5th Symposium on Advances in Approximate Bayesian Inference, 2023 1–20

Expressiveness Remarks for Denoising Diffusion Models and Samplers

Francisco Vargas FAV25@CAM.AC.UK
Teodora Reu TR500@CAM.AC.UK
Anna Kerekes AK2229@CAM.AC.UK
Department of Computer Science, Cambridge University, Cambridge, CB3 0FD, UK.

Abstract

Denoising diffusion models are a class of generative models which have recently achieved state-of-the-art results across many domains. Gradual noise is added to the data using a diffusion process, which transforms the data distribution into a Gaussian. Samples from the generative model are then obtained by simulating an approximation of the time reversal of this diffusion initialized by Gaussian samples. Recent research has explored adapting diffusion models for sampling and inference tasks. In this paper, we leverage known connections to stochastic control akin to the Föllmer drift to extend established neural network approximation results for the Föllmer drift to denoising diffusion models and samplers.

1. Introduction

Let π be a probability density on $\mathbb{R}^d$ of the form

$$\pi(x) = \frac{\gamma(x)}{Z}, \qquad Z = \int_{\mathbb{R}^d} \gamma(x)\mathrm{d}x, \tag{1}$$

where $\gamma : \mathbb{R}^d \to \mathbb{R}^+$ can be evaluated pointwise but the normalizing constant Z is intractable. We are interested in obtaining approximate samples from π. A variety of Monte Carlo techniques has been developed to tackle this problem. Variational techniques are a popular alternative to Markov Chain Monte Carlo (MCMC) (Neal, 2011) where one considers a flexible family of easy-to-sample distributions q^θ whose parameters are optimized by minimizing a suitable cost, such as reverse Kullback–Leibler discrepancy $\mathrm{KL}(q^\theta||\pi)$.

Instead recent work (Vargas et al., 2023) tackles the sampling problem with a new class of samplers coined denoising diffusion samplers (DDS) that leverage Denoising Diffusion Probabilistic Models (DDPM), a powerful class of generative models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021b) to sample from unnormalised densities. In this context, one adds noise progressively to data using diffusion to transform the complex target distribution into a Gaussian distribution. The time-reversal of this diffusion can then be used to transform a Gaussian sample into a sample from the target.

In this work, we will explore in more detail the connection to stochastic control remarked between denoising diffusion models (Ho et al., 2020; Song et al., 2021b) in Vargas et al. (2023) and leverage this to show how the score of VP-SDEs (Song et al., 2021b) can be approximated with neural networks up to an arbitrarily small error and we quantify the induced sampling error. We do this by extending the theoretical results derived in Tzen and Raginsky (2019b).

Our contributions in this paper can be summarized as follows: (1) establishing a connection between the VP-SDE score and OU-semigroup (Section 3.1), (2) exploring novel regularity properties for OU-semigroup (Section 3.2), and (3) demonstrating neural network and sampling approximation results for a simplified VP-SDE (Proposition 1, Remark 2).

2. Background - Denoising Diffusion Models and Samplers

For the purpose of this work, we will introduce Denoising Diffusions and DDS in continuous time. Let $\mathcal{C} = C([0,T], \mathbb{R}^d)$ be the space of continuous functions from $[0,T]$ to $\mathbb{R}^d$ and $\mathcal{B}(\mathcal{C})$ the Borel sets on $\mathcal{C}$. We consider path measures, which are probability measures on $(\mathcal{C}, \mathcal{B}(\mathcal{C}))$ (Léonard, 2014). For synergy with the results in Tzen and Raginsky (2019b) we will introduce DDS (Vargas et al., 2023) with the time reversals flipped, meaning we interchange the backward and the forward processes compared to (Vargas et al., 2023; Ho et al., 2020; Song et al., 2021b; De Bortoli et al., 2021a).

2.1. Backwards diffusion and its time-reversal

Consider the forward noising diffusion given by a time-reversed Ornstein–Uhlenbeck (OU) process (Song et al. (2021b) refer to this SDE as the VP-SDE).

$$\mathrm{d}x_t = -\beta_t x_t \mathrm{d}t + \sigma\sqrt{2\beta_t}\mathrm{d}B_t, \qquad x_0 \sim \pi, \tag{2}$$

where $(B_t)_{t\in[0,T]}$ is a d-dimension Brownian motion and $t \to \beta_t$ is a non-decreasing positive function. This diffusion induces the path-measure $\mathcal{P}$ on the time interval $[0,T]$ and the marginal density of x_t is denoted p_t. The transition density of this diffusion is given by $p_{t|0}(x_t|x_0) = \mathcal{N}(x_t; \sqrt{1-\lambda_t}x_0, \sigma^2\lambda_t I)$, where $\lambda_t = 1 - \exp(-2\int_0^t \beta_s \mathrm{d}s)$. We will always consider a scenario where $\int_0^T \beta_s \mathrm{d}s \gg 1$ so that $p_T(x_T) \approx \mathcal{N}(x_T; 0, \sigma^2 I)$.

From (Haussmann and Pardoux, 1986), its time-reversal $(y_t)_{t\in[0,T]} = (x_{T-t})_{t\in[0,T]}$, where equality is here in distribution, yields the forward time diffusion:

$$\mathrm{d}y_t = \beta_{T-t}\{y_t + 2\sigma^2\nabla \ln p_{T-t}(y_t)\}\mathrm{d}t + \sigma\sqrt{2\beta_{T-t}}\mathrm{d}W_t, \qquad y_0 \sim p_T, \tag{3}$$

where $(W_t)_{t\in[0,T]}$ is another d-dimensional Brownian motion. By definition this time-reversal starts from $y_0 \sim p_T(y_0) \approx \mathcal{N}(y_0; 0, \sigma^2 I)$ and is such that $y_T \sim \pi$. This suggests that approximate simulation of diffusion (3) would result in approximate samples from π. However, putting this idea into practice requires being able to approximate the intractable scores $(\nabla \ln p_t(x))_{t\in[0,T]}$. Unlike DDPM, score matching techniques are not feasible, as sampling from (2) requires sampling $x_0 \sim \pi$, which is impossible by assumption.

2.2. Reference diffusion and value function

In our context, it is useful to introduce a *reference* process defined by the diffusion following (2), but initialized at $p_0^{\mathrm{ref}}(x_0) = \mathcal{N}(x_0; 0, \sigma^2 I)$ rather than $\pi(x_0)$ thus ensuring that the marginals of the resulting path measure $\mathcal{P}^{\mathrm{ref}}$ all satisfy $p_t^{\mathrm{ref}}(x_t) = \mathcal{N}(x_t; 0, \sigma^2 I)$. Following Vargas et al. (2023) we can identify $\mathcal{P}$ as the path measure minimizing the half bridge $\mathcal{P} = \arg\min_{\mathcal{Q}}\{\mathrm{KL}(\mathcal{Q}||\mathcal{P}^{\mathrm{ref}}) : q_T = \pi\}$ (Bernton et al., 2019; Vargas et al., 2021b; De Bortoli et al., 2021b). where representation of $\mathcal{P}^{\mathrm{ref}}$ is given by

$$\mathrm{d}y_t = -\beta_{T-t} y_t \mathrm{d}t + \sigma\sqrt{2\beta_{T-t}}\mathrm{d}W_t, \qquad y_0 \sim p_0^{\mathrm{ref}}. \tag{4}$$

As $\beta_{T-t}y_t + 2\sigma^2\nabla \ln p^{\text{ref}}_{T-t}(y_t) = -\beta_{T-t}y_t$, we can rewrite the time-reversal (3) of $\mathcal{P}$ as

$$\mathrm{d}y_t = -\beta_{T-t}\{y_t - 2\sigma^2\nabla \ln \phi_{T-t}(y_t)\}\mathrm{d}t + \sigma\sqrt{2\beta_{T-t}}\mathrm{d}W_t, \qquad y_0 \sim p_T, \tag{5}$$

where $v_t(x) = -\ln \phi_t(x) = -\ln p_t(x)/p^{\text{ref}}_t(x)$ is known as the value function (Fleming and Rishel, 2012; Pham, 2009; Nüsken and Richter, 2021; Tzen and Raginsky, 2019b).

