
On learning higher-order cumulants in diffusion models

Gert Aarts and Diaa E. Habibi

Department of Physics, Swansea University, Swansea, SA2 8PP, United Kingdom
 g.aarts@swansea.ac.uk (corresponding author), n.e.habibi@swansea.ac.uk

Lingxiao Wang

Interdisciplinary Theoretical and Mathematical Sciences Program (iTHEMS), RIKEN
 Wako, Saitama 351-0198, Japan
 lingxiao.wang@riken.jp

Kai Zhou

School of Science and Engineering, The Chinese University of Hong Kong
 Shenzhen (CUHK-Shenzhen), Guangdong, 518172, China
 Frankfurt Institute for Advanced Studies, D-60438, Frankfurt am Main, Germany
 zhoukai@cuhk.edu.cn

February 28, 2025

Abstract

To analyse how diffusion models learn correlations beyond Gaussian ones, we study the behaviour of higher-order cumulants, or connected n -point functions, under both the forward and backward process. We derive explicit expressions for the moment- and cumulant-generating functionals, in terms of the distribution of the initial data and properties of forward process. It is shown analytically that during the forward process higher-order cumulants are conserved in models without a drift, such as the variance-expanding scheme, and that therefore the endpoint of the forward process maintains nontrivial correlations. We demonstrate that since these correlations are encoded in the score function, higher-order cumulants are learnt in the backward process, also when starting from a normal prior. We confirm our analytical results in an exactly solvable toy model with nonzero cumulants and in scalar lattice field theory.

1 Introduction

Diffusion models [1–4] – see also the review [5] – are a widely used class of deep generative models, able to generate high-quality images and videos via a stochastic denoising process. In diffusion models, images are scrambled during the forward process, by applying random noise drawn from a normal distribution to each pixel. It is often stated that at the end of the forward process the images are close to being fully random, that is, without any correlations remaining among pixels. During the forward process, the change in the logarithm of the distribution function is learned (“score matching”) [6]. In the backward process, this score is applied to initial conditions drawn from a normal distribution and new images are generated (“denoising”) [3]. Diffusion models are powerful and widely employed, see e.g. Stable Diffusion [7] and DALL-E 2 [8].

To deepen the understanding of diffusion models, it is important to understand how correlations beyond Gaussian ones evolve during both the forward and backward process. To address this, we

use generating functionals and lattice field theory as robust and well-understood frameworks. On the one hand, lattice field theories are widely studied and well understood in theoretical high-energy physics where they are used to solve strongly interacting quantum field theories describing nature, such as Quantum Chromodynamics, see e.g. the textbooks [9, 10]. Importantly, interactions between fundamental degrees of freedom are encoded in the higher n -point functions, which can be computed using perturbation theory in terms of Feynman diagrams as well as nonperturbatively using lattice simulations. Lattice field theories are therefore useful playgrounds to assess the feasibility of diffusion models (and other generative methods) to learn higher-order n -point functions. On the other hand, simulations of quantum field theories require the fast generation of ensembles of field configurations and there is a long history [11] of studying strongly interacting quantum field theories numerically by combining the path integral formulation with Monte Carlo methods, after discretisation on a spacetime lattice. Most standard algorithms, such as Hybrid Monte Carlo (HMC) [12] rely on importance sampling, where issues related to critical slowing down remain, see e.g., Ref. [13]. It has been suggested – see e.g. the review [14] – that methods developed in generative AI can provide an alternative approach to generate ensembles, with configurations playing the role of two- or higher-dimensional images. Indeed, a substantial amount of work has been carried out using normalising flow [15–21] and variations thereof, such as continuous normalizing flow [22–25] and stochastic normalizing flow [26, 27]. The use of diffusion models to simulate lattice field theories was suggested recently in Refs. [28, 29] and extended to $U(1)$ gauge theory in Ref. [30], introducing physics-conditioned diffusion models. To further understand whether diffusion models can be employed in this context, it is of paramount importance to investigate whether higher-order correlations are faithfully reproduced. This provides the second motivation for this work, to test the applicability of diffusion models to generate lattice field configurations, via the analysis of higher n -point functions.

We note here that the use of field-theoretical methods to shed light on machine learning methods goes beyond what is sketched above. Ref. [31] proposes to use n -point functions to parametrise the latent layer in an autoencoder. The so-called neural network/field theory correspondence [32, 33] exploits the relation between Gaussian processes (or free fields) and neural networks in the limit of infinite width [34–38]. A relation between deep learning and the AdS/CFT correspondence can be found in Ref. [39]. In quantum field-theoretical machine learning new interaction terms are added on the nodes, to define new systems [40]. Using the language of field theory may also help in understanding the choice of architecture or hyperparameters, as shown for the Gaussian restricted Boltzmann machine [41]. A correspondence between the dynamics of learning and cosmological expansion can be found in Ref. [42] and the relation between learning and Dyson Brownian motion in Ref. [43]. The connection between diffusion models and stochastic quantisation [44–46] was pointed out in Refs. [28, 29]. A further connection with Feynman’s path integral was given in Ref. [47].

The remainder of this paper is organised as follows. In Sec. 2 we summarise the basics of diffusion models, using the language of stochastic differential equations. Moments and cumulants are considered in Sec. 3, where we derive explicit expressions for the generating functionals. Subsequently we verify the analytical results in the case of an exactly solvable model in Sec. 4 and a two-dimensional scalar field theory in Sec. 5. Conclusions are summarised in Sec. 6. App. A contains a brief overview of the numerical implementation of the diffusion model, while App. B gives the analytical solutions for the forward and backward process in the case of a Gaussian target distribution. We consider both score-based, variance-expanding schemes [3, 48] and a wide class of denoising diffusion probabilistic models (DDPMs) [1, 2] in the continuous-time limit.

2 Diffusion models

Diffusion models consist of a forward process, in which images or configurations are made more and more noisy, and a backward process, during which new images or configurations are generated in the denoising process. We use the description in term of stochastic differential equations (SDEs) [4, 49, 50], and consider here for notational simplicity one degree of freedom, x .¹ The forward process is then determined by the stochastic equation,

$$\dot{x}(t) = K(x(t), t) + g(t)\eta(t), \quad (2.1)$$

where $K(x(t), t)$ is a possible drift term, $\eta \sim \mathcal{N}(0, 1)$ is Gaussian noise with variance 1, and $g(t)$ is the time-dependent noise strength. The initial condition for the forward process, $x(0) = x_0$, is

¹In Sec. 5, we will generalise to this to lattice field theory, using the replacement $x(t) \rightarrow \phi(x, t)$.

determined by the target distribution $P_0(x_0)$, i.e. $x_0 \sim P_0(x_0)$, which is either known explicitly or implicitly via a data set. We always consider target distributions for which the first moment vanishes, or has been subtracted, $x_0 \rightarrow x_0 - \mathbb{E}_{P_0}[x_0]$. The forward process runs between $0 \leq t \leq T$, where the usual choice is $T = 1$. Expectation values are taken by an average over both the noise distribution and the target/data distribution $P_0(x_0)$.

The corresponding backward process is written in terms of $\tau = T - t$, such that $0 \leq \tau \leq T$. In this process the drift should contain a time-dependent term which ensures convergence to the target distribution as $\tau \rightarrow T$. This term is given by the change in the logarithm of the distribution $P(x, t)$ in principle. The backward process then reads

$$x'(\tau) = -K(x(\tau), T - \tau) + g^2(T - \tau) \partial_x \log P(x, T - \tau) + g(T - \tau) \eta(\tau). \quad (2.2)$$

The initial conditions for the backward process are drawn from a normal distribution with a variance comparable to the final variance of the forward process. The second term on the RHS is the essential term to determine, which can be achieved during the forward process via score matching. The numerical implementation of score matching relies on neural networks and is summarised in App. A.

There is considerable freedom in choosing the drift $K(x, t)$ and the noise strength $g(t)$, including a linear drift, $K(x(t), t) = -\frac{1}{2}k(t)x(t)$, or no drift term at all, i.e. pure diffusion. A popular choice is the variance-expanding scheme, in which $K(x, t) = 0$ and $g(t) = \sigma^{t/T}$, with σ a tunable but generically large parameter. The variance at the end of the forward process, $\mathbb{E}[x^2(T)] \sim \sigma^2$, should be substantially larger than the variance of the target distribution. In DDPMs the drift is of the form $K(x(t), t) = -\frac{1}{2}g^2(t)x(t)$. In App. B we give the explicit solutions for a Gaussian target distribution in both schemes, in which both the forward and the backward process can be solved analytically.

3 Moments and cumulants

During the forward process the target distribution evolves to a distribution with a predetermined second moment or variance, while during the backward process the distribution is expected to reverse to the target distribution. The question we address here is how higher-order moments or cumulants evolve, both during the forward and the backward process. We note here that it is often stated (see e.g. Refs. [4, 51]) that the final distribution of the forward process approximates a normal distribution, but we will see below that this is not the case for pure diffusion. Knowledge of higher-order cumulants is essential for the application to lattice field theories, since they contain the information on the interactions beyond the free-field limit, as stated in the Introduction.

3.1 Explicit solution for cumulants

We start by determining the explicit evolution of the lowest few higher-order cumulants during the forward process. Expressions to all orders, using moment- and cumulant- generating functions, are derived in the next subsection.

We consider the forward process with a linear drift,

$$\dot{x}(t) = -\frac{1}{2}k(t)x(t) + g(t)\eta(t). \quad (3.1)$$

Here the linear coefficient may be time dependent, constant or zero (pure diffusion). This equation is solved as

$$x(t) = x_0 f(t, 0) + \int_0^t ds f(t, s) g(s) \eta(s), \quad (3.2)$$

where $x_0 \sim P_0(x_0)$ is an initial condition and

$$f(t, s) = e^{-\frac{1}{2} \int_s^t ds' k(s')}. \quad (3.3)$$

Note that for pure diffusion, with $k(t) = 0$, $f(t, s) = 1$.

We denote the time-dependent moments, or n -point functions, as

$$\mu_n(t) = \mathbb{E}[x^n(t)], \quad (3.4)$$

where the expectation value is taken with respect to the target distribution P_0 and the noise distribution. If we only take the expectation with respect to one of these, this will be indicated explicitly with a

subscript P_0 or η respectively. Note that we only consider equal-time expectation values. Cumulants will be denoted with $\kappa_n(t)$ and are obtained easily using the expansion [52]

$$\kappa_n = \mu_n - \sum_{m=2}^{n-2} \binom{n-1}{m-1} \kappa_m \mu_{n-m}, \quad (3.5)$$

with $\mu_1 = \kappa_1 = 0$.

