Stochastic quantization and diffusion models

Kenji Fukushima and Syo Kamata

Department of Physics, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

This is a pedagogical review for the possible connection between the stochastic quantization in physics and the diffusion models in machine learning. For machine-learning applications, the denoising diffusion model has been established as a successful technique, which is formulated in terms of the stochastic differential equation (SDE). In this review, we focus on an SDE approach used in the score-based generative modeling. Interestingly, the evolution of the probability distribution is equivalently described by a particular class of SDEs, and in a particular limit, the stochastic noises can be eliminated. Then, we turn to a similar mathematical formulation in quantum physics, that is, the stochastic quantization. We make a brief overview on the stochastic quantization using a simple toy model of the one-dimensional integration. The analogy between the diffusion model and the stochastic quantization is clearly seen in this concrete example. Finally, we discuss how the sign problem arises in the toy model with complex parameters. The origin of the difficulty is understood based on the Lefschetz thimble analysis. We point out that the SDE is not invariant under the variable change which induces a kernel and a special choice of the kernel guided by the Lefschetz thimble analysis can reduce the sign problem.

1. Introduction

Generative modeling is widely used for practical applications, among which denoising diffusion probabilistic models (DDPMs)[1] based on physical processes of non-equilibrium dynamics[2] appear to be a physics-friendly formulation with the Langevin equation or the stochastic differential equation (SDE). In particular, the approach with score matching[3–5] further extends a physics intuition for denoising processes. The interesting observation is that the reverse of the stochastic noising process also follows the SDE once the score matching is achieved. For a proof of the reverse process in the language of physics, see a recent work.[6]

The denoising process allows for sampling of generated data. It is shown[7] that a wider class of diffusion processes can lead to the equivalent distribution of sampling data, which even include an ordinary differential equation (ODE) that is fully deterministic. Clearly, such a deterministic formulation of denoising diffusion implicit models (DDIMs)[7] can perform sampling much faster. To understand the equivalence, as we will review later, the crucial point is that the probability of samples governed by the Langevin equation should evolve with the Fokker-Planck equation. Therefore, if the same Fokker-Planck equation is derived from a class of SDEs with free parameters, in principle, the generated data should exhibit the same quality.

In this review, we pay attention to an analogy between the diffusion model and the stochastic quantization in physics[8] and discuss a potential interplay. It should be noted that the idea to accelerate configuration generation in lattice field theory has been tested within the generative diffusion model.[9] Then, it has been clearly recognized that the diffusion model can be interpreted as the probability evolution in the stochastic quantization.

In physics, the quantum effect is often called "fluctuation" and in the quantization procedure such quantum fluctuations are integrated out. Then, one may naturally be tempted to associate a Brownian-type motion with the quantum effect. The reformulation of the Schrödinger equation in terms of classical noises is dated back to a seminal work[10] more than half a century ago. In general, however, it is impossible to map all the quantum effect to classical fluctuations unless an extra dimension along the quantum axis is introduced, which underlies a general idea of holographic principle. Such a duality between a quantum theory in d dimensions and a classical theory in $(d+1)$ dimensions is not limited to the gauge-gravity correspondence, but many useful examples can be found in various contexts, such as the renormalization group flow. The stochastic quantization belongs to this category of machinery to quantize theories with classical variables with the quantum axis in an extra dimension.

The key equation in the stochastic quantization is the SDE and the mathematical structures are quite similar to the DDPMs. Thus, it is an intriguing direction of research to think about the mutual interplay between two approaches. Such an attempt has been just launched recently,[9] and this article is expected to serve as a starting point for further productive interactions of two formulations in different communities.

2. A Review of Denoising Diffusion Probabilistic Models

We make a brief review of the DDPMs, especially the score-matching modeling according to the standard literature.[5] The forward noising diffusion is described by the following Langevin equation or the SDE:

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, t)\mathrm{d}t + g(t)\,\mathrm{d}\boldsymbol{w}_t\,. \tag{1}$$

Here, $\boldsymbol{f}(\boldsymbol{x}_t, t)$ represents the drift term and $g(t)$ is the diffusion coefficient. In general $g(t)$ can take a matrix structure but we treat it as a one-component function for simplicity. The last term involves a stochastic variable $\mathrm{d}\boldsymbol{w}_t$ that is the Wiener process. Importantly, both $\boldsymbol{f}(\boldsymbol{x}_t, t)$ and $g(t)$ are time-dependent so that the distribution of $\boldsymbol{x}_t$ converges to the normal distribution after all. In the physics notation, a more familiar form of Eq. (1) would be

$$\dot{\boldsymbol{x}}_t = \boldsymbol{f}(\boldsymbol{x}_t, t) + g(t)\,\boldsymbol{\xi}_t\,. \tag{2}$$

Here, $\mathrm{d}\boldsymbol{w}_t = \boldsymbol{\xi}_t \mathrm{d}t$ and $\mathrm{d}\boldsymbol{w}_t^2 = \mathrm{d}t$ symbolically in the Itô calculus, and $\dot{\boldsymbol{x}}_t$ represents the time-derivative of $\boldsymbol{x}_t$. We note that t will turn out to be a fictitious time in later discussions about the stochastic quantization.