2.3. Learning the Forward Diffusion - Reverse KL Formulation

To approximate (3) $\mathcal{P}$, consider a path measure $\mathcal{Q}^\theta$ which is induced by

$$\mathrm{d}y_t = \beta_{T-t}\{y_t + 2\sigma^2 s_\theta(T-t, y_t)\}\mathrm{d}t + \sigma\sqrt{2\beta_{T-t}}\mathrm{d}W_t, \qquad y_0 \sim \mathcal{N}(0, \sigma^2 I), \tag{6}$$

so that $y_t \sim q^\theta_t$. To obtain $s_\theta(t,x) \approx \nabla \ln p_t(x)$, we parameterize $s_\theta(t,x)$ by a neural network whose parameters are obtained by minimizing

$$\mathrm{KL}(\mathcal{Q}^\theta||\mathcal{P}) = \mathrm{KL}(\mathcal{N}(0,\sigma^2 I)||p_T) + \sigma^2\mathbb{E}_{\mathcal{Q}^\theta}\Big[\int_0^T \beta_{T-t}||s_\theta(T-t,y_t) - \nabla \ln p_{T-t}(y_t)||^2\mathrm{d}t\Big],$$

This expression closely resembles the expression obtained in (Song et al., 2021a, Theorem 1) in the context of DDPM. However unlike DDPM Ho et al. (2020), one cannot get rid of the intractable scores $(\nabla \ln p_t(x))_{t\in[0,T]}$ using score matching ideas. Instead, using (5), Vargas et al. (2023) reparameterize $\mathcal{Q}^\theta$ using.

$$\mathrm{d}y_t = -\beta_{T-t}\{y_t - 2\sigma^2 f_\theta(T-t, y_t)\}\mathrm{d}t + \sigma\sqrt{2\beta_{T-t}}\mathrm{d}W_t, \qquad y_0 \sim \mathcal{N}(0, \sigma^2 I), \tag{7}$$

unlike (6) f_θ approximates $\nabla \ln \phi_t$ rather than the score $\nabla \ln p_t$. Then under this reparameterization Vargas et al. (2023) use standard results on half bridges (Bernton et al., 2019) to express $\mathrm{KL}(\mathcal{Q}^\theta||\mathcal{P})$ in compact form:

$$\mathrm{KL}(\mathcal{Q}^\theta||\mathcal{P}) = \mathbb{E}_{\mathcal{Q}^\theta}\Big[\sigma^2\int_0^T \beta_{T-t}||f_\theta(T-t,y_t)||^2\mathrm{d}t + \ln\Big(\frac{\mathcal{N}(y_T;0,\sigma^2 I)}{\pi(y_T)}\Big)\Big]. \tag{8}$$

Where $q^{\theta^*}_0 = p_T \approx \mathcal{N}(0,\sigma^2 I)$. Then θ minimizing (8), approximate samples from π can be obtained by simulating (7) and returning $y_T \sim q^\theta_T$. Note concurrent work (Berner et al., 2022) also optimises an equivalent reverse KL to (8).

3. Expressiveness and Regularity Results

In this section, we present our main result. We demonstrate that $\nabla \ln \phi_t$ and thus the score of the OU-SDE can be approximated by a multi-layer neural network efficiently.

Theorem 3.1 in Tzen and Raginsky (2019b) provides neural network approximation and sampling guarantees for a different class of SDEs than DDPM (i.e. (3) or (5)). Thus in this section, we will adapt such results to denoising diffusion samplers (Vargas et al., 2023) and via directly relating the approximations to the score of the VP-SDE (2) we motivate how these results extend to DDPM based methods (Song et al., 2021b; Ho et al., 2020; Huang et al., 2021a).

Tzen and Raginsky (2019b) guarantee approximate sampling from a target distribution using a multilayer feedforward neural net drift, assuming the smoothness, Lipschitzness, and boundedness of $f(x) = \frac{d\pi}{d\mathcal{N}(0,\sigma^2 I)}(x)$, (Assumption 2), as well as the smoothness of the activations (Assumption 3) and uniform approximability of f and its gradient by a neural network (Assumption 4). In the following proposition and remark we present our adaption of their results to DDS.

Proposition 1 *Suppose Assumptions in Appendix B are in force. Let L denote the maximum of the Lipschitz constants of f and ∇f. Then for all $0 < \epsilon < 16L^2/c^2$, there exists a neural net $\hat{v} : R^d \times [0,1] \to R^d$ with size polynomial in $1/\epsilon, d, L, c, 1/c$ such that the activation function of each neuron in the set of $\{\sigma, \sigma', ReLU\}$, and the following hold: If $\{\hat{x}_t\}_{t\in[0,1]}$ is the diffusion process governed by the Itô SDE:*

$$d\hat{x}_t = \hat{b}(\hat{x}_t, t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t \tag{9}$$

with $x_0 \sim p_1 \approx \mathcal{N}(0, I)$ with the drift $\hat{b}(x,t) = -(x - 2\hat{v}(x, 1-t))$, then $\hat{\mu} := \mathrm{Law}(\hat{x}_1)$, satisfies $D(\mu||\hat{\mu}) \leq \epsilon$.

Remark 2 *Assuming π satisfies a logarithmic Sobolev inequality, extending the time domain to $t \in [0,T]$ and sampling $\hat{x}_0 \sim \mathcal{N}(0,I)$ approximately, it follows that $D(\mu||\hat{\mu}) \leq e^{-T}\mathrm{KL}(\pi||\mathcal{N}(0,1)) + T\epsilon$.*

The proof will closely follow Tzen and Raginsky (2019b) however key steps must be slightly modified to show that the value function satisfies the required regularity properties to exploit the core results in Tzen and Raginsky (2019a).

3.1. OU Semigroup and Time Reversal

This section introduces the OU semigroup (Metafune et al., 2002) whose logarithmic gradient can be directly connected to the score (Song et al., 2021b) in (2). Based on this reformulation of the score we are able to extend the results from Tzen and Raginsky (2019b) to denoising diffusion via VP-SDEs. In the remainder of this section, we will introduce new results pertaining to the regularity properties of this operator that will enable us to prove Proposition 1.

Definition 3 *We define the VP-SDE semigroup as,*

$$U_t^{\beta_t} f(y) = \mathbb{E}_{Z\sim\mathcal{N}(0,I)}\left[f\left(e^{-\int_0^t \beta_s \mathrm{d}s}y + \sigma(1 - e^{-2\int_0^t \beta_s \mathrm{d}s})^{1/2} Z\right)\right] \tag{10}$$

Then the OU-semigroup (Metafune et al., 2002) (typically defined with $\beta_t = \beta = 1$) is a simpler instance of the above.

$$U_t^{\beta} f(y) = \mathbb{E}_{Z\sim\mathcal{N}(0,I)}\left[f\left(e^{-\beta t}y + \sigma(1 - e^{-2\beta t})^{1/2} Z\right)\right] \tag{11}$$

For the purpose of simplicity we will be working with the OU semi-group when $\beta = 1$ (denoted U_t), however, these results can be extended to the more general case. In the following remark, we highlight the connection between the OU semi-group, the value function and the score in DDPM.

Remark 4 *The time reversal of the VP-SDE (i.e. $b^*(y,t) = -\beta_{T-t}(y - 2\sigma^2\nabla \ln \phi_{T-t}(y))$) can be expressed in terms of the OU semigroup via:*

$$\nabla \ln \phi_{T-t}(y) = \nabla_y \ln U^{\beta_t}_{T-t} f(y), \tag{12}$$

When $f(x) = \frac{\mathrm{d}\pi}{\mathrm{d}\mathcal{N}(0,\sigma^2 I)}(x)$. This in turn can be related to the score

$$\nabla \ln p_{T-t}(y) = -\left(\frac{y}{2\sigma^2} - \nabla \ln \phi_{T-t}(y)\right) = -\left(\frac{y}{2\sigma^2} - \nabla_y \ln U^{\beta_t}_{T-t} f(y)\right). \tag{13}$$

From this stage on we consider the case where $\sigma = \beta = 1$. Notice how the formulation in Remark 4 is reminiscent of the Föllmer drift (Föllmer, 1984; Dai Pra, 1991; Tzen and Raginsky, 2019b; Huang et al., 2021b). Finally, we highlight that it is this very simple remark which facilitates porting over the theoretical results and insights from (Tzen and Raginsky, 2019b) to diffusion-based models. Furthermore we remind the reader that the results in Tzen and Raginsky (2019b) require adapting as they apply to the Föllmer drift and the heat semigroup (i.e.$\nabla_y \ln \phi_t(y) = \nabla_y \ln Q_t f(y)$ with $Q_t f(y) = \mathbb{E}_{Z\sim\mathcal{N}(0,I)}\left[f\left(y + \sqrt{t}Z\right)\right]$).