Recall that the target distribution has a vanishing one-point function $\mathbb{E}_{P_0}[x_0] = 0$ (after subtraction, $x_0 \rightarrow x_0 - \mathbb{E}_{P_0}[x_0]$, if required). Using the solution (3.2), the second moment and cumulant then read

$$\kappa_2(t) = \mu_2(t) = \mu_2(0)f^2(t, 0) + \Xi(t), \quad (3.6)$$

where $\mu_2(0) = \mathbb{E}_{P_0}[x_0^2]$ is the variance of the target distribution and

$$\Xi(t) = \int_0^t ds \int_0^t ds' f(t, s)f(t, s')g(s)g(s')\mathbb{E}_\eta[\eta(s)\eta(s')] = \int_0^t ds f^2(t, s)g^2(s). \quad (3.7)$$

The third moment and cumulant are identical and easy to evaluate,

$$\kappa_3(t) = \mu_3(t) = \kappa_3(0)f^3(t, 0). \quad (3.8)$$

Using the solution (3.2), the fourth moment is given by

$$\mu_4(t) = \mu_4(0)f^4(t, 0) + 6\mu_2(0)f^2(t, 0)\Xi(t) + 3\Xi^2(t). \quad (3.9)$$

The fourth cumulant is given by

$$\kappa_4(t) = \mu_4(t) - 3\mu_2^2(t). \quad (3.10)$$

Inserting the expressions for $\mu_4(t)$ and $\mu_2(t)$ then yields

$$\kappa_4(t) = [\mu_4(0) - 3\mu_2^2(0)]f^4(t, 0) = \kappa_4(0)f^4(t, 0). \quad (3.11)$$

Similarly, after some algebra, the fifth and sixth cumulants read

$$\kappa_5(t) = [\mu_5(0) - 10\mu_3(0)\mu_2(0)]f^5(t, 0) = \kappa_5(0)f^5(t, 0), \quad (3.12)$$

$$\kappa_6(t) = [\mu_6(0) - 15\mu_4(0)\mu_2(0) - 10\mu_3^2(0) + 30\mu_2^3(0)]f^6(t, 0) = \kappa_6(0)f^6(t, 0). \quad (3.13)$$

Importantly, we find therefore that all cumulants with $n > 2$ considered so far take the same form,

$$\kappa_{n>2}(t) = \kappa_n(0)f^n(t, 0), \quad (3.14)$$

i.e. they are equal to the product of the n^{th} cumulant of the target distribution and a simple time-dependent function, raised to the power n . For pure diffusion, $f(t, 0) = 1$ and hence

$$\kappa_{n>2}(t) = \kappa_n(0) \quad (\text{pure diffusion}), \quad (3.15)$$

i.e. the higher-order cumulants are preserved under the forward process. This implies that the final distribution of the forward process is not a normal distribution, but is as correlated as the target distribution, albeit with a different second moment (3.6).

A qualitative different result is obtained for a forward process with a nonzero drift term. Taking for simplicity a time-independent one, $k(t) = k$, such that $f(t, 0) = \exp(-kt/2)$, one finds that the cumulants decay exponentially,

$$\kappa_{n>2}(t) = \kappa_n(0)e^{-\frac{n}{2}kt} \quad (\text{constant drift}). \quad (3.16)$$

Only the second moment remains nonzero,

$$\mu_2(t) = \mu_2(0)e^{-kt} + \int_0^t ds e^{-k(t-s)}g^2(s), \quad (3.17)$$

with the dependence on the target data exponentially suppressed. In this case, the final distribution of the forward process is a normal distribution, up to exponentially suppressed terms.

3.2 Moment- and cumulant-generating functions

The demonstration above was carried out at a given order. It is helpful to prove these results to all orders, using the moment- and cumulant-generating functions, defined by

$$Z[J] = \mathbb{E}[e^{J(t)x(t)}], \quad W[J] = \log Z[J]. \quad (3.18)$$

The normalisation is such that $Z[0] = 1$. We take the solution (3.2) and consider first the average over the noise (see e.g. Ref. [45] for conventions)

$$Z_\eta[J] = \mathbb{E}_\eta[e^{J(t)x(t)}] = \frac{\int D\eta e^{-\frac{1}{2} \int_0^t ds \eta^2(s) + J(t)[x_0 f(t,0) + \int_0^t ds f(t,s)g(s)\eta(s)]}}{\int D\eta e^{-\frac{1}{2} \int_0^t ds \eta^2(s)}}. \quad (3.19)$$

Completing the square in the exponential and performing the integral over η then yields

$$Z_\eta[J] = e^{J(t)x_0 f(t,0) + \frac{1}{2} J^2(t)\Xi(t)}. \quad (3.20)$$

Including now the average over the target distribution yields the final expression for the moment-generating function,

$$Z[J] = \mathbb{E}[e^{J(t)x(t)}] = e^{\frac{1}{2} J^2(t)\Xi(t)} \int dx_0 P_0(x_0) e^{J(t)x_0 f(t,0)}. \quad (3.21)$$

Here we used that the second term in the exponential in Eq. (3.20) is independent of x_0 . The cumulant-generating function immediately follows as

$$W[J] = \log Z[J] = \frac{1}{2} J^2(t)\Xi(t) + \log \int dx_0 P_0(x_0) e^{J(t)x_0 f(t,0)}. \quad (3.22)$$

This expression explains the results derived above. The second moment or cumulant is given by

$$\kappa_2(t) = \left. \frac{d^2 W[J]}{dJ(t)^2} \right|_{J=0} = \Xi(t) + \mathbb{E}_{P_0}[x_0^2] f^2(t,0), \quad (3.23)$$

in agreement with Eq. (3.6). All higher-order cumulants are independent of the stochastic part $\Xi(t)$ and proportional to the cumulants of the target theory, with the replacement $x_0 \rightarrow x_0 f(t,0)$,

$$\kappa_{n>2}(t) = \left. \frac{d^n W[J]}{dJ(t)^n} \right|_{J=0} = \frac{d^n}{dJ(t)^n} \log \mathbb{E}_{P_0}[e^{J(t)x_0 f(t,0)}] \Big|_{J=0} = \kappa_n(0) f^n(t,0). \quad (3.24)$$

This is in complete agreement with the explicit results for the fixed-order cumulants derived in the previous subsection.

We find therefore that the generating functions have a simple structure. They are of the same form as for the original target distribution $P_0(x_0)$, with the modifications:

- The degree of freedom, x_0 , is rescaled with a time-dependent function $f(t,0)$, which results in a multiplication of all connected n -point functions with a factor $f^n(t,0)$. For schemes without a drift (pure diffusion), $f(t,0) = 1$, and there is no rescaling.
- The exception is the two-point function, which contains an additive term $\Xi(t)$, which dominates at the end of the forward process.

This interpretation will be explored further in Sec. 5, where we consider the generating functionals for lattice scalar field theory.

4 Exactly solvable model with nonzero higher-order cumulants

To study the dynamics of higher-order cumulants during the forward and backward processes numerically, we consider as target distribution a linear combination of two normal distributions for one degree of freedom,

$$P_0(x) = \frac{1}{2} [\mathcal{N}(x; \mu_0, \sigma_0^2) + \mathcal{N}(x; -\mu_0, \sigma_0^2)]. \quad (4.1)$$

This distribution has peaks around $x = \pm\mu_0$ and hence resembles a distribution in a double-well potential, but all moments and hence cumulants can be computed exactly. It is also easy to generate ‘configurations’ numerically, using Gaussian random numbers and shifting them by $\pm\mu_0$ equally often. Finally, as we will see below, the time-dependent score is known analytically as well.

Since the distribution is even, all odd moments and cumulants are zero. The even moments can be found using the binomial expansion for $(x \pm \mu_0)^{2n}$ and the standard expression for moments of the normal distribution with vanishing mean. One finds

$$\mu_{2n} \equiv \mathbb{E}[x^{2n}] = \sum_{k=0}^{2n} c_{nk} \sigma_0^{2k} \mu_0^{2n-2k}, \quad c_{nk} = \frac{(2n)!}{(2n-2k)!k!2^k}. \quad (4.2)$$

Cumulants are obtained easily using the expansion (3.5) or via the cumulant-generating function

$$\kappa_n = \left. \frac{d^n W[j]}{dj^n} \right|_{j=0}, \quad W[j] = \log Z[j]. \quad (4.3)$$

The moment-generating function $Z[j]$ is a linear combination of the ones for the normal distribution with mean $\pm\mu_0$,

$$Z[j] = \mathbb{E}[e^{jx}] = e^{\frac{1}{2}\sigma_0^2 j^2} \cosh(\mu_0 j). \quad (4.4)$$

Since

$$W[j] = \frac{1}{2}\sigma_0^2 j^2 + \log \cosh(\mu_0 j), \quad (4.5)$$

only the second cumulant depends on σ_0^2 , and the few lowest cumulants are given by

$$\kappa_2 = \mu_0^2 + \sigma_0^2, \quad \kappa_4 = -2\mu_0^4, \quad \kappa_6 = 16\mu_0^6, \quad \kappa_8 = -272\mu_0^8. \quad (4.6)$$

We have analysed this model numerically using the variance-expanding scheme and a particular choice of a denoising diffusion probabilistic model.

4.1 Variance-expanding scheme

We start with the variance-expanding scheme, using $k(t) = 0$ and $g(t) = \sigma^{t/T}$, where $\sigma = 10$ and $T = 1$. The parameters in the target distribution are $\mu_0 = 1$ and $\sigma_0 = 1/4$ throughout. Some details on the implementation are given in App. A. In Fig. 1 we show the evolution of the second moment (or second cumulant), as $\kappa_2/\kappa_2^{\text{exact}} - 1$, during the forward and backward process, using 10^6 trajectories. For the latter, we use the score determined by the diffusion model as well as the analytical score (to be discussed below). As expected, the second cumulant increases during the forward process as

$$\kappa_2(t) = \kappa_2^{\text{target}} + \frac{T}{\log \sigma^2} \left[\sigma^{2t/T} - 1 \right] \sim \frac{T}{\log \sigma^2} \sigma^{2t/T}, \quad (4.7)$$

and decreases correspondingly during the backward process. The target value is obtained linearly as $\tau \rightarrow T$, as shown in the inset, see App. B for details.