It is convenient to establish a relation between the SDE and the evolution equation for the probability distribution that $\boldsymbol{x}_t$ follows. The evolution equation is known as the Fokker-Planck-Kolmogorov (FPK) equation, which is commonly called just the Fokker-Planck equation in physics or the forward Kolmogorov equation in stochastic theory. The explanation we summarize below is based on the argument in the review[8] in which the Stratonovich calculus is assume to use the ordinary chain rule. To see the correspondence to the Fokker-Planck equation, we consider the average of some function $A(\boldsymbol{x}_t)$ with respect to the noise $\boldsymbol{\xi}_t$ as denoted in physics by

$$\langle A(\boldsymbol{x}_t)\rangle := \int D\boldsymbol{\xi}\, A(\boldsymbol{x}_t)\, \exp\Big(-\frac{1}{2}\int \|\boldsymbol{\xi}_t\|^2 \mathrm{d}t\Big), \tag{3}$$

where $\boldsymbol{x}_t$ is a solution of Eq. (2) and thus $\boldsymbol{\xi}_t$ dependence is implicitly implemented through $\boldsymbol{x}_t$. Using this, we can also introduce the probability $p_t(\boldsymbol{x})$ from

$$\langle A(\boldsymbol{x}_t)\rangle =: \int \mathrm{d}\boldsymbol{x}\, A(\boldsymbol{x}) p_t(\boldsymbol{x})\,. \tag{4}$$

The right-hand size of the above is often denoted as $\mathbb{E}_{p_t}[A(\boldsymbol{x})]$. We note that the t-dependence is incorporated only in $p_t(\cdot)$ and $\boldsymbol{x}$ is just the integration variable. Now, our task is to find an equation to describe the time-evolution of $p_t(\boldsymbol{x})$ that is consistent with Eq. (2). The time derivative of $\langle A(\boldsymbol{x}_t)\rangle$ is immediately given by

$$\begin{aligned}\int \mathrm{d}\boldsymbol{x}\, A(\boldsymbol{x})\, \dot{p}_t(\boldsymbol{x}) &= \Big\langle \frac{\partial A(\boldsymbol{x}_t)}{\partial \boldsymbol{x}_t}\cdot \dot{\boldsymbol{x}}_t \Big\rangle \\ &= \Big\langle \frac{\partial A(\boldsymbol{x}_t)}{\partial \boldsymbol{x}_t}\cdot \big(\boldsymbol{f}(\boldsymbol{x}_t,t) + g(t)\,\boldsymbol{\xi}_t\big)\Big\rangle\,. \end{aligned}\tag{5}$$

From the definition (3), we can see:

$$\Big\langle \frac{\partial A(\boldsymbol{x}_t)}{\partial \boldsymbol{x}_t}\cdot \boldsymbol{\xi}_t \Big\rangle = \Big\langle \frac{\partial}{\partial \boldsymbol{\xi}_t}\cdot \frac{\partial A(\boldsymbol{x}_t)}{\partial \boldsymbol{x}_t}\Big\rangle = \Big\langle \frac{\partial \boldsymbol{x}_t}{\partial \boldsymbol{\xi}_t}\frac{\partial^2 A(\boldsymbol{x}_t)}{\partial \boldsymbol{x}_t^2}\Big\rangle\,. \tag{6}$$

Here, we used the integration by part so that $\boldsymbol{\xi}_t$ comes down from the exponential function. Using $\partial \boldsymbol{x}_t/\partial \boldsymbol{\xi}_t = g(t)/2$, where $1/2$ appears from $\theta(0) = 1/2$ with the Heavisite step function, we arrive at

$$\begin{aligned}&\int \mathrm{d}\boldsymbol{x}\, A(\boldsymbol{x})\, \dot{p}_t(\boldsymbol{x}) \\ &= \int \mathrm{d}\boldsymbol{x}\, A(\boldsymbol{x}) \frac{\partial}{\partial \boldsymbol{x}}\cdot\Big(-\boldsymbol{f}(\boldsymbol{x},t) + \frac{g(t)^2}{2}\frac{\partial}{\partial \boldsymbol{x}}\Big) p_t(\boldsymbol{x})\,. \end{aligned}\tag{7}$$

The above is a heuristic argument for physicists, while a more secure derivation utilizes the Itô's formula. In any case, because this expression should hold for any function $A(\boldsymbol{x})$, we conclude the following Fokker-Planck equation:

$$\dot{p}_t(\boldsymbol{x}) = -\boldsymbol{\nabla}\cdot\Big[\boldsymbol{f}(\boldsymbol{x},t)\, p_t(\boldsymbol{x}) - \frac{g(t)^2}{2}\boldsymbol{\nabla} p_t(\boldsymbol{x})\Big]. \tag{8}$$

Here, we use $\boldsymbol{\nabla}$ instead of $\partial/\partial \boldsymbol{x}$ for notational brevity. It is clear that the Langevin equation (2) and the Fokker-Planck equation (8) have equivalent physical contents. Nevertheless, solving the Fokker-Planck equation numerically demands huge computational costs, and the Langevin equation is more tractable. We comment that the Fokker-Planck equation appears in a wide range of physical systems including even the high-energy scattering process of quarks and gluons,[11] for which the Fokker-Planck equation in non-Abelian group space called the JIMWLK equation is an established theoretical tool. It is impossibly difficult to solve such a complicated functional equation, but the equivalent rewriting to the Langevin equation paves a path for feasible numerical simulations.[12,13] We leave this comment with a hope that the generative modeling could be useful even for such resummation programs in high-energy small-x physics.