3.2. Regularity Properties

In this section, we will prove regularity properties pertaining to the OU semigroup which will allow us to extend the theoretical guarantees in Tzen and Raginsky (2019b) to denoising diffusion models and samplers (Song et al., 2021b; Ho et al., 2020; Vargas et al., 2023). Moving forward we prove a basic auxiliary result regarding the commutativity of the OU-semigroup with partial derivatives. From this result, by using Corollary 12, we were able to bound the OU-semigroup norm when differentiated.

Lemma 5 *OU semigroup is commutative with the gradient operator that is for $f : \mathbb{R}^d \to \mathbb{R}$ we have $\partial_{y_i} U_t f(y) = U_t \partial_{y_i} f(y)$.*

3.2.1. Terminal Cost

This section derives the regularity properties of $g_{x,t}(z) = g(e^{-t}x + (1 - e^{-2t})^{1/2}z)$, which we will refer to as the terminal cost. We want to underline to the reader that the optimal drift can be expressed in terms of the OU-semigroup is applied to the terminal cost ($\nabla \ln \phi_t(x) = \nabla \ln U_t g_{x,t}(z)$) when $g = f$.

- First we prove that a centred version of the terminal cost is $\mathscr{L}^2(Q)$ Lipchitz with respect to a newly defined metric. This will allow us to obtain a bound for the covering number of a function class induced by the terminal cost.
- We then derive an envelope for the terminal cost. This in conjunction with further results on covering numbers allows us to control Dudley's entropy integral (Dudley, 1967). This in turn provides us with results from empirical process theory (Giné and Nickl, 2021) that quantify the error for an empirical estimate of the OU semigroup.

Lemma 6 *($\mathscr{L}^2$ Lipchitz condition) Let $\bar{g}_{t,x}(z) = g(e^{-t}x + (1 - e^{-2t})^{1/2}z) - g(0)$ then it follows that:*

$$||\bar{g}_{t,x}(z) - \bar{g}_{t',x'}(z)||_{\mathscr{L}^2(Q)} \leq L\left(1 + \sqrt{2}||z||_{\mathscr{L}^2(Q)}\right)\rho_{OU}((t,x),(t',x'))$$

such that $\rho_{OU}((t,x),(t',x')) = ||e^{-t}x - x'e^{-t'}|| + |t - t|^{1/2}$.

Lemma 7 *Let $g : R^d \to R$ to L-Lipschitz with respect to the Euclidean norm. Then for $F(z) := L((R \vee 1) + \sqrt{2}||z||)$.*

$$\left| g\left(e^{-t}x + (1 - e^{-2t})^{1/2} z\right) - g(0) \right| \leq F(z) \tag{14}$$

3.2.2. Covering Number

The $\mathscr{L}^2(Q)$ covering number of the function space $\mathcal{G}$ is defined by:

$$N\left(\mathcal{G}, \mathscr{L}^2(Q), \varepsilon\right) := \min \left\{ K : \exists f_1, \ldots, \exists f_K \in \mathscr{L}^2(Q) \text{ s.t. } \sup_{g \in \mathcal{G}} \min_{k \leq K} \|g - f_k\|_{L^2(P)} \leq \varepsilon \right\},$$

in general the covering number $N(\mathcal{A}, \rho, \varepsilon)$ is the smallest number of balls of size ϵ wrt to the metric ρ that cover the set $\mathcal{A}$. Once we obtain the appropriate bound on $N\left(\mathcal{G}, \mathscr{L}^2(Q), \varepsilon\right)$ the results from (Tzen and Raginsky, 2019b) follow with minor modifications and thus Proposition 1 will follow. In this section we will be bounding the $\mathscr{L}^2(Q)$ covering number of the function space $\mathcal{G} := \left\{\bar{g}_{x,t} : x \in \mathrm{B}^d(R), t \in [0, 1]\right\}$.

Lemma 8 *Given the metric space $\left([0, T] \times B^d(R), \rho_{OU}\right)$ where:*

$$\rho_{OU}((t, x), (t', x')) = ||e^{-t}x - x'e^{-t'}|| + |t - t'|^{1/2} \tag{15}$$

and $||(t, x)||_{OU} = \rho_{OU}((t, x), (0, 0)) = ||e^{-t}x|| + |t|^{1/2}$. It follows that:

$$N(\mathcal{G}, \mathscr{L}^2(Q), \epsilon ||F||_{\mathscr{L}^2(Q)}) \leq N([0, T] \times B^d(R), \rho_{OU}, \epsilon) \tag{16}$$

Lemma 9 *Given the metric space $\left([0, T] \times B^d(R), \rho_{OU}\right)$ it follows that*

$$N([0, T] \times B^d(R), \rho_{OU}, \epsilon) \leq N([0, T], |\cdot|, \epsilon^2/4) N(B^d(R), ||\cdot||, \epsilon/2). \tag{17}$$

From Lemmas 8, 9 it follows that :

$$N(\mathcal{G}, \mathscr{L}^2(Q), \epsilon ||F||_{\mathscr{L}^2(Q)}) \leq N([0, T], |\cdot|, \epsilon^2/4) N(B^d(R), ||\cdot||, \epsilon/2) \tag{18}$$

and thus it follows that Lemmas C.4 and thus Theorem C.1 in Tzen and Raginsky (2019b) hold true in our setting. This provides us with the tools required to show the neural network approximation results (see Appendix E) which in turn enable our main result Proposition 1.

4. Conclusion

We establish a connection between the VP-SDE score and the OU-semigroup, revealing similarities between Föllmer drift-based and DDPM-based sampling approaches. Using this connection, we demonstrate how the VP-SDE score can be approximated efficiently by multilayer neural networks, under fairly general assumptions on the target distribution. In order to exploit previous results on the Föllmer drift (Tzen and Raginsky, 2019b) we establish novel regularity properties for the OU-semigroup that allow us to adapt the results in Tzen and Raginsky (2019b) to our setting. Although our results are derived for target distributions represented as densities, rather than empirical distributions as in DDPM, however, these results may apply in the large sample limit, given suitable assumptions.

References

Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Julius Berner, Lorenz Richter, and Karen Ullrich. An optimal control perspective on diffusion-based generative modeling. *arXiv preprint arXiv:2211.01364*, 2022.

Espen Bernton, Jeremy Heng, Arnaud Doucet, and Pierre E Jacob. Schrödinger bridge samplers. *arXiv preprint arXiv:1912.13170*, 2019.

Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected langevin monte carlo. *Discrete & Computational Geometry*, 59: 757–783, 2018.

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. *arXiv preprint arXiv:2209.11215*, 2022.

Paolo Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. *arXiv preprint arXiv:2106.01357*, 2021a.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*, 2021b.

Richard M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Wendell H Fleming and Raymond W Rishel. *Deterministic and stochastic optimal control*, volume 1. Springer Science & Business Media, 2012.

Hans Föllmer. An entropy approach to the time reversal of diffusion processes. *Lecture Notes in Control and Information Sciences*, 69:156–163, 1984.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

Ulrich G Haussmann and Etienne Pardoux. Time reversal of diffusions. *The Annals of Probability*, 14(3):1188–1205, 1986.

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

Chin-Wei Huang, Jae Hyun Lim, and Aaron Courville. A variational perspective on diffusion-based generative models and score matching. *arXiv preprint arXiv:2106.02808*, 2021a.

Jian Huang, Yuling Jiao, Lican Kang, Xu Liao, Jin Liu, and Yanyan Liu. Schrödinger-Föllmer sampler: Sampling without ergodicity. *arXiv preprint arXiv:2106.10880*, 2021b.

Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems-Series A*, 34(4):1533–1574, 2014.

Giorgio Metafune, Diego Pallara, and Enrico Priola. Spectrum of ornstein-uhlenbeck operators in lp spaces with respect to invariant measures. *Journal of Functional Analysis*, 196(1):40–60, 2002.

Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011.

Nikolas Nüsken and Lorenz Richter. Solving high-dimensional Hamilton–Jacobi–Bellman PDEs using neural networks: perspectives from the theory of controlled diffusions and measures on path space. *Partial Differential Equations and Applications*, 2(4):1–48, 2021.