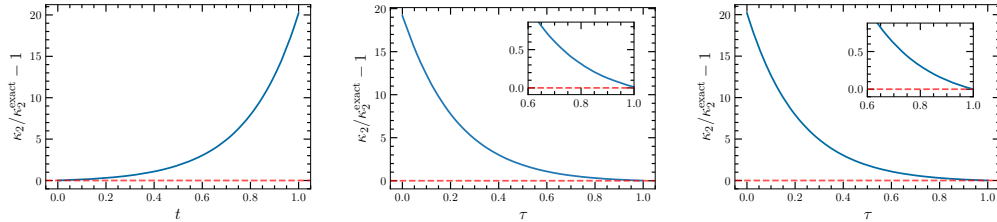


Figure 1: Evolution of the normalised second moment or cumulant, presented as $\kappa_2/\kappa_2^{\text{exact}} - 1$, in the two-peak model in the variance-expanding scheme, with $\mu_0 = 1$ and $\sigma_0 = 1/4$, during the forward process (left), the backward process with the score determined by the diffusion model (middle), and with the analytical score (right), all using 10^6 trajectories. The insets zoom in at $0.6 < \tau < 1$.

In Fig. 2 we show the forward and backward evolution of the fourth, sixth and eighth moments, as $\mu_n/\mu_n^{\text{exact}} - 1$. These moments increase (decrease) as $\sigma^{nt/T}$ ($\sigma^{n(\tau-T)/T}$), and hence grow very

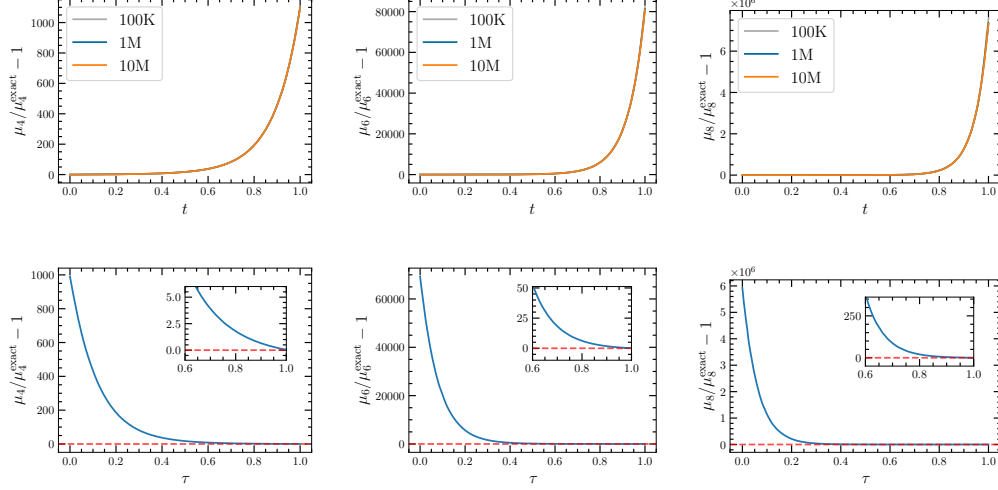


Figure 2: Evolution of the normalised 4th (left), 6th (middle) and 8th (right) moments, presented as $\mu_n/\mu_n^{\text{exact}} - 1$, in the two-peak model in the variance-expanding scheme, during the forward process using 10^5 , 10^6 and 10^7 trajectories (above), and during the backward process with the score determined by the diffusion model, using 10^6 trajectories (below). Other parameters as above.

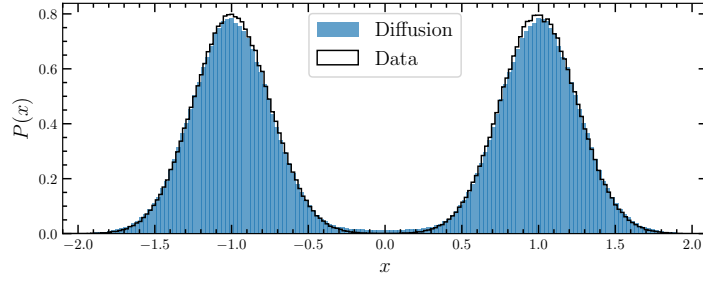


Figure 3: Distribution created by sampling from the target distribution with $\mu_0 = 1$ and $\sigma_0 = 1/4$ (Data) and from the trained diffusion model in the variance-expanding scheme (Diffusion), using 10^6 samples in each case.

large. For the forward process, we show results for 10^5 , 10^6 and 10^7 trajectories, which cannot be distinguished. The insets indicate that the target values are obtained at the end of the backward process. Before turning to the cumulants, we show in Fig. 3 the distribution, as obtained by direct sampling from Eq. (4.1) (Data) and as produced by the diffusion model (Diffusion). By eye, the distributions are matching, but we make this precise below.

The forward and backward evolution of the higher moments is dominated by the evolution of the second moment. To study the properties of the distributions in more detail, it is necessary to follow the higher-order cumulants. In Fig. 4 (top row) we show the fourth, sixth and eighth cumulants, as $\kappa_n/\kappa_n^{\text{exact}} - 1$, during the forward process. The prediction from the previous section is that these should be conserved during the evolution. We observe that they are indeed approximately constant, except towards the end of the forward process. The latter can be understood as the effect of incomplete cancellations. Higher-order cumulants are obtained as differences between (higher-order) moments $\mu_n(t)$, which each grow large, as shown in Fig. 2. The time evolution hence depends on precise cancellations, which requires sufficient statistics. This is demonstrated by including expectation values with 10^5 , 10^6 and 10^7 trajectories. The apparent numerical instability is reduced as the number of trajectories increases. This supports the analytical result that the higher-order cumulants are

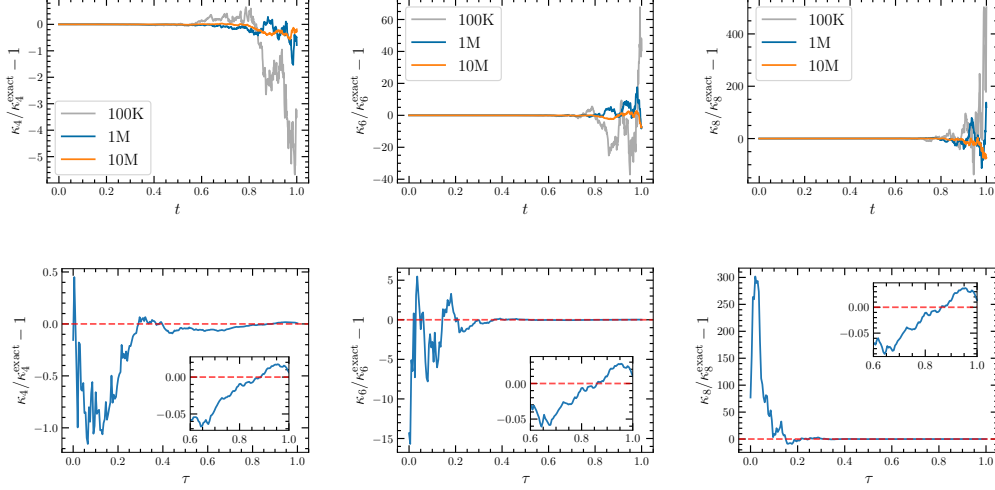


Figure 4: Evolution of the normalised 4th, 6th and 8th cumulants, presented as $\kappa_n/\kappa_n^{\text{exact}} - 1$, in the two-peak model in the variance-expanding scheme, during the forward process, using 10^5 , 10^6 and 10^7 trajectories (above), and during the backward process, with the score determined by the diffusion model, using 10^6 trajectories (below). Other parameters as above.

preserved, and that the distribution at the end of the forward process is as correlated as the target distribution (in the limit of an infinite number of trajectories).

The same effect is observed during the backward process, as shown in Fig. 4 (bottom row), using 10^6 trajectories. Initial conditions are drawn from a normal distribution and hence the cumulants are expected to be zero initially. Here we observe the reverse behaviour. Near the start of the backward process, the second and higher moments are large, leading to only a partial cancellation with a finite number of trajectories. After some time, however, the cumulants become approximately constant and approximately equal to the target value. The fluctuations shown in the inset reflect the stochastic evolution of predominantly the second moment, which approximates the target value as $\tau \rightarrow T$, as do the cumulants.

One possibility is that the large fluctuations at the start of the backward process are the result of a poorly learned score. This can be tested in this model, since the time-dependent distribution and hence the score can be obtained analytically. It is then possible to follow the backward process with the exact drift term, using a finite number of trajectories. The time-dependent distribution during the forward process in the case of pure diffusion reads

$$P(x, t) = \frac{1}{2} [\mathcal{N}(x; \mu_0, \sigma^2(t)) + \mathcal{N}(x; -\mu_0, \sigma^2(t))], \quad (4.8)$$

where $\sigma^2(t) = \sigma_0^2 + \Xi(t)$. The proof follows via the application of Eq. (4.5): with a time-dependent $\sigma^2(t)$, only the second cumulant is time dependent,

$$\mu_2(t) = \kappa_2(t) = \mu_0^2 + \sigma^2(t) = \mu_0^2 + \sigma_0^2 + \Xi(t), \quad (4.9)$$

while all the higher-order cumulants are constant. The drift of the backward process, i.e. the score, then follows from

$$-\partial_x \log P(x, t) = \frac{x}{\sigma^2(t)} - \frac{\mu_0}{\sigma^2(t)} \tanh\left(\frac{\mu_0 x}{\sigma^2(t)}\right), \quad (4.10)$$

with $t \rightarrow T - \tau$. Note that this implies that all the terms in a polynomial expansion of the drift depend on time via $\sigma^2(t)$, but that these are resummed in such a way that only the second cumulant is time dependent.