The Fokker-Planck equation tells us a condition for the drift and the noise terms. If we naïvely disturb $\boldsymbol{x}_t$ with noise, the distribution of $\boldsymbol{x}_t$ may simply spread widely. To make the reverse process well organized, we require an asymptotic form at $t \to \infty$ to take the normal distribution, i.e., $p_{t\to\infty}(\boldsymbol{x}) \propto e^{-\|\boldsymbol{x}\|^2/(2\sigma^2)}$. We can plug this form of the normal distribution into the Fokker-Planck equation, and then we can deduce the condition; we should choose $\boldsymbol{f}(\boldsymbol{x},t)$ and $g(t)$ such that $\lim_{t\to\infty} \boldsymbol{f}(\boldsymbol{x},t)/g(t)^2 = -\boldsymbol{x}/(2\sigma^2)$. For the normal distribution with $\sigma^2 = 1$, the simple choice of time-dependent coefficients, i.e., the SDE schedulings is:

$$\boldsymbol{f}(\boldsymbol{x},t) = f(t)\,\boldsymbol{x} = -\frac{\beta}{2}t\,\boldsymbol{x}\,, \qquad g(t) = \sqrt{\beta t}\,. \tag{9}$$

Technically, it is notable that the drift term, $\boldsymbol{f}(\boldsymbol{x},t)$, is a linear function of $\boldsymbol{x}$, i.e., it is *affine*. This choice enables us to derive some useful analytical formulas. We can arbitrarily take β which just controls the scale of time evolution. In this work, we fix $\beta = 20$ following the convention.[6] The important feature of the affine drift term is that the mean, $\boldsymbol{m}(t)$, and the variance, $\sigma(t)^2$, of the conditional probability, $p_{t|0}(\boldsymbol{x}|\boldsymbol{x}_0)$, is solved. From the Fokker-Planck equation (8), it is straightforward to derive the following differential equations:

$$\dot{\boldsymbol{m}} = \mathbb{E}[\boldsymbol{f}(\boldsymbol{x},t)] = f(t)\,\boldsymbol{m}\,, \tag{10}$$

$$\dot{\sigma^2} = 2\mathbb{E}[\boldsymbol{f}(\boldsymbol{x},t)\cdot(\boldsymbol{x}-\boldsymbol{m})] + g(t)^2 = 2f(t)\,\sigma^2 + g(t)^2\,. \tag{11}$$

The solution of above two equations is found to be

$$\boldsymbol{m} = \alpha(t)\,\boldsymbol{x}_0\,, \qquad \sigma(t)^2 = \alpha(t)^2\int_0^t \frac{g(\xi)^2}{\alpha(\xi)^2}\mathrm{d}\xi\,, \tag{12}$$

where

$$\alpha(t) = \mathrm{e}^{\int_0^t f(\xi)\mathrm{d}\xi}\,. \tag{13}$$

These expressions will appear in the loss function in what follows below.

From now, we shall continue the explanation by taking a concrete example. From the analogy to the stochastic quantization as we discuss later, we set up the simplest one-dimensional problem such that the probability distribution is given by a double-well potential form, i.e.,

$$p_0(x) = Z^{-1}(a,b)\,\mathrm{e}^{-S(x;a,b)}\,, \qquad S(x;a,b) = ax^2 + bx^4\,. \tag{14}$$

We assume $a, b \in \mathbb{R}$ for the moment, and when we analyze the sign problem later, we will generalize them to complex numbers. The normalization is $Z(a,b) = \int \mathrm{d}x\, \mathrm{e}^{-S(x;a,b)}$ which converges for $b > 0$ or $a > 0$ if $b = 0$. This integral is expressed in terms of special functions, if $a > 0$ and $b > 0$,

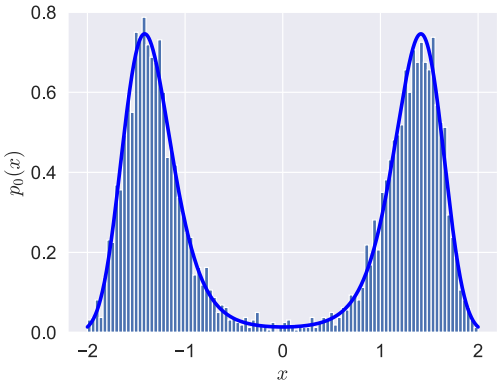

Fig. 1. Randomly sampled 4000 points according to the given $p_0(x)$ with $a = -4$ and $b = 1$ that is shown by the solid line. The histogram is plotted from $x = -2$ to $+2$ with the bin size equally divided by 100 intervals.

as

$$Z(a,b) = \frac{1}{2}\sqrt{\frac{a}{b}}\mathrm{e}^{z}K_{1/4}(z) \tag{15}$$

with $z = a^2/(8b)$. Interestingly, the above expression holds for complex a and b as long as $\mathrm{Re}a > 0$ and $\mathrm{Re}b > 0$.