Huyên Pham. *Continuous-time stochastic control and optimization with financial applications*, volume 61. Springer Science & Business Media, 2009.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

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

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*, 2021b.

Belinda Tzen and Maxim Raginsky. Neural stochastic differential equations: Deep latent Gaussian models in the diffusion limit. *arXiv preprint arXiv:1905.09883*, 2019a.

Belinda Tzen and Maxim Raginsky. Theoretical guarantees for sampling and inference in generative models with latent diffusions. In *Conference on Learning Theory*, 2019b.

Francisco Vargas, Andrius Ovsianas, David Fernandes, Mark Girolami, Neil Lawrence, and Nikolas Nüsken. Bayesian learning via neural Schrödinger–Föllmer flows. *arXiv preprint arXiv:2111.10510*, 2021a.

Francisco Vargas, Pierre Thodoroff, Austen Lamacraft, and Neil Lawrence. Solving Schrödinger bridges via maximum likelihood. *Entropy*, 23(9), 2021b. ISSN 1099-4300. doi: 10.3390/e23091134. URL https://www.mdpi.com/1099-4300/23/9/1134.

Francisco Vargas, Will Sussman Grathwohl, and Arnaud Doucet. Denoising diffusion samplers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=8pvnfTAbu1f.

Appendix A. List of Detailed Contributions

Our contributions are.

- Our overall contribution is porting over the expressiveness results from Tzen and Raginsky (2019b) to denoising diffusion-based models in the setting where the target distribution admits a density.
- To facilitate this connection we provide Remark 4 which expresses the score in DDPM in terms of the well-known OU semigroup. This expression is more akin to the Föllmer drift and thus motivates the connection to the results in (Tzen and Raginsky, 2019b).
- In order to do this we prove 3 novel results, specifically Lemmas 5-9 and Remark 10 and highlight how they allow us to use Theorem C.1 of Tzen and Raginsky (2019b).
- For completeness we provide the adapted sketches for Theorem 3.2 of Tzen and Raginsky (2019b) and our Proposition 1 where we highlight the differences to (Tzen and Raginsky, 2019b) in magenta.
- Finally we provide Remark 2 which quantifies the error from initialising $\hat{x}_0$ at $\mathcal{N}(0, I)$ rather than p_T. The result combines the derived expressiveness/score approximation error with the mixing error of the OU-process.

A.1. Limitations

We would like to highlight that all our results are in continuous time, and additional work would be required to analyse them under a given discretisation (e.g. Euler Maruyama). It is possible with additional assumptions to apply results directly such as Theorem 2 in Chen et al. (2022) or it may be possible to adapt Corollary 2 in Vargas et al. (2021a).

Appendix B. Assumptions

Assumption 1 *Throughout all this work we assume that the target distribution π has a density that is it is absolutely continuous wrt to the Lebesgue measure on $\mathbb{R}^d$.*

Assumption 2 *The function f is differentiable, both f and ∇f are L-Lipschitz, and there exists a constant $c \in (0, 1]$ such that $f \geq c$ everywhere.*

Assumption 3 *The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is differentiable. Moreover, there exists $c_\sigma > 0$ depending only on σ, such that the following holds: For any L-Lipschitz function $h : \mathbb{R} \to \mathbb{R}$ which is constant outside the interval $[-R, R]$ and for any $\delta > 0$, there exist real numbers a, $\{\alpha_i, \beta_i, \gamma_i\}_{i=1}^m$ where $m \leq c_\sigma \frac{RL}{\delta}$, such that the function $\tilde{h}(x) = a + \sum \alpha_i \sigma(\beta_i x + \gamma_i)$ satisfies $\sup_{x \in \mathbb{R}} |\tilde{h}(x) - h(x)| \leq \delta$.*

Finally as per Tzen and Raginsky (2019b) we introduce the assumption pertaining to the approximability of f by neural nets. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a fixed nonlinearity. Given a vector $w \in \mathbb{R}^n$ and scalars α, β, define the function

$$N^\sigma_{w,\alpha,\beta} : \mathbb{R}^n \to \mathbb{R}, \quad N^\sigma_{w,\alpha,\beta}(x) := \alpha \cdot \sigma\left(w^T x + \beta\right).$$

For $\ell \geq 2$, we define the class $\mathcal{N}_\ell^\sigma$ of ℓ-layer feedforward neural nets with activation function σ recursively as follows: $\mathcal{N}_2^\sigma$ consists of all functions of the form $x \mapsto \sum_{i=1}^m N_{w_i,\alpha_i,\beta_i}^\sigma(x)$ for all $m \in \mathbb{N}, w_1, \ldots, w_m \in \mathbb{R}^d$, $\alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_m \in \mathbb{R}$, and, for each $\ell \geq 2$,

$$\mathcal{N}_{\ell+1}^\sigma := \bigcup_{k\geq 1} \bigcup_{m \geq 1} \Bigg\{ x \mapsto \sum_{i=1}^m N_{w_i,\alpha_i,\beta_i}^\sigma \left(h_1(x), \ldots, h_k(x)\right) : \\ \alpha_1, \ldots, \alpha_m, \beta_1, \ldots, \beta_m \in \mathbb{R}, w_1, \ldots, w_m \in \mathbb{R}^k, h_1, \ldots, h_k \in \mathcal{N}_\ell^\sigma \Bigg\}.$$

Assumption 4 *For any $R > 0$ and $\epsilon > 0$, there exist a neural net $\hat{f} \in \mathcal{N}_{l,s}^\sigma$ with $l, s <$ poly$(1/\epsilon, d, L, R)$, such that*

$$\sup_{x \in B^d(R)} |f(x) - \hat{f}(x)| \leq \epsilon \quad \textit{and} \quad \sup_{x \in B^d(R)} ||\nabla f(x) - \nabla \hat{f}(x)|| \leq \epsilon. \tag{19}$$

Appendix C. Regularity Results

Remark 4 *The time reversal of the VP-SDE (i.e. $b^*(y,t) = -\beta_{T-t}(y - 2\sigma^2 \nabla \ln \phi_{T-t}(y))$) can be expressed in terms of the OU semigroup via:*

$$\nabla \ln \phi_{T-t}(y) = \nabla_y \ln U_{T-t}^{\beta_t} f(y), \tag{20}$$

When $f(x) = \frac{\pi}{\mathcal{N}(0,\sigma^2 I)}(x)$. This in turn can be related to the score

$$\nabla \ln p_{T-t}(y) = -\left(\frac{y}{2\sigma^2} - \nabla \ln \phi_{T-t}(y)\right) = -\left(\frac{y}{2\sigma^2} - \nabla_y \ln U_{T-t}^{\beta_t} f(y)\right). \tag{21}$$

Proof Consider the OU semigroup evaluated on the appropriate RND:

$$\begin{aligned} U_t^{\beta_t} f(y) &= \mathbb{E}_{Z \sim \mathcal{N}(0,I)} \left[\frac{\pi}{\mathcal{N}(0,\sigma^2 I)} \left(e^{-\beta t} y + \sigma (1 - e^{-2\beta t})^{1/2} Z \right) \right] \\ &= \mathbb{E}_{x_T \sim p_{T|t}^{\mathrm{ref}}(\cdot|x)} \left[\frac{\pi}{\mathcal{N}(0,\sigma^2 I)} (x_T) \right] \\ &= \int p_{T|t}^{\mathrm{ref}}(x_T|x) \frac{\pi}{\mathcal{N}(0,\sigma^2 I)} (x_T) \, \mathrm{d}x_T \\ &= \int \frac{p_{t|T}^{\mathrm{ref}}(x|x_T) p_T^{\mathrm{ref}}(x_T)}{p_t^{\mathrm{ref}}(x)} \frac{\pi}{\mathcal{N}(0,\sigma^2 I)} (x_T) \, \mathrm{d}x_T \\ &= \int \frac{p_{t|T}^{\mathrm{ref}}(x|x_T)}{p_t^{\mathrm{ref}}(x)} \pi(x_T) \, \mathrm{d}x_T = \frac{p_t(x)}{p_t^{\mathrm{ref}}(x)} \end{aligned}$$

and thus it follows that

$$\nabla \ln p_{T-t}(y) = -\left(\frac{y}{2\sigma^2} - \nabla_y \ln U_{T-t}^{\beta_t} f(y)\right). \tag{22}$$

relating the score and the OU semi-group as required. ■

Lemma 5 *OU semigroup is commutative with the gradient operator that is for $f : \mathbb{R}^d \to \mathbb{R}$ we have $\partial_{y_i} U_t f(y) = U_t \partial_{y_i} f(y)$.*

Proof It suffices to show that

$$d(x,z) = \delta^{-1}(f(e^{-t}x + (1-e^{-2t})^{1/2}z) - f(e^{-t}(x+\delta \boldsymbol{e}_i) + (1-e^{-2t})^{1/2}z)), \tag{23}$$

is dominated, where $[\boldsymbol{e}_i]_j = \delta_{ij}$. As f is Lipchitz by assumption it follows that

$$|d(x,z)| \leq L|\delta^{-1}e^{-t}\delta| = Le^{-t} \leq L \tag{24}$$

As L is integrable under $\mathcal{N}(0,I)$ we have shown $d(x,z)$ is dominated for all δ and thus the partial derivative operator and the OU semigroup commute. ■

The choice of $F(z) := L((R \vee 1) + \sqrt{2}||z||)$ with these specific constants arises from the following result.