In Fig. 5, we show the evolution of the normalised cumulants during the backward process, using the analytical score with 10^6 trajectories – the second moment is shown in Fig. 1 (right). We observe the same behaviour as in the case with the learned score, see Fig. 4 (below). We hence conclude that the

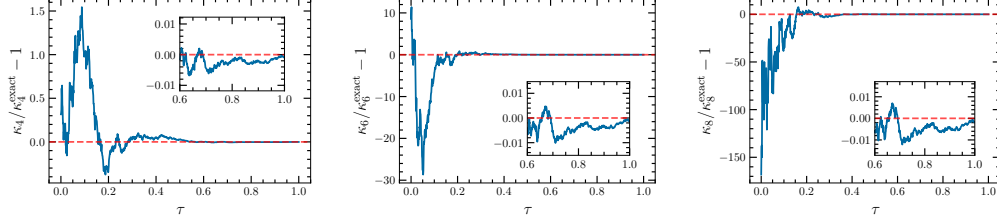


Figure 5: As in the preceding figure, employing the analytical score during the backward process, using 10^6 trajectories.

observed fluctuations during the first part of the backward process are due to the finite number of trajectories and not due to a poorly learned score, which is reassuring.

To assess quantitatively how well the cumulants are determined, we have estimated the first four nonvanishing cumulants κ_n , with the statistical uncertainty determined by a bootstrap analysis. The results are shown in Table 1. The first row contains the exact values and the second row the values generated by direct sampling (target ensemble). The third row contains the results obtained with the diffusion model in the variance-expanding scheme, as discussed above. We observe excellent agreement for the higher-order cumulants. For the second cumulant a small deviation is observed, which is however less than 1%. It is interesting that better agreement is observed for the higher-order cumulants.

	κ_2	κ_4	κ_6	κ_8
Exact	1.0625	-2	16	-272
Data	1.0624(5)	-2.000(2)	16.00(2)	-272.0(6)
Variance expanding	1.0692(6)	-2.001(2)	16.03(3)	-272.7(6)
Variance preserving (DDPM)	1.0609(5)	-1.976(2)	15.72(2)	-265.6(6)

Table 1: First four nonvanishing cumulants κ_n in the two-peak model, as obtained from training data and from diffusion models without a drift (variance expanding) and with a drift (variance preserving, DDPM). Statistical errors are computed by bootstrap resampling of a 10^6 size dataset with 1000 bins.

We conclude that the higher-order cumulants are learned correctly, within numerical uncertainty, with a noticeable effect for a finite number of trajectories near the start of the backward process. One may infer from this behaviour that the first part of the backward process is not that relevant for the evolution towards the final stages, which is worth exploring further.

4.2 Denoising diffusion probabilistic models

Next we turn to the class known as denoising diffusion probabilistic models, or DDPMs, in the continuous-time limit. These models have a nonzero drift, which leads to qualitatively different time dependence: the distribution at the end of the forward process is expected to be a normal distribution, and hence all the higher-order cumulants should go to zero. We use a linear drift, with coefficient $k(t) = g^2(t)$. The qualitative features do not depend on the choice for $g(t)$, but for the numerics shown below we have taken $g(t) = \sigma^{t/T}$, with $\sigma = 10$, $T = 1$. The specific choice of $g(t)$ determines the time profile, via the definition

$$u(t) = \int_0^t ds g^2(s) = \frac{T}{\log \sigma^2} \left[\sigma^{2t/T} - 1 \right]. \quad (4.11)$$

Also in this case the analytical score is available and the time-dependent distribution reads

$$P(x, t) = \frac{1}{2} \left[\mathcal{N}(x; \mu(t), \sigma^2(t)) + \mathcal{N}(x; -\mu(t), \sigma^2(t)) \right], \quad (4.12)$$

where

$$\mu(t) = \mu_0 f(t, 0), \quad \sigma^2(t) = \sigma_0^2 f^2(t, 0) + \Xi(t), \quad (4.13)$$

with

$$f(t, s) = e^{-\frac{1}{2} \int_s^t ds' k(s')} = e^{-\frac{1}{2} u(t) + \frac{1}{2} u(s)}, \quad (4.14)$$

$$\Xi(t) = \int_0^t ds f^2(t, s) g^2(s) = 1 - f^2(t, 0). \quad (4.15)$$

The solution (4.12) describes the evolution of the distribution during the forward process. The proof is the same as above; with this distribution the second and higher cumulants evolve as

$$\kappa_2(t) = \mu^2(t) + \sigma^2(t) = (\mu_0^2 + \sigma_0^2 - 1) f^2(t, 0) + 1, \quad (4.16)$$

$$\kappa_{n>2}(t) = \kappa_n(0) f^n(t, 0), \quad (4.17)$$

as it should be. At $\tau = T$, $\kappa_2(T) \rightarrow 1$ and $\kappa_{n>2}(T) \rightarrow 0$, up to exponentially suppressed terms. The distribution then becomes normal, $P(x, T) = \mathcal{N}(x; 0, 1)$, again up to exponentially suppressed terms, see also App. B.

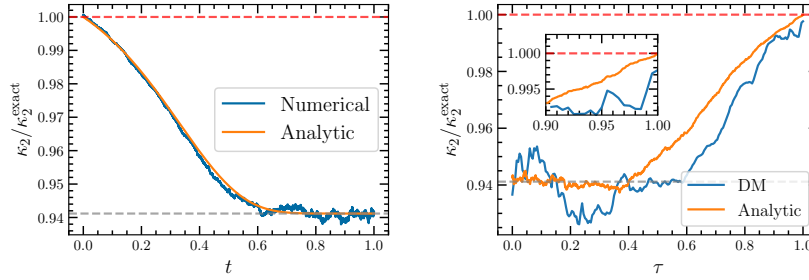


Figure 6: Evolution of the normalised 2nd cumulant, presented as $\kappa_2/\kappa_2^{\text{exact}}$, in the two-peak model in the DDPM, during the forward process (left) and during the backward process (right), with the score determined by the diffusion model and with the analytical score, using 10^6 trajectories in all cases. In the forward process, the analytical solution is shown as well. Other parameters as above.

The evolution of the second cumulant is shown in Fig. 6. Note that the cumulant is normalised with the target value, $\kappa_2(0) = \mu_0^2 + \sigma_0^2 = 1.0625$, such that the expected value at the end of the forward process is $1/1.0625 = 0.9412$. In the forward process we also show the analytical solution (4.16), to make clear that the observed noisy behaviour is due to the finite number of trajectories. For the backward process, we show the evolution using the learned score and the analytical score from the time-dependent solution (4.12), using an equal number of trajectories. The behaviours are initially somewhat different, but for $\tau/T > 0.5$ both processes converge towards the target value.

Higher-order cumulants are presented in Fig. 7, for the forward (top row) and backward (bottom row) process. The cumulants evolve from the target value to zero, and vice versa, in a much more controlled manner than in the variance-expanding scheme; the cancellations required above are not needed here. The results, with statistical error, are given in Table 1, fourth row. We note an agreement which is slightly worse compared to the variance-expanding scheme. This may be due to the qualitatively different behaviour observed during the evolution: while fluctuations are suppressed, there is a stronger dependence on the value obtained around, say, $\tau/T = 0.5$, to reach the expected value at $\tau = T$. It would be interesting to combine features of the two schemes, as the effects due to a finite number of trajectories affect the second cumulant and higher cumulants in opposite ways.

5 Lattice field theory

We now move to the case of many degrees of freedom and consider a Euclidean field theory, discretised on a lattice. For a real scalar field $\phi(x)$, the target probability distribution reads

$$P_0[\phi] = \frac{1}{Z} e^{-S[\phi]}, \quad Z = \int D\phi e^{-S[\phi]}, \quad (5.1)$$

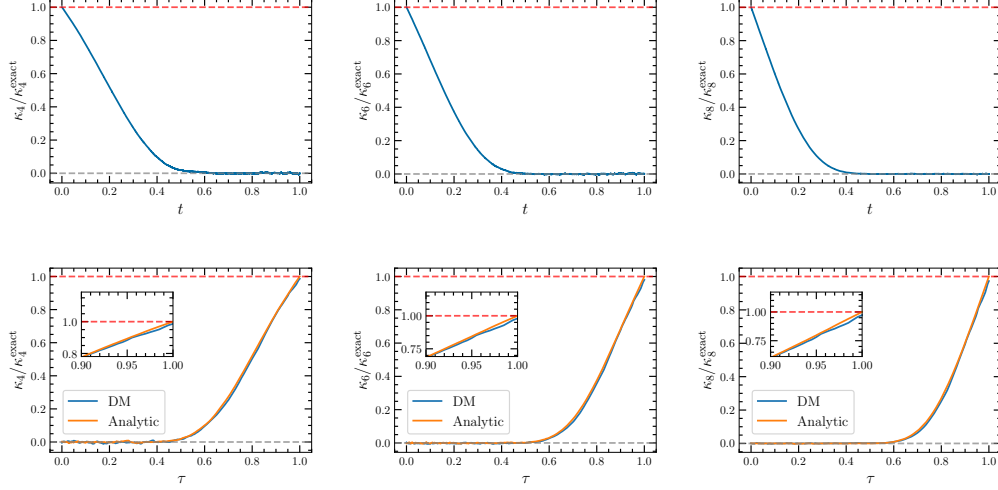


Figure 7: Evolution of the normalised 4th, 6th and 8th cumulants, presented as $\kappa_n/\kappa_n^{\text{exact}}$, in the two-peak model in the DDPM, during the forward process (above), and during the backward process (below), with the score determined by the diffusion model and with the analytical score, using 10^6 trajectories in all cases. Other parameters as above.

where $S[\phi]$ is the Euclidean action and Z denotes the partition function. The integration is over all field configurations and $D\phi$ denotes the discretised measure,

$$\int D\phi = \prod_x \int_{-\infty}^{\infty} d\phi_x, \quad (5.2)$$

where the product is over all spacetime points, forming a hypercubic (square) lattice in n (two) dimensions. The simplest interacting field theory is a $\lambda\phi^4$ theory, with continuum action

$$S[\phi] = \int d^n x \left[\frac{1}{2} \sum_{\nu=1}^n (\partial_\nu \phi(x))^2 + \frac{1}{2} m_0^2 \phi^2(x) + \frac{1}{4} \lambda_0 \phi^4(x) \right]. \quad (5.3)$$

As stated we use Euclidean signature, with ∂_ν ($\nu = 1, \dots, n$) denoting partial derivatives. Following the standard route to discretisation [9], the corresponding lattice action reads

$$S[\phi] = \sum_x \left[-2\kappa \sum_{\nu=1}^n \phi_x \phi_{x+\hat{\nu}} + \phi_x^2 + \lambda (\phi_x^2 - 1)^2 \right]. \quad (5.4)$$

We use periodic boundary conditions. The relation between the continuum and lattice fields, the lattice spacing, and the parameters m_0 , λ_0 and κ , λ can be found in Refs. [9, 28]. Note that κ is the so-called hopping parameter, not to be confused with a cumulant. The actions (5.3, 5.4) are non-Gaussian and hence lead to nonvanishing higher-order cumulants. Below we denote the field with $\phi(x)$.