In this study we rather consider a bit more nontrivial case with $a < 0$. To make our analyses concrete, we specifically choose $a = -4$ and $b = 1$ to visualize the symmetry-breaking-type potential.

In this toy model the true distribution $p_0(x)$ is already known in Eq. (14), and let us explain the step-by-step procedures in the score-based diffusion model.

First, we sample points from $p_0(x)$, which corresponds to sampling image data for the training. Although we already have $p_0(x)$ in the present exercise, we usually do not know what $p_0(x)$ looks like. We can collect the data for the training purpose, assuming the existence of some probability distribution that rules the data. At first glance, it may sound like an unmanageable task, but amazingly, the model can be trained efficiently with the given training data. This is a vital feature for the practical usage. In Fig. 1, we show the example of random sampling according to $x \sim p_0$.

These sampled points, $\{x_i\}$, give the initial values; $x_{i,t=0} = x_i$, and the SDE leads to $x_{i,t}$ for later time. We shall integrate the SDE up to $t = T$. For sufficiently large T, it is expected that $x_{i,T} \sim \mathcal{N}(0,1)$; remember that $f(\boldsymbol{x},t)$ and $g(t)$ were chosen in such a way. The non-trivial question is how $p_t(x)$ should behave in the intermediate time region. In principle, the Fokker-Planck equation (8) uniquely solves $p_t(x)$ for a given initial condition. However, it is generally difficult to solve $p_t(x)$ as it is, and we can translate the problem into the optimization problem of the score function,

$$\boldsymbol{s}_t(\boldsymbol{x};\boldsymbol{\theta}) \approx \boldsymbol{\nabla}\ln p_t(\boldsymbol{x})\,, \tag{16}$$

where $\boldsymbol{\theta}$ denotes fitting parameters to approximate $p_t(\boldsymbol{x})$. Thus, in the DDPM, the loss function is the deviation of $\boldsymbol{s}_t$ from $\boldsymbol{\nabla}\ln p_t$. The L^2-norm yields the loss function as follows:

$$L(\boldsymbol{\theta}) = \mathbb{E}_{p_t}\Big[\|\boldsymbol{s}_t(\boldsymbol{x};\boldsymbol{\theta}) - \boldsymbol{\nabla}\ln p_t(\boldsymbol{x})\|^2\Big]. \tag{17}$$

In the case of the Implicit Score Matching (ISM), we can evaluate this loss function only with $\{\boldsymbol{x}_i\}$ through the following rewriting:

$$\begin{aligned} L_{\mathrm{ISM}}(\boldsymbol{\theta}) &= \int \mathrm{d}\boldsymbol{x}_t\, p_t(\boldsymbol{x})\Big(\boldsymbol{s}_t^2 - 2\boldsymbol{s}_t\cdot\frac{\boldsymbol{\nabla}p_t}{p_t}\Big) + (\text{const.}) \\ &= 2\mathbb{E}_{p_t}\Big[\|\tfrac{1}{2}\boldsymbol{s}_t^2 + \boldsymbol{\nabla}\cdot\boldsymbol{s}_t\|^2\Big] + (\text{const.}) \end{aligned} \tag{18}$$

Here, (const.) represents the terms involving not $\boldsymbol{s}_t$ but p_t only which are independent of $\boldsymbol{\theta}$. From the first to the second line, the integration by part is performed to move $\boldsymbol{\nabla}$ onto $\boldsymbol{s}_t$. In this final form, remarkably, there is no explicit $p_t(\boldsymbol{x})$ in the function itself but it appears in the weight. Thus, this expectation value is well approximated by the training data $\{\boldsymbol{x}_i\}$ under the assumption that the training data obey $p_t(\boldsymbol{x})$. On the other hand, in the method referred to as the Denoising Score Matching (DSM), the loss function is given by

$$\begin{aligned} L_{\mathrm{DSM}}(\boldsymbol{\theta}) &= \frac{1}{N}\sum_{i=1}^{N}\|\boldsymbol{s}_t(\boldsymbol{x}_{i,t};\boldsymbol{\theta}) - \boldsymbol{\nabla}\ln p_{t|0}(\boldsymbol{x}_{i,t}|\boldsymbol{x}_{i,0})\| \\ &= \frac{1}{N}\sum_{i=1}^{N}\Big\|\boldsymbol{s}_t(\boldsymbol{x}_{i,t};\boldsymbol{\theta}) + \frac{\boldsymbol{x}_{i,t} - \alpha(t)\boldsymbol{x}_{i,0}}{\sigma(t)^2}\Big\|, \end{aligned} \tag{19}$$

where the explicit solution of $p_{t|0}(\boldsymbol{x}_{i,t}|\boldsymbol{x}_{i,0})$ is used with $\boldsymbol{m}$ and $\sigma(t)^2$ in Eq. (12). This is again trainable with $\{x_{i,t}\}$. This form (19) is more advantageous than Eq. (18) because of the lack of the gradient.