Lemma 6 *($\mathscr{L}^2$ Lipchitz condition) Let $\bar{g}_{t,x}(z) = g(e^{-t}x + (1-e^{-2t})^{1/2}z) - g(0)$ then it follows that:*

$$||\bar{g}_{t,x}(z) - \bar{g}_{t',x'}(z)||_{\mathscr{L}^2(Q)} \leq L\left(1 + \sqrt{2}||z||_{\mathscr{L}^2(Q)}\right)\rho_{OU}((t,x),(t',x'))$$

such that:

$$\rho_{OU}((t,x),(t',x')) = ||e^t x - x' e^{t'}|| + |t-t'|^{1/2} \tag{25}$$

Proof

$$\begin{aligned}
||\bar{g}_{t,x}(z) - \bar{g}_{t',x'}(z)||_{\mathscr{L}^2(Q)} &\leq L||||e^{-t}x + (1-e^{-2t})^{1/2}z - e^{-t'}x' - (1-e^{-2t'})^{1/2}z||||_{\mathscr{L}^2(Q)} \\
&\leq L\left|\left| ||e^{-t}x - e^{-t'}x'|| + |(1-e^{-2t})^{1/2} - (1-e^{-2t'})^{1/2}| \cdot ||z|| \right|\right|_{\mathscr{L}^2(Q)} \\
&\leq L\left(||e^{-t}x - e^{-t'}x'|| + |(1-e^{-2t})^{1/2} - (1-e^{-2t'})^{1/2}| \cdot ||z||_{\mathscr{L}^2(Q)} \right) \\
&\leq L\left(||e^{-t}x - e^{-t'}x'|| + |e^{-2t} - e^{-2t'}|^{1/2} \cdot ||z||_{\mathscr{L}^2(Q)} \right) \\
&\leq L\left(||e^{-t}x - e^{-t'}x'|| + \sqrt{2}|t-t'|^{1/2} \cdot ||z||_{\mathscr{L}^2(Q)} \right)
\end{aligned}$$

Where in the last line we use that $\sup_{t\in[0,T]} |(e^{-2t})'| = 2$ and thus e^{-2t} is 2-Lipchitz. ■

Lemma 7 *Let $g : R^d \to R$ to L-Lipschitz with respect to the Euclidean norm. Then for $F(z) := L((R \vee 1) + \sqrt{2}||z||)$.*

$$\left| g\left(e^{-t}x + (1-e^{-2t})^{1/2}z\right) - g(0) \right| \leq F(z) \tag{26}$$

Proof By Lipschitz continuity for all $z \in R^d, X \in B^d(R), t \in [0,T]$ we have:

$$|g\left(e^{-t}x + (1-e^{-2t})^{1/2}z\right) - g(0)| \le L||e^{-t}x + (1-e^{-2t})^{1/2}z|| \tag{27}$$
$$\le L(e^{-t}||x|| + (1-e^{-2t})^{1/2}||z||) \tag{28}$$

Since both e^{-t} and $(1-e^{-2t})^{1/2}$ are strictly smaller than 1, we have:

$$L(e^{-t}||x|| + (1-e^{-2t})^{1/2}||z||) \le L(R + ||z||) \tag{29}$$
$$\le L((R \vee 1) + ||z||) \le F(z) \tag{30}$$

■

Appendix D. Covering Number Results

Remark 10 *The space* $([0,T] \times B^d(R), \rho_{OU})$ *is a metric space, where*

$$\rho_{OU}((t,x),(t',x')) = ||e^{-t}x - x'e^{-t'}|| + |t-t'|^{1/2}. \tag{31}$$

Proof

- **Positive definiteness**:

$$\rho_{OU}((t,x),(t',x')) = 0 \iff \tag{32}$$
$$||e^{-t}x - x'e^{-t'}|| + |t-t'|^{1/2} = 0 \iff \tag{33}$$
$$x = x' \text{ and } t = t. \tag{34}$$

  Since in (33) both terms are positive on the LHS, each has to be 0 to get the RHS, thus we get (34).

- **Symmetry**:

$$\rho_{OU}((t,x),(t',x')) = \rho_{OU}((t',x'),(t,x)). \tag{35}$$

- **Triangle inequality**: we show triangle inequality on $(t,x),(t',x')$ and (t'',x''). First let us note, that $||e^{-t}x - x'e^{-t'}|| + ||e^{-t'}x' - x''e^{-t''}|| \ge ||e^{-t}x - x''e^{-t''}||$, since $||\cdot||$ has the triangle inequality. Now:

$$|t-t'|^{1/2} + |t'-t''|^{1/2} \ge |t-t''|^{1/2} \iff \tag{36}$$
$$|t-t'| + 2|t-t'|^{1/2}|t'-t''|^{1/2} + |t'-t''| \ge |t-t''|. \tag{37}$$

  (37) is true, since $|\cdot|$ has the triangle inequality and $2|t-t'|^{1/2}|t'-t''|^{1/2} \ge 0$.

■

Lemma 8 *Given the metric space* $\big([0,T]\times B^d(R), \rho_{OU}\big)$ *where:*

$$\rho_{OU}((t,x),(t',x')) = ||e^{-t}x - x'e^{-t'}|| + |t-t'|^{1/2} \tag{38}$$

and

$$||(t,x)||_{OU} = \rho_{OU}((t,x),(0,0)) = ||e^{-t}x|| + |t|^{1/2} \tag{39}$$

It follows that:

$$N(\mathcal{G}, \mathscr{L}^2(Q), \epsilon||F||_{\mathscr{L}^2(Q)}) \leq N([0,T]\times B^d(R), \rho_{OU}, \epsilon) \tag{40}$$

Proof

Consider the ϵ-cover $A_{\rho_{OU}}$ with respect to ρ_{OU} of $[0,T]\times B^d(R)$ it follows that for any $(t,x)\in[0,T]\times B^d(R)$ we have that there exists $(t',x')\in A_{\rho_{OU}}$ such that $\rho_{OU}((t,x),(t',x'))\leq \epsilon$ then by Lemma 6 it follows that

$$\begin{aligned} ||\bar{g}_{t,x}(z) - \bar{g}_{t',x'}(z)||_{\mathscr{L}^2(Q)} &\leq L\left(1+\sqrt{2}||z||_{\mathscr{L}^2(Q)}\right)\rho_{OU}((t,x),(t',x')) & (41)\\ &\leq ||F||_{\mathscr{L}^2(Q)}\rho_{OU}((t,x),(t',x')) & (42)\\ &\leq ||F||_{\mathscr{L}^2(Q)}\epsilon & (43)\end{aligned}$$

Hence the set:

$$\mathcal{G}_{\rho_{OU}} = \{\bar{g}_{t,x} : (t,x)\in A_{\rho_{OU}}\} \tag{44}$$

is an $||F||\epsilon$ cover of $\mathcal{G}$ with respect to the metric ρ_{OU} ■

Lemma 9 *We have that*

$$N([0,T]\times B^d(R), \rho_{OU}, \epsilon) \leq N([0,T], |\cdot|, \epsilon^2/4)N(B^d(R), ||\cdot||, \epsilon/2) \tag{45}$$