Returning to the diffusion model, the equations discussed above for one degree of freedom can be taken over directly. The forward process reads

$$\partial_t \phi(x, t) = K[\phi(x, t), t] + g(t) \eta(x, t), \quad (5.5)$$

with the backward process

$$\partial_\tau \phi(x, \tau) = -K[\phi(x, \tau), T - \tau] + g^2(T - \tau) \nabla_\phi \log P(\phi, T - \tau) + g(T - \tau) \eta(x, \tau). \quad (5.6)$$

Here $K[\phi, t]$ is the possible drift term and $\eta \sim \mathcal{N}(0, 1)$ is Gaussian noise with variance 1, applied locally at each lattice coordinate, i.e.

$$\mathbb{E}_\eta[\eta(x, s) \eta(x', s')] = \delta(s - s') \delta(x - x'). \quad (5.7)$$

Note that $g(t)$ still only depends on time, but one could introduce x dependence as well. As above, we assume the first moment vanishes, or has been subtracted, $\phi(x) \rightarrow \phi(x) - \mathbb{E}_{P_0}[\phi(x)]$.

With a linear drift, $K[\phi(x, t), t] = -\frac{1}{2}k(t)\phi(x, t)$, and the initial condition $\phi_0 \sim P_0[\phi_0]$, the forward equation is solved by

$$\phi(x, t) = \phi_0(x)f(t, 0) + \int_0^t ds f(t, s)g(s)\eta(x, s), \quad (5.8)$$

where $f(t, s)$ was defined in Eq. (3.3). The equal-time two-point function (or two-point correlation function or propagator) then reads

$$G(x, y; t) \equiv \mathbb{E}[\phi(x, t)\phi(y, t)] = \mathbb{E}_{P_0}[\phi_0(x)\phi_0(y)]f^2(t, 0) + \Xi(t)\delta(x - y), \quad (5.9)$$

where

$$G_{\text{target}}(x, y) \equiv \mathbb{E}_{P_0}[\phi_0(x)\phi_0(y)] \quad (5.10)$$

is the full two-point function in the target theory. $\Xi(t)$ is exactly the same as in Eq. (3.7), having used Eq. (5.7).

Here we focus on moments and cumulants, involving products of the field at coinciding spacetime points – we come back to the propagator (5.9) below. Moments are then defined as

$$\mu_n(x, t) = \mathbb{E}[\phi^n(x, t)]. \quad (5.11)$$

Under the usual assumption that the target theory is translationally invariant, moments and cumulants are independent of x and the x -label may be dropped.

Since the noise is applied at each spacetime point separately, the computation for the moments and cumulants is exactly as before. We hence give immediately the results for the generation functionals. Moments are generated by

$$Z[J] = \mathbb{E}[e^{J(x, t)\phi(x, t)}] = e^{\frac{1}{2}J^2(x, t)\Xi(t)} \int D\phi_0 P_0[\phi_0] e^{J(x, t)\phi_0(x)f(t, 0)}, \quad (5.12)$$

and the cumulant-generating function reads

$$W[J] = \log Z[J] = \frac{1}{2}J^2(x, t)\Xi(t) + \log \int D\phi_0 P_0[\phi_0] e^{J(x, t)\phi_0(x)f(t, 0)}. \quad (5.13)$$

The second moment or cumulant is given by Eq. (5.9), evaluated at $x = y$.² All higher-order cumulants are given by

$$\kappa_{n>2}(t) = \frac{\delta^n W[J]}{\delta J(x, t)^n} \Big|_{J=0} = \frac{\delta^n}{\delta J(x, t)^n} \log \mathbb{E}_{P_0}[e^{J(x, t)\phi_0(x)f(t, 0)}] \Big|_{J=0}, \quad (5.14)$$

and are hence equal to the cumulants in the target theory, multiplied with the time-dependent function $f^n(t, 0)$. In particular, for pure diffusion we find again that

$$\kappa_{n>2}(t) = \kappa_n(0) \quad (\text{pure diffusion}). \quad (5.15)$$

We will now verify these results in the scalar field theory introduced above, defined on a two-dimensional lattice, using the implementation of the variance-expanding diffusion model without drift, previously discussed in this context in Refs. [28, 29]. The results shown below are obtained on a 32×32 lattice, with hopping parameter $\kappa = 0.4$ and coupling $\lambda = 0.022$. The theory is in the symmetric phase. We used 10^5 configurations to train the model using the variance-expanding scheme with $\sigma = 25$ and $T = 1$, and also 10^5 configurations to evolve the cumulants during the forward and backward processes.

Fig. 8 shows the 2nd and 4th cumulant, normalised with the numerically computed target value, during the forward (left) and backward (right) process. As expected, the 2nd moment or cumulant increases (decreases) as in Eq. (4.7). During the backward process, the second cumulant approaches the target result linearly from above, as expected, see App. B. The 4th cumulant is approximately constant, with the effect of a finite number of trajectories visible towards the end (start) of the forward (backward) process, as above. The forward and backward evolution of the 6th order cumulant is shown in Fig. 9. The cumulant is approximately conserved during the first (second) part of the forward (backward) evolution and suffers from incomplete cancellations for the remainder, similar to

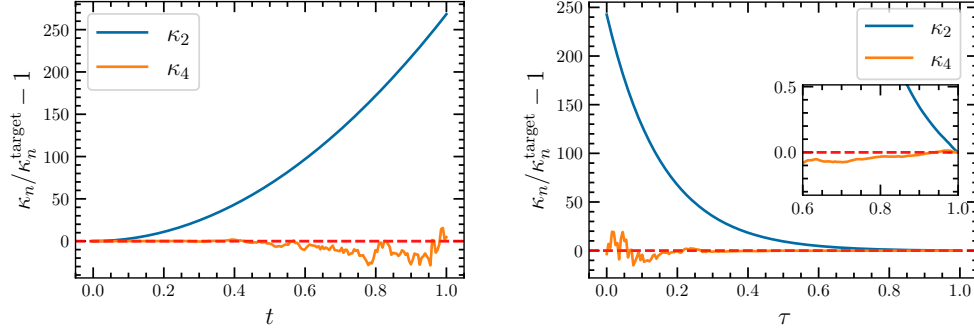


Figure 8: Evolution of the normalised 2nd and 4th cumulant, presented as $\kappa_n/\kappa_n^{\text{target}} - 1$, in the two-dimensional ϕ^4 theory, during the forward (left) and backward (right) process with the score determined by the diffusion model.

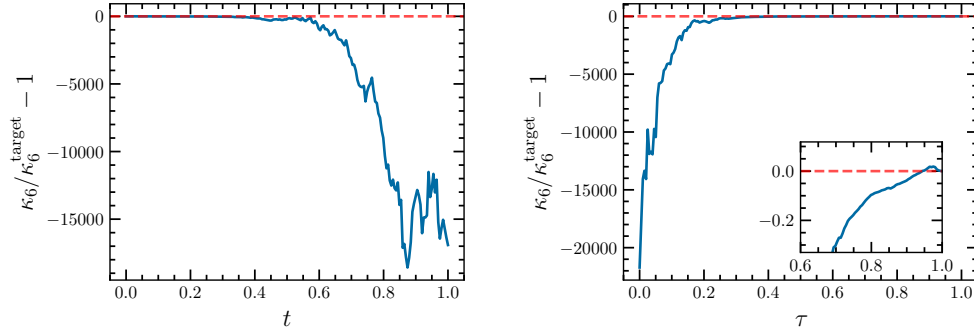


Figure 9: Evolution of the normalised 6th cumulant, presented as $\kappa_6/\kappa_6^{\text{target}} - 1$, in the two-dimensional ϕ^4 theory, during the forward (left) and backward (right) process with the score determined by the diffusion model.

	κ_2	κ_4	κ_6	κ_8
HMC (normalised)	0.39597(4)	-0.29453(6)	0.90108(28)	-5.8689(25)
Diffusion model	0.39598(4)	-0.29454(7)	0.90113(32)	-5.8694(28)

Table 2: First four nonvanishing cumulants κ_n in the scalar ϕ^4 field theory, with $\kappa = 0.4$, $\lambda = 0.022$ and 10^5 configurations on a 32^2 lattice, using normalised HMC data and as obtained from the diffusion model in the variance-expanding scheme. Statistical errors are computed by bootstrapping.

the evolution in the two-peak model. We note that with $32^2 \sim 10^3$ lattice sites, the total number of trajectories is $32^2 \times 10^5 \sim 10^8$. The insets show the evolution at the end of the backward process.

Estimates for the first four nonvanishing cumulants are presented in Table 2, with statistical uncertainty. We note here that the cumulant values are given for the normalised data. To revert to the unnormalised (i.e. original) data, the n^{th} cumulant κ_n has to be multiplied with the n^{th} power of $\phi_{\text{max}} = -\phi_{\text{min}}$, see App. A for details. Importantly, we observe that the cumulants are reproduced to high precision, including κ_2 . This illustrates that the diffusion model is capable of learning higher-order cumulants for a system with many degrees of freedom, such as a lattice field theory.

²The delta function should be understood as defined on the discretised lattice, $\delta(x - y) \rightarrow \delta_{x,y}$, where $\delta_{x,y}$ is the Kronecker delta with $\delta_{x,x} = 1$.

Returning finally to the propagator (5.9) in the case without a drift, $f(t, 0) = 1$, and using translational invariance, we note that it reads in momentum space

$$G(p, t) = G_{\text{target}}(p) + \Xi(t) = \frac{1}{p^2 + m^2 + \Sigma(p)} + \Xi(t). \quad (5.16)$$

Here we used the Dyson equation for the full propagator of the target theory,

$$G_{\text{target}}^{-1}(p) = G_0^{-1}(p) + \Sigma(p), \quad G_0^{-1}(p) = p^2 + m^2. \quad (5.17)$$

We note that during the forward process, with $\Xi(t)$ increasing, the most ultraviolet (large momentum) modes are destroyed first, and reversely, during the backward process, with $\Xi(t)$ decreasing, the most infrared (small momentum) modes are denoised first. The direction of this flow of information (from large spatial scales to small ones, or vice versa) is interesting to explore further, e.g. in the context of the Renormalisation Group (see e.g. Refs. [53–55] for connections between diffusion and (inverse) Renormalisation Group flows) or of nonequilibrium phase transitions and symmetry breaking/restoration (see e.g. Ref. [56] for phase transitions in diffusion models applied to ImageNet data).