The python code for two-dimensional numerical simulations in the preceding work[6] is provided in public, and we adapted the code for our one-dimensional simulations. We employ a neural network to represent $s_t(x)$ as defined in the original code. Specifically, in the previous literature,[6] the choice of the neural network is $[x,t] \to$ (`Dense`(128) $\to$ `Swish`)^3 $\to$ `Dense`(1) $\to s_t(x)$. Here, `Swish` is a function also called SiLU given by $x/(1+\mathrm{e}^{-x})$ that looks like ReLU but has no vanishing gradient. Since our current problem is one-dimensional, one might think that far simpler neural networks could work fine, but we numerically found that the performance was not satisfactory if we reduced the layer size. We then train the model from $t = t_0 = 0.01$ to $t = T = 1$ with $\Delta t = (T - t_0)/N_t$ where $N_t = 10^3$. We choose the loss function in Eq. (19) and take the batch size as $N = 32$. For the training, the epoch number is 300. In our present problem, $p_0(x)$ is set by hand, and the corresponding one-dimensional score is $\partial_x \ln p_0(x) = -2ax - 4bx^3$. Figure 2 shows the trained score functions at $t = t_0$, 0.1, 0.2, respectively, together with the exact answer. We can confirm that $s_0(x)$ at $t = t_0$ well approximates $\partial_x \ln p_0(x)$. From the imposed condition, $\lim_{t\to\infty}\boldsymbol{f}(\boldsymbol{x},t)/g(t)^2 = -\boldsymbol{x}/(2\sigma^2)$, we see that $\boldsymbol{s}_{t\to\infty}(\boldsymbol{x}) = -\boldsymbol{x}$ asymptotically. Actually, in Fig. 2, the score function $s_t(x)$ at $t = 0.2$ is already close to this asymptotic behavior of $s_{t\to\infty}(x) = -x$.

Once $s_t(x)$ is trained well, the denoising or the reverse process is described by the following SDE:

$$\dot{\boldsymbol{x}}_t = \boldsymbol{f}(\boldsymbol{x}_t,t) - g(t)^2\boldsymbol{s}_t(\boldsymbol{x}_t;\boldsymbol{\theta}) + g(t)\bar{\boldsymbol{\xi}}_t\,, \tag{20}$$

where t decreases from $t = T$ to $t = t_0$. Figure 3 shows randomly sampled 30 trajectories with initial $x_{i,T} \sim \mathcal{N}(0,1)$.

The physical meaning is transparent; $\ln p_t$ is regarded as a negative potential energy and $\partial_x \ln p_t$ is thus a force. Therefore, in the x-region with $s_t(x) > 0$, the particle at $x_{i,t}$ tends

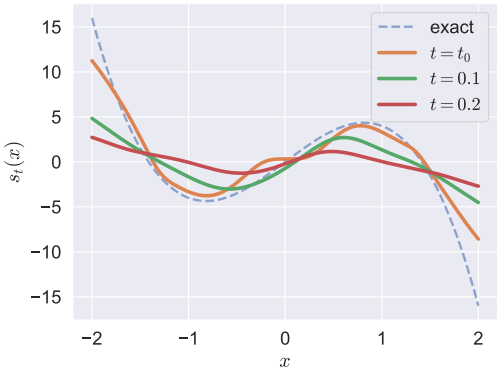

Fig. 2. Trained score functions at $t = t_0$, 0.1, 0.2, respectively, with the exact answer, $\partial_x \ln p_0(x) = -2ax - 4bx^3$, overlaid by the dashed line.

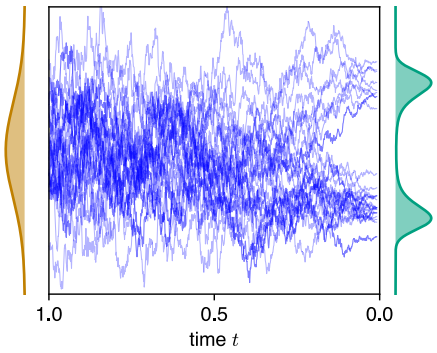

Fig. 3. Randomly sampled 30 trajectories for the reverse process to recover the original probability distribution (depicted by the green shaded region) from the normal distribution (depicted by the brown shaded region).

to move to the positive-x direction, while in the x-region with $s_t(x) < 0$, the tendency is opposite. As a consequence, the distribution becomes denser near x_0 with $s_t(x_0) \approx 0$ when $\partial_x s_t(x_0) < 0$, and the distribution is diluted around x_0 with $s_t(x_0) \approx 0$ when $\partial_x s_t(x_0) > 0$. In this way, one can easily associate $s_t(x)$ in Fig. 2 with the two-peak structure of $p_0(x)$ in Fig. 1.