Proof

Let $B^d_{r_0}(R)$ denote a euclidean d-dimensional ball of radius R centered at r_0 and let $B^{d+1}_{t_0\oplus x_0,\rho}(R')$[1] denote it's counterpart with respect to the metric ρ. Now notice that if $||e^{-t}x - e^{-t_0}x_0|| + |t-t_0|^{1/2}\leq\epsilon$ then $||e^{-t_0}(x-x_0)||\leq\epsilon$ and $||x-x_0||\leq e^{t_0}\epsilon$, thus,

$$\{t_0\}\times B^d_{x_0}(\epsilon)\subseteq\{t_0\}\times B^d_{x_0}(e^{t_0}\epsilon)\subseteq B^{d+1}_{t_0\oplus x_0,\rho}(\epsilon), \tag{46}$$

then since $\{t_0\}\times B^d_{x_0}(e^{t_0}\epsilon)\subseteq B^{d+1}_{t_0\oplus x_0,\rho}(\epsilon)$ we can construct an ϵ cover namely A_{t_0} of $\{t_0\}\times B^d(R)$ with $N(B^d(R), ||\cdot||, \epsilon e^{t_0})$ balls. Finally notice that if $||e^{-t}x - e^{-t_0}x_0|| + |t-t_0|^{1/2}\leq\epsilon$ it follows that $|t-t_0|^{1/2}\leq\epsilon$ thus $[0,T]$ can be covered in $N([0,T], |\cdot|, \epsilon^2)\leq T\epsilon^{-2}$ sub intervals.

1. $a\oplus b$ denotes the concatenation of a and b.

Let U_T be the smallest cover containing $N([0,T],|\cdot|,\epsilon^2)$ intervals u_n each centered at t_n , then:

$$A = \bigcup_{u_n \in U_T} A_{t_n} \tag{47}$$

is an ϵ cover of $[0,T] \times B^d(R)$ (with respect to the metric ρ_{OU}), notice this follows as $\forall x \in B^d(R)$ there exists an x_0 such that

$$[t_n - \epsilon^2, t_n + \epsilon^2] \times \{x\} \subseteq B^{d+1}_{t_n \oplus x_0, \rho}(\epsilon) \in A_{t_n} \tag{48}$$

Now we can see that

$$\begin{aligned} |A| \leq |U_T||A_0| = N([0,T],|\cdot|,\epsilon^2)N(B^d(R),||\cdot||,\epsilon e^{t_0}), && (49)\\ \leq N([0,T],|\cdot|,\epsilon^2/4)N(B^d(R),||\cdot||,\epsilon/2), && (50) \end{aligned}$$

where $|A_0| = \max_n |A_{t_n}|$, completing our proof.

■

From Lemmas 8, 9 it follows that :

$$N(\mathcal{G}, \mathscr{L}^2(Q), \epsilon ||F||_{\mathscr{L}^2(Q)}) \leq N([0,T],|\cdot|,\epsilon^2/4)N(B^d(R),||\cdot||,\epsilon/2) \tag{51}$$

and thus it follows (see the start of Page 18 in Tzen and Raginsky (2019a)) that Lemmas C.4 and thus Theorem C.1 in Tzen and Raginsky (2019b) hold true in our setting, with the modified choice of

$$N = \left\lceil \left(\frac{C\sqrt{d}}{\varepsilon} \cdot L((R \vee 1) + \sqrt{2d} + \sqrt{6}) \cdot (16\sqrt{6\pi R d} + 5\sqrt{\ln 4(d+1)}) \right)^2 \right\rceil, \tag{52}$$

for Theorem C.1., which we will restate now for completeness.

Corollary 11 *(Theorem C.1. from Tzen and Raginsky (2019b))*
For any $\varepsilon > 0$ and any $R > 0$, there exist $N = \mathrm{poly}(1/\varepsilon, d, L, R)$ points $z_1, \ldots, z_N \in \mathbb{R}^d$, for which the following holds:

$$\max_{n \leq N} \|z_n\| \leq 8\sqrt{(d+6)\ln N}$$

$$\sup_{x \in \mathrm{B}^d(R)} \sup_{t \in [0,1]} \left| \frac{1}{N} \sum_{n=1}^{N} f\left(e^{-t}x + (1-e^{-2t})^{1/2} z_n\right) - U_t f(x) \right| \leq \varepsilon$$

$$\sup_{x \in \mathrm{B}^d(R)} \sup_{t \in [0,1]} \left\| \frac{1}{N} \sum_{n=1}^{N} \nabla f\left(e^{-t}x + (1-e^{-2t})^{1/2} z_n\right) - \nabla U_t f(x) \right\| \leq \varepsilon$$

We now have everything that is required to show the neural network approximation results.

Appendix E. Neural Network Approximation

Corollary 12 *Under Assumption 2, the vector field $\nabla \ln U_t f(x)$ is bounded in norm by $\frac{L}{c}$ and is Lipschitz with constant $\frac{L}{c} + \frac{L^2}{c^2}$ where L is the max of the Lip constant of f and ∇f.*

Proof By direct application of Lemma B.1. (Tzen and Raginsky (2019b)) and our Lemma 5, which assures that OU semi-group commutes with the gradient operator, we have that the results of this Corollary hold. ∎

We now proceed to adapt one of the main theorems in Tzen and Raginsky (2019b). Whilst the changes are minor to the sketch in Tzen and Raginsky (2019a) some are subtle thus we have incorporated this proof for completeness. We highlight in magenta the subtle changes required to adapt the result.

Corollary 13 *(Tzen and Ragisnky) Let $0 < \varepsilon < 4L/c$ and $R > 0$ be given. Then there exists a neural net $\widehat{v} : \mathbb{R}^d \times [0, 1] \to \mathbb{R}^d$ of size polynomial in $1/\varepsilon, d, L, R, c, 1/c$, such that the activation function of each neuron is an element of the set $\{\sigma, \sigma', \mathrm{ReLU}\}$, and the following holds:*

$$\sup_{x \in \mathrm{B}^d(R)} \sup_{t \in [0,1]} \|\widehat{v}(x, t) - \nabla \ln U_t f(x)\| \leq \varepsilon$$

and

$$\max_{i \in [d]} \sup_{x \in \mathbb{R}^d} \sup_{t \in [0,1]} |\widehat{v}_i(x, t)| \leq \frac{2L}{c}.$$

Proof Let $\delta = \frac{c^2\varepsilon}{16L}$. By Theorem C.1 (which has been proved to hold true in our settings in Appendix C), there exist points $z_1, \ldots, z_N \in \mathbb{R}^d$ with $N = \mathrm{poly}(1/\delta, d, L, R)$, such that $R_{N,d} := \max_{n \leq N} \|z_n\| \leq 8\sqrt{(d + 6) \ln N}$, and the function $\varphi : \mathbb{R}^d \times [0, 1] \to \mathbb{R}$ defined by

$$\varphi(x, t) := \frac{1}{N} \sum_{n=1}^{N} f\left(e^{-t} x + (1 - e^{-2t})^{1/2} z_n\right) \tag{53}$$

satisfies

$$\sup_{x \in \mathrm{B}^d(R)} \sup_{t \in [0,1]} |\varphi(x, t) - U_t f(x)| \leq \delta \quad \text{and} \quad \sup_{x \in \mathrm{B}^d(R)} \sup_{t \in [0,1]} \|\nabla\varphi(x, t) - \nabla U_t f(x)\| \leq \delta$$