6 Conclusion and outlook

In this paper we investigated how higher-order moments and cumulants are learnt in diffusion models, by deriving exact expressions for the moment- and cumulant-generating functionals. We have demonstrated that higher-order cumulants are exactly conserved in models without a drift, such as the variance-expanding scheme. The distribution at the end of the forward process is therefore as correlated as the target distribution. In models with a drift, such as DDPMs, higher-order cumulants go to zero and the distribution at the end of the forward process is normal, up to exponentially suppressed terms. In both cases, the score incorporates the knowledge of the higher-order correlations, which are therefore regenerated during the backward process, when starting from a normal distribution. These predictions were subsequently verified in an exactly-solvable but nontrivial model with one degree of freedom and in a lattice scalar field theory. The use of the latter is highly relevant for the application of diffusion models to generate field configurations as an alternative to standard Monte Carlo-based approaches. Since higher-order cumulants contain the information on interactions between fundamental degrees of freedom, it is indeed of utmost importance that these can be encoded and learnt properly.

In the numerical implementation of the variance-expanding scheme, we observed that at the final (initial) stages of the forward (backward) process the higher-order cumulants suffer from incomplete cancellations between trajectories and therefore appear “noisy”. We have demonstrated that this is not due to a poorly learned score but due to finite statistics. It will therefore be interesting to investigate how relevant the first stage of the backward process is in the generation of new configurations and whether efficiency can be gained by starting the backward process slightly later. For efficiency gains by adapting the noise scheduler, see e.g. Refs. [57–59]. We also note here that we have not incorporated any acceptance/reject step in this study, which is possible in principle [28] and a further area to explore in cases where minor deviations between the target and generated data need to be addressed. Finally, an analysis of the two-point function in field theory indicated an interesting interplay between momentum scales in the propagator and the scale of the noise during the forward and backward processes, which is worth exploring further, e.g. using the framework of the Renormalisation Group or nonequilibrium phase transitions.

Although the main thrust of our paper is theoretical, looking forward more generally we note that understanding the preservation and learning of complex correlations in diffusion models, as encoded in higher-order cumulants, might inform the development of more robust generative models across various domains. Additionally, modifying the training or sampling procedures of diffusion models to explicitly account for these may enhance their performance in capturing intricate data structures. Such advancements could lead to more accurate and efficient models in fields ranging from image and signal processing to the simulation of complex physical systems, including lattice field theory.

Acknowledgements – GA is supported by STFC Consolidated Grant ST/X000648/1. DEH is supported by the UKRI AIMLAC CDT EP/S023992/1. LW thanks the DEEP-IN working group at RIKEN-iTHEMS for its support in the preparation of this paper. KZ is supported by the CUHK-Shenzhen university development fund under grant No. UDF01003041 and UDF03003041, and Shenzhen Peacock fund under No. 2023TC0179.

We acknowledge the support of the Supercomputing Wales project, which is part-funded by the European Regional Development Fund (ERDF) via Welsh Government.

Research Data and Code Access – The code and data used for this manuscript are available from Ref. [60].

Open Access Statement – For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

A Implementation of the diffusion model

In this Appendix, we give some details on the numerical implementation of the diffusion model.

Estimating the score – To estimate the score of a distribution we may train a score-based model based on some sample dataset of the target distribution with score-matching [6]. The score function approximation $s_\theta(x, t)$ can be trained with the Fisher divergence objective such as in Ref. [4],

$$\mathcal{L}(\theta, \lambda) := \frac{1}{2} \int_0^T dt \mathbb{E}_{P_t(x)} \left[\lambda(t) \left\| s_\theta(x, t) - \nabla \log P_t(x) \right\|_2^2 \right], \quad (\text{A.1})$$

where the weight $\lambda(t)$ is chosen to be the variance of the noise at time t which, for $g(t) = \sigma^{t/T}$, equals

$$\lambda(t) = \frac{T}{2 \log \sigma} \left(\sigma^{2t/T} - 1 \right). \quad (\text{A.2})$$

A practical and computationally easier training objective is to approximate the score function of the marginal probability at some time t , $\nabla \log P_t(x)$, with the score of a transition kernel $\nabla \log P_t(x_t|x_0)$. In the case of an affine drift $K(x, t)$, the transition kernel is always a Gaussian distribution kernel such that $P_t(x_t|x_0) = \mathcal{N}(x_t; x_0, \lambda(t))$, and $P_t(x_t) = \int dx_0 P_t(x_t|x_0) P_{\text{data}}(x_0)$.

In a given epoch, we sample a random time step from the uniform distribution, $t \sim \mathcal{U}(\varepsilon, T)$, with ε close to 0, for every batch in the data set and perturb our sample x_0 with noise $\mathcal{N}(0, \lambda(t))$. For an affine drift, the score of the transition kernel in the training objective can be written as $\nabla \log P_t(x_t|x_0) = -(x_t - x_0)/\lambda(t)$, and using Eq. (A.1), we obtain the simplified loss function

$$\begin{aligned} \mathcal{L}(\theta, \lambda) &= \frac{1}{2} \sum_{t=0}^T \mathbb{E}_{P_t(x)} \left[\lambda(t) \left\| s_\theta(x, t) + \frac{x_t - x_0}{\lambda(t)} \right\|_2^2 \right] \\ &= \frac{1}{2} \sum_{t=0}^T \mathbb{E}_{P_t(x)} \left[\left\| s_\theta(x, t) \sqrt{\lambda(t)} + z_t \right\|_2^2 \right], \quad z_t \sim \mathcal{N}(0, 1). \end{aligned} \quad (\text{A.3})$$

Having a trained score model $s_\theta^*(x, t)$, we can obtain samples from the target distribution by numerically solving the backwards stochastic process, c.f. Eq. 2.2,

$$x_{\tau+\Delta\tau} = x_\tau + \left[-K(x_\tau, T - \tau) + g^2(T - \tau) s_\theta^*(x_\tau, T - \tau) \right] \Delta\tau + g(T - \tau) \sqrt{\Delta\tau} \eta_\tau, \quad (\text{A.4})$$

where $\Delta\tau$ is the step size, $\eta_\tau \sim \mathcal{N}(0, 1)$ and the equation is solved from $\tau = 0$ to T .

Two-peak model with one degree of freedom – To model the score $s_\theta(x, t)$, we use a fully connected neural network conditioned on the time information using Gaussian Fourier feature mapping [61]. For inference, we choose to run the backward process using 1000 steps for 10^6 trajectories. Our choice of hyperparameters is summarised in table 3.

Lattice field theory – Here we use the same setup as in Ref. [28], see App. A of that paper.

For training purposes, the HMC data has been normalised using the reversible transformation

$$\tilde{\phi}(x) = 2 \left(\frac{\phi(x) - \phi_{\min}}{\phi_{\max} - \phi_{\min}} - \frac{1}{2} \right), \quad \phi(x) = \frac{1}{2} \left(\tilde{\phi}(x) + 1 \right) (\phi_{\max} - \phi_{\min}) + \phi_{\min}, \quad (\text{A.5})$$

where $\phi_{\min, \max}$ are the minimal and maximal value of the field over the entire ensemble for fixed lattice parameters. For a symmetric distribution, n^{th} order moments and cumulants for unnormalised and normalised data are related via a multiplication or division by $\phi_{\max} = -\phi_{\min}$. For our ensemble $\phi_{\max} \sim -\phi_{\min} = 5.711$.

Model Hyperparameters		Training Hyperparameters	
Hyperparameter	Value	Hyperparameter	Value
Layers	[64, 64]	Learning Rate	1e-4
Time Embedding dims	128	Batch Size	512
Activation Function	LeakyReLU	Optimizer	Adam
Weight Initialization	PyTorch default	Max Epochs	200

Table 3: Model and training hyperparameters used in the training of the two-peak model. We save the weights with the best loss during the training process and set the training to stop early if the loss has not improved within 50 epochs. An early stop was observed occurring after an average of 100 epochs for a dataset of 10^6 realisations.

B Gaussian target distribution

In this Appendix we discuss the case of a Gaussian target distribution in some detail. Since the distribution is a Gaussian throughout the forward and backward process (with vanishing mean), the only dynamical quantity is the variance, i.e. the second moment or cumulant. It is important to distinguish the variances during the process:

- variance of the target distribution: σ_{target}^2 ,
- initial condition of the forward process: $\sigma_{\text{fw}}^2(0) = \sigma_{\text{target}}^2$,
- (final) variance during the forward process: $\sigma_{\text{fw}}^2(t), \sigma_{\text{fw}}^2(T)$,
- initial condition of the backward process: $\sigma_{\text{bw}}^2(0) = \sigma_{\text{target}}^2$,
- (final) variance during the backward process: $\sigma_{\text{bw}}^2(\tau), \sigma_{\text{bw}}^2(T)$.

A diffusion model works well if the final result of the backward process is (approximately) equal to the variance of the target distribution, $\sigma_{\text{bw}}^2(T) \sim \sigma_{\text{target}}^2$.