Now, we are ready to alter the SDE with a free parameter. The Fokker-Planck equation corresponding to the SDE (20) is immediately deduced from the above-mentioned derivation of the Fokker-Planck equation as

$$\dot{p}_t(\boldsymbol{x}) = -\boldsymbol{\nabla}\cdot\left[\Big(\boldsymbol{f}(\boldsymbol{x},t) - g(t)^2 \boldsymbol{s}_t(\boldsymbol{x})\Big)p_t(\boldsymbol{x}) + \frac{g(t)^2}{2}\boldsymbol{\nabla} p_t(\boldsymbol{x})\right]. \quad (21)$$

Here, we dropped $\boldsymbol{\theta}$ to simplify the notation. We can split the last term as

$$\frac{g(t)^2}{2}\boldsymbol{\nabla} p_t(\boldsymbol{x}) = \frac{\lambda^2 g(t)^2}{2}\boldsymbol{\nabla} p_t(\boldsymbol{x}) + \frac{1-\lambda^2}{2} g(t)^2 \boldsymbol{\nabla} p_t(\boldsymbol{x}). \quad (22)$$

The latter term is further rewritten as

$$\frac{1-\lambda^2}{2} g(t)^2 \boldsymbol{\nabla} p_t(\boldsymbol{x}) = \frac{1-\lambda^2}{2} g(t)^2 (\boldsymbol{\nabla} \ln p_t(\boldsymbol{x}))\, p_t(\boldsymbol{x}). \quad (23)$$

Because this final form is proportional to $p_t(\boldsymbol{x})$, we can regard

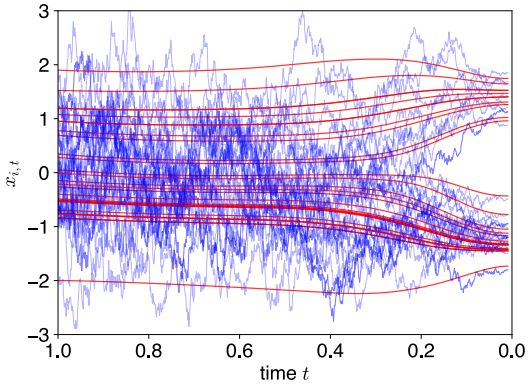

Fig. 4. Deterministic trajectories of the solutions of the ODE (26) with the same initial values as chosen in Fig. 3.

it as a part of the term involving $\boldsymbol{f}(\boldsymbol{x},t) - g(t)^2\, \boldsymbol{s}_t(\boldsymbol{x})$. Then, we can read the corresponding SDE back from the rewritten Fokker-Planck equation, yielding

$$\dot{\boldsymbol{x}}_t = \boldsymbol{f}(\boldsymbol{x}_t,t) - g(t)^2\left(\boldsymbol{s}_t(\boldsymbol{x}_t) - \frac{1-\lambda^2}{2}\boldsymbol{\nabla}\ln p_t(\boldsymbol{x}_t)\right) + \lambda\, g(t)\,\bar{\boldsymbol{\xi}}_t. \quad (24)$$

If the learning is ideally perfect to realize $\boldsymbol{s}_t(\boldsymbol{x}) = \boldsymbol{\nabla}\ln p_t(\boldsymbol{x})$, we can replace $\boldsymbol{\nabla}\ln p_t(\boldsymbol{x}_t)$ in the above SDE with $\boldsymbol{s}_t(\boldsymbol{x}_t)$. After all, with this replacement, we arrive at the following class of SDEs with a free parameter λ:

$$\dot{\boldsymbol{x}}_t = \boldsymbol{f}(\boldsymbol{x}_t,t) - \frac{1+\lambda^2}{2} g(t)^2 \boldsymbol{s}_t(\boldsymbol{x}_t;\boldsymbol{\theta}) + \lambda\, g(t)\,\bar{\boldsymbol{\xi}}_t. \quad (25)$$

Interestingly, if we choose $\lambda = 0$, then the stochastic noise is completely removed from the differential equation; that is, we can perform the sampling procedure using the deterministic equation:[7)]

$$\dot{\boldsymbol{x}}_t = \boldsymbol{f}(\boldsymbol{x}_t,t) - \frac{1}{2} g(t)^2 \boldsymbol{s}_t(\boldsymbol{x}_t;\boldsymbol{\theta}). \quad (26)$$

Without noises, the solutions of the ODE behave smoothly, as shown by the red lines in Fig. 4. We note that the red lines start with the initial values, $\{x_{i,T}\}$, chosen to be the same as in Fig. 3.

Now, it is important to note that the time-dependence in $\boldsymbol{s}_t(\boldsymbol{x}_t)$ should be properly adjusted within the finite time interval $(0, T]$. For example, if we continue solving Eq. (26) until equilibration is achieved, then $\dot{\boldsymbol{x}}_t = 0$ at $t \to -\infty$ leads to the condition, $\boldsymbol{s}_t(\boldsymbol{x}_t) = 2\boldsymbol{f}(\boldsymbol{x}_t)/g(t)^2|_{t\to-\infty} = \boldsymbol{x}_t/\sigma^2$. Hence, all of $\{x_{i,t}\}$ eventually converges to discrete points where the condition is satisfied. Also, if we skip the training procedure and simply plug $\boldsymbol{s}_0(\boldsymbol{x})$ in Eq. (26), the solution of Eq. (26) does not follow $p_0(\boldsymbol{x})$. At the same time, these features imply that the convergence of the ODE solution to follow $p_0(\boldsymbol{x})$ could be even more improved with refinement of $\boldsymbol{f}(\boldsymbol{x}_t,t)$ and $g(t)$, though we do not discuss this possibility in this review.

