrm{DDPM}}(\theta; x_0) \coloneqq \int_0^T \int \big\{\|\nabla \log P_{\theta,t}\|^2 - 2\,\langle \nabla \log P_{\theta,t}, \nabla \log Q_{t|0}(\cdot \mid x_0)\rangle\big\}\,\mathrm{d}Q_{t|0}(\cdot \mid x_0)\,\mathrm{d}t\,.$$

This is now amenable to empirical risk minimization, and hence we define the empirical version

$$\widehat{\mathcal{R}}_n^{\mathrm{DDPM}}(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^n \ell^{\mathrm{DDPM}}(\theta; x_0^{(i)})$$

where $x_0^{(1)}, \ldots, x_0^{(n)}$ are samples.

Finally, to see that this is equivalent to Definition 1, we note that by well-known properties of the OU semigroup, if $z_t \sim \mathcal{N}(0, I)$ is independent of x_0 and $x_t \coloneqq \exp(-t)\, x_0 + \sqrt{1-\exp(-2t)}\, z_t$, then $x_t \sim Q_{t|0}(\cdot \mid x_0)$. Then, we can write

$$\nabla \log Q_{t|0}(x_t \mid x_0) = -\frac{x_t - \exp(-t)\, x_0}{1-\exp(-2t)} = -\frac{z_t}{\sqrt{1-\exp(-2t)}}\,.$$

C Further Related Work

In this section, we present further related work.

Statistical Guarantees for Diffusion Models for Learning. Recent works have established rigorous statistical guarantees for diffusion models and related score matching estimators. For example, Oko et al. (2023) bound the estimation error when using a neural network and demonstrated that diffusion models are nearly minimax-optimal estimators in both the total variation and the Wasserstein distance of order one, provided that the target density belongs to the Besov space. Cui et al. (2024) employ a two-layer neural network to learn score functions and, in the special case where the target distribution is a mixture of two Gaussians, they establish an error guarantee of $\Theta(1/n)$ for the estimated mean. Further, Wibisono et al. (2024) consider subgaussian densities with Lipschitz-continuous score functions and provide optimal rates for estimating the scores in the L_2-norm. In a related direction, Koehler & Vuong (2024) show that pseudolikelihood methods can be used to learn low-rank Ising models, which is an example of using score matching for designing learning methods with provable statistical guarantees. Also, Mei & Wu (2025) study the statistical efficiency of neural networks to approximate score functions focusing on the setting of graphical models and variational inference algorithms. Moreover, Dou et al. (2024) study the score matching (SM) estimator in detail and establish the sharp minimax rate of score estimation for smooth, compactly supported densities using sophisticated techniques. In contrast to these works, which focus on the vanilla SM estimator and variants, we investigate the statistical properties of the DDPM estimator and, hence, use a very different set of tools.

Computational Properties of the DDPM Estimator. Beyond the immense practical success of DDPM estimators and the growing interest from statisticians, surprisingly, DDPM estimators are also

leading to new *provably* efficient algorithms for sampling and distribution learning Shah et al. (2023); Chen et al. (2024); **?**.

Sampling Guarantees for Diffusion Models. Finally, a rapidly growing body of work establishes sampling guarantees for SGMs under the assumption that the score functions are accurately estimated. Since the literature is vast and orthogonal to the statistical concerns in this paper, we do not survey it here.