We consider the forward process

$$\dot{x}(t) = -\frac{1}{2}k(t)x(t) + g(t)\eta(t), \quad (\text{B.1})$$

where the drift includes the cases considered above: $k(t) = 0, g^2(t)$. The corresponding Fokker-Planck equation (FPE) reads

$$\partial_t P(x, t) = \frac{1}{2} \partial_x [g^2(t) \partial_x + k(t)x] P(x, t). \quad (\text{B.2})$$

With a linear drift, the solution is a Gaussian distribution,

$$P(x, t) = \frac{e^{-x^2/2\sigma_{\text{fw}}^2(t)}}{\sqrt{2\pi\sigma_{\text{fw}}^2(t)}}, \quad \sigma_{\text{fw}}^2(t) = \mathbb{E}_{P(x,t)}[x^2(t)] = \int dx P(x, t) x^2. \quad (\text{B.3})$$

Substituting this Ansatz in the FPE yields the equation $\sigma_{\text{fw}}^2(t)$ has to satisfy,

$$\dot{\sigma}_{\text{fw}}^2(t) = -k(t)\sigma_{\text{fw}}^2(t) + g^2(t). \quad (\text{B.4})$$

Using the notation in Sec. 3, the equation above is solved by

$$\sigma_{\text{fw}}^2(t) = \sigma_{\text{target}}^2 f^2(t, 0) + \Xi(t), \quad (\text{B.5})$$

in terms of the initial variance σ_{target}^2 , and

$$f(t, s) = e^{-\frac{1}{2} \int_s^t ds' k(s')}, \quad \Xi(t) = \int_0^t ds f^2(t, s) g^2(s). \quad (\text{B.6})$$

The backward process (with $\tau = T - t$) is

$$x'(\tau) = \frac{1}{2}k(T - \tau)x(\tau) + g^2(T - \tau)\partial_x \log P(x, T - \tau) + g(T - \tau)\eta(\tau). \quad (\text{B.7})$$

With

$$\partial_x \log P(x, t) = -x(t)/\sigma_{\text{fw}}^2(t), \quad (\text{B.8})$$

the drift for the backward process is linear, such that

$$x'(\tau) = -\frac{1}{2}k_{\text{bw}}(\tau)x(\tau) + g(T - \tau)\eta(\tau), \quad (\text{B.9})$$

with the coefficient

$$k_{\text{bw}}(\tau) = \frac{2g^2(T - \tau)}{\sigma_{\text{fw}}^2(T - \tau)} - k(T - \tau), \quad (\text{B.10})$$

and the corresponding FPE,

$$\partial_\tau P(x, \tau) = \frac{1}{2}\partial_x [g^2(T - \tau)\partial_x + k_{\text{bw}}(\tau)x] P(x, \tau). \quad (\text{B.11})$$

The solution is of the same form as for the forward process, see Eq. (B.5),

$$\sigma_{\text{bw}}^2(\tau) = \sigma_{0,\text{bw}}^2 f^2(\tau, 0) + \Xi(\tau), \quad (\text{B.12})$$

but with a more involved coefficient $k_{\text{bw}}(\tau)$ and

$$f(\tau, s) = e^{-\frac{1}{2} \int_s^\tau ds' k_{\text{bw}}(s')}, \quad \Xi(\tau) = \int_0^\tau ds f^2(\tau, s) g^2(T - s). \quad (\text{B.13})$$

We now consider two special cases:

1. Variance-expanding scheme without a drift term. We take $k(t) = 0, g(t) = \sigma^{t/T}$, such that $f(t, s) = 1$ and

$$\sigma_{\text{fw}}^2(t) = \sigma_{\text{target}}^2 + \int_0^t ds g^2(s) = \sigma_{\text{target}}^2 + \frac{T}{2 \log \sigma} [\sigma^{2t/T} - 1]. \quad (\text{B.14})$$

For $t/T \ll 1$, this reads $\sigma_{\text{fw}}^2(t) = \sigma_{\text{target}}^2 + t$, a linear increase in time. The memory of the initial distribution is suppressed at $t = T$, provided that $D\sigma^2 \gg \sigma_{\text{target}}^2$ and large, such that

$$\sigma_{\text{fw}}^2(T) \approx D\sigma^2, \quad \text{with} \quad D = \frac{g^2(t)}{dg^2(t)/dt} = \frac{T}{\log \sigma^2}. \quad (\text{B.15})$$

The solution (B.12) of the backward process has the somewhat cumbersome coefficients

$$f(\tau, 0) = \frac{D(\sigma^{2(1-\tau/T)} - 1) + \sigma_{\text{target}}^2}{D(\sigma^2 - 1) + \sigma_{\text{target}}^2}, \quad (\text{B.16})$$

$$\Xi(\tau) = D\sigma^{2(1-\tau/T)} (\sigma^{2\tau/T} - 1) f(\tau, 0). \quad (\text{B.17})$$

Since $f(0, 0) = 1$ and $\Xi(0) = 0$, the solution satisfies the correct initial condition. At the end of the backward process, we find

$$f(T, 0) = \frac{\sigma_{\text{target}}^2}{D(\sigma^2 - 1) + \sigma_{\text{target}}^2}, \quad \Xi(T) = D(\sigma^2 - 1) f(T, 0). \quad (\text{B.18})$$

In the same limit as above, $D\sigma^2 \gg \sigma_{\text{target}}^2$ and large, one finds that

$$f(T, 0) = 0 + \frac{\sigma_{\text{target}}^2}{D(\sigma^2 - 1)} + \dots, \quad (\text{B.19})$$

$$\Xi(T) = \sigma_{\text{target}}^2 \left[1 - \frac{\sigma_{\text{target}}^2}{D(\sigma^2 - 1)} + \dots \right], \quad (\text{B.20})$$

such that the desired outcome is indeed obtained, $\sigma_{\text{bw}}^2(T) \approx \sigma_{\text{target}}^2$.

Writing $\tau = T - \epsilon$, the approach to the value at $\tau = T$ can be analysed, using

$$f(T - \epsilon, 0) = f(T, 0) \left[1 + \frac{\epsilon}{\sigma_{\text{target}}^2} + \mathcal{O}(\epsilon^2) \right], \quad (\text{B.21})$$

$$\Xi(T - \epsilon) = \Xi(T) \frac{f(T - \epsilon, 0)}{f(T, 0)} \left[1 - \frac{\epsilon^2}{D} + \mathcal{O}(\epsilon^3) \right], \quad (\text{B.22})$$

which yields

$$\sigma_{\text{bw}}^2(T - \epsilon) = \sigma_{\text{bw}}^2(T) \left[1 + \frac{\Xi(T) + 2f^2(T, 0)\sigma_{0,\text{bw}}^2}{\Xi(T) + f^2(T, 0)\sigma_{0,\text{bw}}^2} \frac{\epsilon}{\sigma_{\text{target}}^2} + \mathcal{O}(\epsilon^2) \right]. \quad (\text{B.23})$$

Using again that $D\sigma^2 \gg \sigma_{\text{target}}^2$, this reduces to

$$\sigma_{\text{bw}}^2(T - \epsilon) \approx \sigma_{\text{bw}}^2(T) + \frac{\sigma_{\text{bw}}^2(T)}{\sigma_{\text{target}}^2} \epsilon + \mathcal{O}(\epsilon^2) \approx \sigma_{\text{bw}}^2(T) + \epsilon + \mathcal{O}(\epsilon^2), \quad (\text{B.24})$$

i.e. the target value is approached linearly with slope 1, as expected from the forward process at early times.

2. Denoising diffusion probabilistic models (DDPMs). We take $k(t) = g^2(t)$, which incorporates various examples of DDPMs in the continuous-time limit. The FPE simplifies considerably,

$$\partial_t P(x, t) = \frac{1}{2} g^2(t) \partial_x (\partial_x + x) P(x, t), \quad (\text{B.25})$$

which suggests to redefine time as

$$u(t) = \int_0^t ds g^2(s), \quad (\text{B.26})$$

such that

$$\partial_u P(x, u) = \frac{1}{2} \partial_x (\partial_x + x) P(x, u). \quad (\text{B.27})$$

With

$$f(t, s) = e^{-\frac{1}{2}u(t) + \frac{1}{2}u(s)}, \quad \Xi(t) = 1 - f^2(t, 0), \quad (\text{B.28})$$

the variance during the forward process is given by

$$\sigma_{\text{fw}}^2(t) = 1 + e^{-u(t)} (\sigma_{\text{target}}^2 - 1), \quad (\text{B.29})$$

with $\sigma_{\text{fw}}^2(0) = \sigma_{\text{target}}^2$. At the end of the forward process, the variance is unity, up to exponentially suppressed terms.

For the backward process, we introduce

$$v(\tau) = \int_0^\tau ds g^2(T - s) = u(T) - u(T - \tau). \quad (\text{B.30})$$

We then find

$$f(\tau, s) = e^{\frac{1}{2}v(s) - \frac{1}{2}v(\tau)} \frac{e^{v(T)} + ce^{v(\tau)}}{e^{v(T)} + ce^{v(s)}}, \quad (\text{B.31})$$

$$\Xi(\tau) = \left(1 - e^{-v(\tau)} \right) \frac{1 + ce^{v(\tau) - v(T)}}{1 + ce^{-v(T)}}, \quad (\text{B.32})$$

with $c = \sigma_{\text{target}}^2 - 1$. Inserting these in Eq. (B.12) yields the solution of the backward process. At the end of the backward process, we obtain

$$f(T, 0) = \frac{e^{-\frac{1}{2}v(T)}}{1 + ce^{-v(T)}} \sigma_{\text{target}}^2 \rightarrow 0, \quad \Xi(T) = \frac{1 - e^{-v(T)}}{1 + ce^{-v(T)}} \sigma_{\text{target}}^2 \rightarrow \sigma_{\text{target}}^2, \quad (\text{B.33})$$

which implies that the desired outcome is again obtained, $\sigma_{\text{bw}}^2(T) \approx \sigma_{\text{target}}^2$, up to exponentially suppressed terms. Note that this is independent of the choice of $g(t)$.

We conclude that in the case of a Gaussian target distribution both schemes lead to the correct result, by explicit computation. Noticeably, the manner in which this is achieved is quite different: in the scheme without a drift, the variance at the end of the forward process should become large, and in particular much larger than target variance. Corrections are suppressed as $\sigma_{\text{target}}^2/\sigma^2$, where σ^2 is the strength of the noise at $t = T$. In the scheme with a drift this requirement is not needed: the variance at the end of the forward process becomes unity and deviations are suppressed exponentially.