3. A Review of Stochastic Quantization

In physics, among many, a direct analogue of the framework of the diffusion model is found in the theory of stochastic quantization. Let us suppose that we have a 0-dimensional field-theoretical model, that is, a partition function given by a one-dimensional integral. To make a connection to the pre-

vious section about the DDPMs based on Eq. (14), we shall take the following model:

$$Z(a,b) = \int \mathrm{d}\phi\, \mathrm{e}^{-S(\phi;a,b)}\,, \qquad S(\phi;a,b) = a\phi^2 + b\phi^4\,. \quad (27)$$

Then, as a matter of fact, we are going to solve the same problem as in the previous section using a different language and convention. Here, ϕ has no coordinate dependence for simplicity, but the generalization to a more realistic physical model is straightforward. This model is useful as a prototype of the quantum field-theoretical problem. If ϕ has only t dependence, then the $(0+1)$ dimensional model can be regarded as the quantum mechanical system.

For Rea $<$ 0 which is physically more interesting than Rea $>$ 0, the analytical expression of the partition function is slightly changed from Eq. (15). We can compute the nonzero expectation values of physical observables as

$$\langle\phi^{2n}\rangle_{\mathrm{c}} = (-1)^n \frac{\partial^n}{\partial a^n} \ln Z(a,b)\,. \quad (28)$$

In particular, for $(a,b) = (-4,1)$, we can find the analytical expression for $\langle\phi^2\rangle_{\mathrm{c}}$:

$$\langle\phi^2\rangle_{\mathrm{c}} = \frac{9}{8} + \frac{I_{3/4}(z) + I_{5/4}(z) + \frac{1}{8}I_{1/4}(z) - \frac{1}{8}I_{-1/4}(z)}{I_{1/4}(z) + I_{-1/4}(z)}\,. \quad (29)$$

We note that this setup with Rea $<$ 0 corresponds to a physical system with spontaneous symmetry breaking. Precisely speaking, there is no spontaneous symmetry breaking unless the degrees of freedom are infinitely large, and yet, we can see a bifurcation in numerical simulations.

Since the problem is as elementary as one-dimensional integral, the numerical integration is not difficult at all. This situation will be totally changed once the sign problem occurs as we will address later. For our present analysis with $(a,b) = (-4,1)$, we can easily figure out the first two expectation values, for example, as

$$\langle\phi^2\rangle_{\mathrm{c}} \approx 1.83534\,, \qquad \langle\phi^4\rangle_{\mathrm{c}} \approx 0.552211\,. \quad (30)$$

We could continue the calculations to higher powers if necessary, but for the purpose of benchmark test, these first two expectation values should suffice.

Now, let us explain how the stochastic quantization works. In the stochastic quantization, instead of performing the functional integral, the key ingredient is the SDE along the fictitious time (or the quantum axis), which is denoted by τ here. The stochastic field variable is defined by the solution of the SDE as

$$\dot\phi_\eta(x,\tau) = -\frac{\delta S[\phi_\eta]}{\delta\phi_\eta(x,\tau)} + \eta(x,\tau)\,. \quad (31)$$

Generally, the field variables are functions of spacetime, x, including both spatial and temporal coordinates. Although our toy model has no x dependence, in the review part, we shall keep the general notation with d-dimensional spacetime. Here, $\eta(x,\tau)$ is the stochastic noise which satisfies,

$$\begin{aligned}&\langle\eta(x,\tau)\rangle_\eta = 0\,,\\ &\langle\eta(x_1,\tau_1)\eta(x_2,\tau_2)\rangle_\eta = 2\delta^{(d)}(x_1-x_2)\delta(\tau_1-\tau_2)\,,\end{aligned} \quad (32)$$

where $\langle\cdots\rangle_\eta$ denotes the average over the noise $\eta(x,\tau)$. In the path-integral formalism, this can be represented as

$$\langle A[\eta]\rangle_\eta := \frac{\displaystyle\int \mathcal{D}\eta\, A[\eta]\, \exp\Big[-\frac{1}{4}\int \mathrm{d}^d x\, \mathrm{d}\tau\, \eta(x,\tau)^2\Big]}{\displaystyle\int \mathcal{D}\eta\, \exp\Big[-\frac{1}{4}\int \mathrm{d}^d x\, \mathrm{d}\tau\, \eta(x,\tau)^2\Big]}\,. \quad (33)$$

Then, the quantum expectation value is obtained from the noise expectation value at infinitely large τ:

$$\langle\phi(x_1)\cdots\phi(x_k)\rangle = \lim_{\tau\to\infty}\langle\phi_\eta(x_1,\tau)\cdots\phi_\eta(x_k,\tau)\rangle_\eta\,. \quad (34)$$

This is the calculation scheme in the stochastic quantization.