By Assumption 4, there exists a neural net $\widehat{f} : \mathbb{R}^d \to \mathbb{R}$ be that approximates f and the gradient of f to accuracy δ on the blown-up ball $\mathrm{B}^d(R + R_{N,d})$. Then the function

$$\widehat{\varphi} : \mathbb{R}^d \times [0, 1] \to \mathbb{R}, \quad \widehat{\varphi}(x, t) := \frac{1}{N} \sum_{n=1}^{N} \widehat{f}\left(e^{-t} x + (1 - e^{-2t})^{1/2} z_n\right)$$

can be computed by a neural net of size $N \cdot \mathrm{poly}(1/\delta, d, L, R)$, such that

$$
\begin{aligned}
&\sup_{x\in \mathrm{B}^d(R)} \sup_{t\in[0,1]} |\widehat{\varphi}(x,t) - U_t f(x)| \\
&\le \sup_{x\in \mathrm{B}^d(R)} \sup_{t\in[0,1]} |\widehat{\varphi}(x,t) - \varphi(x,t)| + \sup_{x\in \mathrm{B}^d(R)} \sup_{t\in[0,1]} |\varphi(x,t) - U_t f(x)| \\
&\le \sup_{x\in \mathrm{B}^d(R)} \sup_{t\in[0,1]} \left| \frac{1}{N}\sum_{n=1}^{N} \widehat{f}\left(x + (1-e^{-2t})^{1/2} z_n\right) - \frac{1}{N}\sum_{n=1}^{N} f\left(x + (1-e^{-2t})^{1/2} z_n\right) \right| \\
&\qquad + \sup_{x\in \mathrm{B}^d(R)} \sup_{t\in[0,1]} |\varphi(x,t) - U_t f(x)| \\
&\le \sup_{x\in \mathrm{B}^d\left(R+R_{N,d}\right)} |\widehat{f}(x) - f(x)| + \sup_{x\in \mathrm{B}^d(R)} \sup_{t\in[0,1]} |\varphi(x,t) - U_t f(x)| \le 2\delta
\end{aligned}
$$

where the third inequality follows since $e^{-t} \in [0,1]$ and the final inequality follows since

$$
\max_n \sup_{t\in[0,1]} (1-e^{-2t})^{1/2} ||z_n|| = \max_n ||z_n|| = R_{N,d}
$$

Similarly

$$
\begin{aligned}
&\sup_{x\in \mathrm{B}^d(R)} \sup_{t\in[0,1]} \|\nabla\widehat{\varphi}(x,t) - \nabla U_t f(x)\| \\
&\le \sup_{x\in \mathrm{B}^d(R)} \sup_{t\in[0,1]} \|\nabla\widehat{\varphi}(x,t) - \nabla\varphi(x,t)\| + \sup_{x\in \mathrm{B}^d(R)} \sup_{t\in[0,1]} \|\nabla\varphi(x,t) - \nabla U_t f(x)\| \\
&\le \sup_{x\in \mathrm{B}^d\left(R+R_{N,d}\right)} \|\nabla\widehat{f}(x) - \nabla f(x)\| + \sup_{x\in \mathrm{B}^d(R)} \sup_{t\in[0,1]} \|\nabla\varphi(x,t) - \nabla U_t f(x)\| \le 2\delta.
\end{aligned}
$$

Since f is L-Lipschitz and bounded below by c, we have $U_t f(x) \ge \mathbb{E}_{Z\sim\mathcal{N}(0,I)}[c] = c$, and

$$
\begin{aligned}
U_t f(x) = \mathbb{E}_{Z\sim\mathcal{N}(0,I)}\left[f(e^{-t}x + (1-e^{-2t})^{1/2} Z)\right] &\le \mathbb{E}_{Z\sim\mathcal{N}(0,I)}\left[L(||x|| + \sqrt{2}||z||) + f(0)\right] \\
&= L||x|| + f(0) + L\sqrt{2}\mathbb{E}[||z||] \\
&\le L(||x|| + \sqrt{2d}) + f(0)
\end{aligned}
$$

Thus it follows that $c \le U_t f(x) \le L(\|x\| + \sqrt{2d}) + f(0)$ for any $x \in \mathbb{R}^d$ and $t \in [0,1]$. Therefore, on $\mathrm{B}^d(R) \times [0,1]$,

$$
\frac{c}{2} \le \widehat{\varphi}(x,t) \le L(R + \sqrt{2d}) + f(0) + \frac{c}{2}
$$

where we use $\delta \le c/4$. Without loss of generality, we may assume that $L \ge 1$. Then, for any $x \in \mathrm{B}^d(R)$ and $t \in [0,1]$

$$
\begin{aligned}
&\|\nabla \ln \widehat{\varphi}(x,t) - \nabla \ln U_t f(x)\| \\
&= \left\| \frac{\nabla\widehat{\varphi}(x,t)}{\widehat{\varphi}(x,t)} - \frac{\nabla U_t f(x)}{U_t f(x)} \right\| \\
&\le \frac{1}{\widehat{\varphi}(x,t)} \|\nabla\widehat{\varphi}(x,t) - \nabla U_t f(x)\| + \left\| \frac{\nabla U_t f(x)}{U_t f(x)} \right\| \frac{|\widehat{\varphi}(x,t) - U_t f(x)|}{\widehat{\varphi}(x,t)} \\
&\le \frac{2L}{c} \cdot 2\delta + \frac{L}{c} \cdot \frac{2}{c} \cdot 2\delta \\
&\le \frac{\varepsilon}{2},
\end{aligned}
$$

where we have used Corollary 12 to bound $\left\|\frac{\nabla U_t f}{U_t f}\right\| \le L/c$. In other words, $\nabla \ln \widehat{\varphi}(x,t)$ approximates $\nabla \ln U_t f(x)$ to accuracy $\varepsilon/2$ uniformly on $\mathrm{B}^d(R) \times [0,1]$. It remains to approximate $\nabla \ln \widehat{\varphi}(x,t)$ by a neural net to accuracy $\varepsilon/2$.

To that end, we first represent $\nabla \ln \widehat{\varphi}(x,t)$ as a composition of several elementary operations and then approximate each step by a neural net. Specifically, the computation of $v_i = \partial_i \ln \widehat{\varphi}(x,t)$ can be represented as a computation graph with the following structure:

1. Compute $a = \widehat{\varphi}(x,t)$.
2. Compute $b_i = \partial_i \widehat{\varphi}(x,t)$.
3. Compute $r = 1/a$.
4. Compute $v_i = r b_i$.

Given x and t, a is computed by a neural net with activation function σ, of size $\mathrm{poly}(1/\delta, d, L, R)$ and depth poly $(1/\delta, d, L, R)$. Therefore, by the cheap gradient principle (Lemma D.1 from Tzen and Raginsky (2019b)), b_i can be computed by a neural net of size poly $(1/\delta, d, L, R)$, where the activation function of each neuron is an element of the set $\{\sigma, \sigma'\}$. Next, since a takes values in $[c/2, L(R+\sqrt{2d}) + f(0) + c/2]$, by Lemma D.2 from Tzen and Raginsky (2019b) the reciprocal $r = 1/a$ can be computed to accuracy $\varepsilon/(4L\sqrt{d})$ by a 2 -layer neural net with activation function σ and of size

$$\mathcal{O}\left(\frac{4}{c^2} \cdot (L(R+\sqrt{2d}) + f(0) + c/2) \cdot \frac{4L\sqrt{d}}{\varepsilon}\right) \le \mathrm{poly}(1/\varepsilon, d, L, R, c, 1/c)$$

Let $\widehat{r}$ denote the resulting approximation. Then, since $|b_i| \le 2L$ and $|\widehat{r}| \le 2/c + \varepsilon/(4L\sqrt{d}) \le 4/c$, by Lemma D.2 the product $\widehat{r} b_i$ can be approximated to accuracy $\varepsilon/4\sqrt{d}$ by a 2-layer neural net with activation function σ and with at most

$$\mathcal{O}\left((4/c \vee 2L)^2 \cdot \frac{4\sqrt{d}}{\varepsilon}\right) \le \mathrm{poly}(1/\varepsilon, d, L, 1/c)$$

neurons. The overall accuracy of the approximation is

$$|\widehat{v}_i - v_i| \le |\widehat{v}_i - \widehat{r} b_i| + |\widehat{r} b_i - r b_i| \le \frac{\varepsilon}{2\sqrt{d}}$$

Thus, the vector $v = (v_1, \ldots, v_d)$ can be $\varepsilon/2$-approximated by $\tilde{v}(x,t)$, where $\tilde{v} : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ is a neural net with vector-valued output that has the size $\mathrm{poly}(1/\varepsilon, d, L, R, c, 1/c)$. Finally, since $\sup_{x \in \mathrm{B}^d(R)} \sup_{t \in [0,1]} |\tilde{v}_i(x,t)| \le 2L/c$, the function

$$\widehat{v}_i(x,t) := \min\left\{\max\left\{\tilde{v}_i(x,t), -2L/c\right\}, 2L/c\right\}$$

is continuous, takes values in $[-2L/c, 2L/c]$ and coincides with $\tilde{v}_i$ on $\mathrm{B}^d(R) \times [0,1]$. Moreover, the min and max operations can each be implemented exactly using $\mathcal{O}(1)$ ReLU neurons. ■

Proposition 1 *Suppose Assumptions 1-3 are in force. Let L denote the maximum of the Lipschitz constants of f and ∇f. Then for all $0 < \epsilon < 16L^2/c^2$, there exists a neural net $\hat{v} : R^d \times [0,1] \to R^d$ with size polynomial in $1/\epsilon, d, L, c, 1/c$ such that the activation function of each neuron in the set of $\{\sigma, \sigma', ReLU\}$, and the following hold: If $\{\hat{x}_t\}_{t\in[0,1]}$ is the diffusion process governed by the Itô SDE:*

$$d\hat{x}_t = \hat{b}(\hat{x}_t, t)\mathrm{d}t + \sqrt{2\beta}\mathrm{d}W_t \tag{54}$$

with $x_0 \sim p_1 \approx \mathcal{N}(0, I)$ with the drift $\hat{b}(x,t) = -(x - 2\hat{v}(x, 1-t))$, then $\hat{\mu} := \mathrm{Law}(\hat{x}_1)$, satisfies $D(\mu||\hat{\mu}) \leq \epsilon$.