References

- [1] J. Sohl-Dickstein, E.A. Weiss, N. Maheswaranathan and S. Ganguli, *Deep unsupervised learning using nonequilibrium thermodynamics*, in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.* - Vol. 37, pp. 2256–2265, 2015 [1503.03585].
- [2] J. Ho, A. Jain and P. Abbeel, *Denoising diffusion probabilistic models*, in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, pp. 6840–6851, 2020 [2006.11239].
- [3] Y. Song and S. Ermon, *Generative Modeling by Estimating Gradients of the Data Distribution*, 1907.05600.
- [4] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon and B. Poole, *Score-based generative modeling through stochastic differential equations*, 2011.13456.
- [5] L. Yang, Z. Zhang, S. Hong, R. Xu, Y. Zhao, Y. Shao et al., *Diffusion Models: A Comprehensive Survey of Methods and Applications*, 2209.00796.
- [6] A. Hyvärinen, *Estimation of Non-Normalized Statistical Models by Score Matching*, *Journal of Machine Learning Research* **6** (2005) 695.
- [7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, *High-Resolution Image Synthesis with Latent Diffusion Models*, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022) 10684 [2112.10752].
- [8] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu and M. Chen, *Hierarchical Text-Conditional Image Generation with CLIP Latents*, 2204.06125.
- [9] J. Smit, *Introduction to quantum fields on a lattice: A robust mate*, vol. 15, Cambridge University Press (1, 2011).
- [10] C. Gattringer and C.B. Lang, *Quantum chromodynamics on the lattice*, vol. 788, Springer, Berlin (2010), 10.1007/978-3-642-01850-3.
- [11] M. Creutz, *Monte Carlo Study of Quantized SU(2) Gauge Theory*, *Phys. Rev. D* **21** (1980) 2308.
- [12] S. Duane, A.D. Kennedy, B.J. Pendleton and D. Roweth, *Hybrid Monte Carlo*, *Phys. Lett. B* **195** (1987) 216.
- [13] ALPHA collaboration, *Critical slowing down and error analysis in lattice QCD simulations*, *Nucl. Phys. B* **845** (2011) 93 [1009.5228].
- [14] K. Cranmer, G. Kanwar, S. Racanière, D.J. Rezende and P.E. Shanahan, *Advances in machine-learning-based sampling motivated by lattice quantum chromodynamics*, *Nature Rev. Phys.* **5** (2023) 526 [2309.01156].
- [15] D.J. Rezende and S. Mohamed, *Variational Inference with Normalizing Flows*, in *International conference on machine learning*, pp. 1530–1538, PMLR, 2015 [1505.05770].
- [16] F. Noé, S. Olsson, J. Köhler and H. Wu, *Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning*, *Science* **365** (2019) eaaw1147 [1812.01729].
- [17] M.S. Albergo, G. Kanwar and P.E. Shanahan, *Flow-based generative models for Markov chain Monte Carlo in lattice field theory*, *Phys. Rev. D* **100** (2019) 034515 [1904.12072].
- [18] K.A. Nicoli, S. Nakajima, N. Strodthoff, W. Samek, K.-R. Müller and P. Kessel, *Asymptotically unbiased estimation of physical observables with neural samplers*, *Phys. Rev. E* **101** (2020) 023304 [1910.13496].
- [19] G. Kanwar, M.S. Albergo, D. Boyda, K. Cranmer, D.C. Hackett, S. Racanière et al., *Equivariant Flow-Based Sampling for Lattice Gauge Theory*, *Phys. Rev. Lett.* **125** (2020) 121601 [2003.06413].
- [20] K.A. Nicoli, C.J. Anders, L. Funcke, T. Hartung, K. Jansen, P. Kessel et al., *Estimation of thermodynamic observables in lattice field theories with deep generative models*, *Phys. Rev. Lett.* **126** (2021) 032001 [2007.07115].
- [21] K.A. Nicoli, C.J. Anders, T. Hartung, K. Jansen, P. Kessel and S. Nakajima, *Detecting and mitigating mode-collapse for flow-based sampling of lattice field theories*, *Phys. Rev. D* **108** (2023) 114501 [2302.14082].
- [22] R.T. Chen, Y. Rubanova, J. Bettencourt and D.K. Duvenaud, *Neural ordinary differential equations*, *Advances in neural information processing systems* **31** (2018) [1806.07366].

- [23] P. de Haan, C. Rainone, M.C.N. Cheng and R. Bondesan, *Scaling Up Machine Learning For Quantum Field Theory with Equivariant Continuous Flows*, 2110.02673.
- [24] M. Gerdes, P. de Haan, C. Rainone, R. Bondesan and M.C.N. Cheng, *Learning lattice quantum field theories with equivariant continuous flows*, *SciPost Phys.* **15** (2023) 238 [2207.00283].
- [25] M. Caselle, E. Cellini and A. Nada, *Sampling the lattice Nambu-Goto string using Continuous Normalizing Flows*, *Journal of High Energy Physics* **02** (2024) 048 [2307.01107].
- [26] H. Wu, J. Köhler and F. Noé, *Stochastic normalizing flows*, *Advances in Neural Information Processing Systems* **33** (2020) 5933 [2002.06707].
- [27] M. Caselle, E. Cellini, A. Nada and M. Panero, *Stochastic normalizing flows as non-equilibrium transformations*, *JHEP* **07** (2022) 015 [2201.08862].
- [28] L. Wang, G. Aarts and K. Zhou, *Diffusion models as stochastic quantization in lattice field theory*, *JHEP* **05** (2024) 060 [2309.17082].
- [29] L. Wang, G. Aarts and K. Zhou, *Generative Diffusion Models for Lattice Field Theory*, in *37th Conference on Neural Information Processing Systems*, 11, 2023 [2311.03578].
- [30] Q. Zhu, G. Aarts, W. Wang, K. Zhou and L. Wang, *Diffusion models for lattice gauge field simulations*, in *38th conference on Neural Information Processing Systems*, 2024 [2410.19602].
- [31] D.S. Berman, M.S. Klinger and A.G. Stapleton, *Ncoder – a quantum field theory approach to encoding data*, 2402.00944.
- [32] J. Halverson, A. Maiti and K. Stoner, *Neural Networks and Quantum Field Theory*, *Mach. Learn. Sci. Tech.* **2** (2021) 035002 [2008.08601].
- [33] M. Demirtas, J. Halverson, A. Maiti, M.D. Schwartz and K. Stoner, *Neural network field theories: non-Gaussianity, actions, and locality*, *Mach. Learn. Sci. Tech.* **5** (2024) 015002 [2307.03223].
- [34] R.M. Neal, *Bayesian learning for neural networks*, PhD Thesis, University of Toronto, 1995 (1995).
- [35] J. Lee, Y. Bahri, R. Novak, S.S. Schoenholz, J. Pennington and J. Sohl-Dickstein, *Deep Neural Networks as Gaussian Processes*, 1711.00165.
- [36] A.G. de G. Matthews, M. Rowland, J. Hron, R.E. Turner and Z. Ghahramani, *Gaussian Process Behaviour in Wide Deep Neural Networks*, 1804.11271.
- [37] G. Yang, *Tensor Programs I: Wide Feedforward or Recurrent Neural Networks of Any Architecture are Gaussian Processes*, 1910.12478.
- [38] K. Hashimoto, Y. Hirono, J. Maeda and J. Totsuka-Yoshinaka, *Neural network representation of quantum systems*, *Mach. Learn. Sci. Tech.* **5** (2024) 045039 [2403.11420].
- [39] K. Hashimoto, S. Sugishita, A. Tanaka and A. Tomiya, *Deep learning and the AdS/CFT correspondence*, *Phys. Rev. D* **98** (2018) 046019 [1802.08313].
- [40] D. Bachtis, G. Aarts and B. Lucini, *Quantum field-theoretic machine learning*, *Phys. Rev. D* **103** (2021) 074510 [2102.09449].
- [41] G. Aarts, B. Lucini and C. Park, *Scalar field restricted Boltzmann machine as an ultraviolet regulator*, *Phys. Rev. D* **109** (2024) 034521 [2309.15002].
- [42] S. Krippendorff and M. Spannowsky, *A duality connecting neural network and cosmological dynamics*, *Mach. Learn. Sci. Tech.* **3** (2022) 035011 [2202.11104].
- [43] G. Aarts, B. Lucini and C. Park, *Stochastic weight matrix dynamics during learning and Dyson Brownian motion*, *Phys. Rev. E* **111** (2025) 015303 [2407.16427].
- [44] G. Parisi and Y.S. Wu, *Perturbation theory without gauge fixing*, *Sci. China, A* **24** (1980) 483.
- [45] P.H. Damgaard and H. Hüffel, *Stochastic quantization*, *Phys. Rept.* **152** (1987) 227.
- [46] M. Namiki, *Basic ideas of stochastic quantization*, *Prog. Theor. Phys. Suppl.* **111** (1993) 1.
- [47] Y. Hirono, A. Tanaka and K. Fukushima, *Understanding Diffusion Models by Feynman’s Path Integral*, 2403.11262.
- [48] Y. Song and S. Ermon, *Improved techniques for training score-based generative models*, 2006.09011.

- [49] Y. Song, C. Durkan, I. Murray and S. Ermon, *Maximum likelihood training of score-based diffusion models*, 2101.09258.
- [50] T. Karras, M. Aittala, T. Aila and S. Laine, *Elucidating the design space of diffusion-based generative models*, 2206.00364.
- [51] P. Nakkiran, A. Bradley, H. Zhou and M. Advani, *Step-by-Step Diffusion: An Elementary Tutorial*, 2406.08929.
- [52] Peter J. Smith, *A Recursive Formulation of the Old Problem of Obtaining Moments from Cumulants and Vice Versa*, *The American Statistician* **49** (1995) 217.
- [53] J. Cotler and S. Rezhikov, *Renormalizing Diffusion Models*, 2308.12355.
- [54] D.S. Berman and M.S. Klinger, *The Inverse of Exact Renormalization Group Flows as Statistical Inference*, *Entropy* **26** (2024) 389 [2212.11379].
- [55] D.S. Berman, M.S. Klinger and A.G. Stapleton, *Bayesian renormalization*, *Mach. Learn. Sci. Tech.* **4** (2023) 045011 [2305.10491].
- [56] A. Sclocchi, A. Favero and M. Wyart, *A Phase Transition in Diffusion Models Reveals the Hierarchical Nature of Data*, 2402.16991.
- [57] T. Chen, *On the Importance of Noise Scheduling for Diffusion Models*, 2301.10972.
- [58] T. Hang and S. Gu, *Improved Noise Schedule for Diffusion Training*, 2407.03297.
- [59] K. Ikeda, T. Uda, D. Okanojara and S. Ito, *Speed-accuracy trade-off for the diffusion models: Wisdom from nonequilibrium thermodynamics and optimal transport*, 2407.04495.
- [60] D.E. Habibi, G. Aarts, L. Wang and K. Zhou,
“DiaaEddinH/On-learning-higher-order-cumulants-in-diffusion-models: v1.0.2.”
10.5281/zenodo.14041604.
- [61] M. Tancik, P.P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal et al., *Fourier features let networks learn high frequency functions in low dimensional domains*, 2006.10739.