Proof

For any $R > 0$, Corollary 13 guarantees the existence of a neural net $\widehat{v} : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ that satisfies

$$\sup_{x\in \mathrm{B}^d(R)} \sup_{t\in[0,1]} \|\widehat{v}(x,t) - \nabla \ln U_t f(x)\| \leq \sqrt{\varepsilon} \tag{55}$$

and

$$\max_{i\in[d]} \sup_{x\in\mathbb{R}^d} \sup_{t\in[0,1]} |\widehat{v}_i(x,t)| \leq \frac{2L}{c}. \tag{56}$$

Let $\boldsymbol{\mu} := \mathrm{Law}\left(x_{[0,1]}\right)$ and $\widehat{\boldsymbol{\mu}} := \mathrm{Law}\left(\widehat{x}_{[0,1]}\right)$. The Girsanov formula gives

$$\mathrm{KL}(\boldsymbol{\mu}\|\widehat{\boldsymbol{\mu}}) = \frac{1}{2}\int_0^1 \mathbf{E}\left\|b\left(x_t,t\right) - \widehat{b}\left(x_t,t\right)\right\|^2 \mathrm{d}t$$

where the interchange of the integral and the expectation follows from Fubini's theorem because both b and $\widehat{b}$ are bounded by Corollary 12 and (56). We now proceed to estimate the integrand. For each $t \in [0,1]$

$$\begin{aligned}
&\mathbf{E}\left\|b\left(x_t,t\right) - \widehat{b}\left(x_t,t\right)\right\|^2 \\
&= \mathbf{E}\left[\left\|b\left(x_t,t\right) - \widehat{b}\left(x_t,t\right)\right\|^2 \cdot \mathbf{1}\left\{x_t \in \mathrm{B}^d(R)\right\}\right] + \mathbf{E}\left[\left\|b\left(x_t,t\right) - \widehat{b}\left(x_t,t\right)\right\|^2 \cdot \mathbf{1}\left\{x_t \notin \mathrm{B}^d(R)\right\}\right] \\
&=: T_1 + T_2,
\end{aligned}$$

where $T_1 \leq \varepsilon$ by (56). To estimate T_2, we first observe that, since the OU drift is bounded in norm by L/c by 12, we have

$$\mathbf{P}\left\{\sup_{t\in[0,1]} \|x_t\| \geq R\right\} \leq \frac{\sqrt{d} + L/c}{R}$$

(Bubeck et al. (2018), Lemma 3.8). Therefore,

$$T_2 \leq \frac{9dL^2}{c^2} \cdot \frac{\sqrt{d} + L/c}{R}$$

Since some of the bounds differ from the original Tzen and Raginsky (2019b) we verify that the bound still holds for our drift. We used that $d \geq 2$.

$$T_2 = \mathbf{E}\left[\left\|b\left(x_t,t\right)-\widehat{b}\left(x_t,t\right)\right\|^2\cdot\mathbf{1}\left\{x_t\notin \mathrm{B}^d(R)\right\}\right] = \int_{x_t\notin \mathrm{B}^d(R)}\|b\left(x_t,t\right)-\widehat{b}\left(x_t,t\right)\|^2 dP_{x_t} =$$

$$= \int_{x_t\notin \mathrm{B}^d(R)} 2\|b\left(x_t,t\right)\|^2 + 2\|\widehat{b}\left(x_t,t\right)\|^2 dP_{x_t} \leq \int_{x_t\notin \mathrm{B}^d(R)} 2\|b\left(x_t,t\right)\|^2 + 2d\left(\frac{2L}{c}\right)^2 dP_{x_t} \leq$$

$$\leq \int_{x_t\notin \mathrm{B}^d(R)} 2\|\nabla \ln U_t f(x_t)\|^2 + 8d\left(\frac{L}{c}\right)^2 dP_{x_t} = \int_{x_t\notin \mathrm{B}^d(R)} 2\left\|\frac{\nabla U_t f(x_t)}{U_t f(x_t)}\right\|^2 + 8d\left(\frac{L}{c}\right)^2 dP_{x_t} \leq$$

$$\leq \int_{X_t\notin \mathrm{B}^d(R)} 2\frac{L}{c}^2 + 8d\left(\frac{L}{c}\right)^2 dP_{x_t} \leq 9d\frac{L^2}{c^2}P\left\{\sup_{t\in[0,1]}\|x_t\|\geq R\right\} \leq \frac{9dL^2}{c^2}\cdot\frac{\sqrt{d}+L/c}{R}$$

Choosing R large enough to guarantee $T_2 \leq \varepsilon$ and putting everything together, we obtain $D(\boldsymbol{\mu}\|\widehat{\boldsymbol{\mu}}) \leq \varepsilon$. Therefore, $D(\mu\|\widehat{\mu}) \leq D(\boldsymbol{\mu}\|\widehat{\boldsymbol{\mu}}) \leq \varepsilon$ by the data processing inequality. ■

Finally, we would like to highlight what happens when we sample $\hat{x}_0 \sim \mathcal{N}(0,1)$ rather than p_T. Whilst our results are done for $t\in[0,1]$ one can see that the overall approximation results will hold for $t \in [0,T]$.

Remark 2 *Assuming π satisfies a logarithmic Sobolev inequality we extend the time domain to $t\in[0,T]$ and sampling $\hat{x}_0 \sim \mathcal{N}(0,I)$ approximately, it follows that $D(\mu||\hat{\mu}) \leq e^{-T}\mathrm{KL}(\pi||\mathcal{N}(0,1)) + T\epsilon$*

Proof

First, we remark that the estimation results and the results in Proposition 1 apply to the $t\in[0,T]$ setting, however, they will introduce a polynomial dependency in T for the size of the network.

As in the above proof, we apply the Girsanov theorem to control the path KL, however here the starting distributions of the two Ito processes are no longer the same thus we get an extra term from the chain rule:

$$\mathrm{KL}(\boldsymbol{\mu}\|\widehat{\boldsymbol{\mu}}) = \mathrm{KL}(p_T||\mathcal{N}(0,1)) + \frac{1}{2}\int_0^T \mathbf{E}\left\|b\left(x_t,t\right)-\widehat{b}\left(x_t,t\right)\right\|^2 \,\mathrm{d}t \tag{57}$$

$$\leq \mathrm{KL}(p_T||\mathcal{N}(0,1)) + T\epsilon \tag{58}$$

$$\leq e^{-T}\mathrm{KL}(\pi||\mathcal{N}(0,1)) + T\epsilon \tag{59}$$

Where the final inequality follows from Theorem 5.2.1 in Bakry et al. (2014) under the assumption that π satisfies a log-Sobolev inequality. This completes the circle and fully extends Theorem 3.1 from Tzen and Raginsky (2019b) to our denoising diffusion setting. ■

Finally, note that if we assume that $\operatorname{supp}\pi \subseteq \mathrm{B}^d(R)$ from Theorem 2 of Chen et al. (2022) it follows that:

$$\mathrm{TV}\left(\mathrm{Law}\hat{x}_t,\pi\right) \leq \mathcal{O}\left(\sqrt{\mathrm{KL}\left(\pi\|\mathcal{N}(0,I)\right)}\exp(-T) + \epsilon\sqrt{T}\right). \tag{60}$$

This result complements Proposition 1 very nicely as unlike Chen et al. (2022) we no longer require assuming an ϵ error on the score but instead prove such error can be attained.