Alternatively, we can introduce the probability distribution from

$$\langle\phi_\eta(x_1,\tau)\cdots\phi_\eta(x_k,\tau)\rangle_\eta =: \int \mathcal{D}\phi\, p_\tau(\phi)\,\phi(x_1)\cdots\phi(x_k)\,. \quad (35)$$

We are now sufficiently experienced to write down the corresponding Fokker-Planck equation immediately. That is, the identification of $\boldsymbol{f} = -\boldsymbol{\nabla}_\phi S[\phi]$ and $g(t) = \sqrt{2}$ in Eq. (8) leads to

$$\dot p_\tau = \int \mathrm{d}^d x\, \frac{\delta}{\delta\phi(x,\tau)}\Big(\frac{\delta S}{\delta\phi(x,\tau)} + \frac{\delta}{\delta\phi(x,\tau)}\Big) p_\tau[\phi]\,. \quad (36)$$

We can introduce a trick here to deform this evolution equation with a free parameter. Then, the SDE is also modified as

$$\dot\phi_\eta(x,\tau) = -\frac{\delta S[\phi_\eta]}{\delta\phi_\eta(x,\tau)} - (1-\lambda^2)\frac{\delta}{\delta\phi_\eta(x,\tau)}\ln p_\tau(\phi_\eta) + \lambda\eta(x,\tau)\,. \quad (37)$$

Here again, we can take the limit of $\lambda = 0$, so that the τ-evolution should become deterministic. That is,

$$\dot\phi(x,\tau) = -\frac{\delta S[\phi]}{\delta\phi(x,\tau)} - \frac{\delta}{\delta\phi(x,\tau)}\ln p_\tau[\phi]\,. \quad (38)$$

Obviously, the τ-evolution of $\phi(x,\tau)$ ceases when $p_\tau[\phi] \approx \mathrm{e}^{-S[\phi]}$. So, once $p_\tau[\phi]$ or the derivative of $\ln p_\tau[\phi]$ (which is a counterpart of the score function) is trained, the stochastic nature can be completely eliminated from the formalism. Here again, we emphasize that the quantum nature is captured by the entire τ-evolution of $p_\tau[\phi]$. In this review, we will not pursue this issue, and the interested readers can consult the recent work; see especially Sec. 3.3 in the literature[9] for the effective action from the probability flow ODE formulation. Although Eq. (38) is not well known in physics, it deserves extensive investigations.

For the practical application, we need the initial condition. In the literature, the null initial condition is the conventional choice, i.e.,

$$p_{\tau=0}[\phi] = \prod_x \delta[\phi(x)]\,, \quad (39)$$

but this form is not mandatory. Actually, the above choice of the singular form is numerically inconvenient and we can better start with a more regular form such as the normal distribution. Then, Fig. 5 shows the randomly sampled trajectories when we choose $\Delta\tau = T/N_{\mathrm{step}}$ with $N_{\mathrm{step}} = 2^{10}$. We immediately realize that the convergence from the normal distribution to the original distribution, $p_0(x)$, is qualitatively similar to the behavior we have seen in Fig. 3. The quantitative differences are caused by the absence of $\boldsymbol{f}(\boldsymbol{x},t)$ and $g(t)$ in the standard formulation of the stochastic quantization.

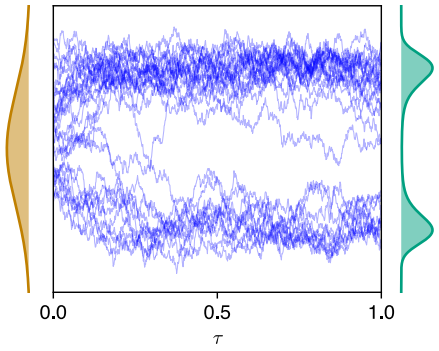

Fig. 5. Randomly sampled 30 trajectories in the stochastic quantization with the initial condition given by the normal distribution.

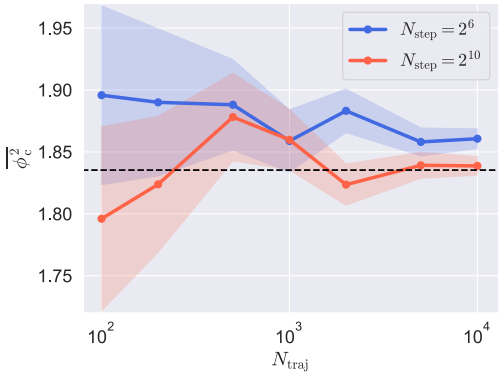

Fig. 6. Ensemble average of ϕ^2 over randomly sampled $N_{\rm traj}$ trajectories with different time steps. The 1σ band is estimated from $N_{\rm traj}$ trajectories. The dashed black line represents the exact answer.

It should be noted that a technique similar to introducing $g(t)$ is known in the context of the stochastic quantization. The noise fluctuation in Eq. (33) could be regularized by a kernel as follows,

$$\int \mathcal{D}\eta\, A[\eta]\, \exp\Big[-\frac{1}{4}\int \mathrm{d}^d x\, \eta(x,\tau)^2\Big] \to \int \mathcal{D}\eta_K\, A[\eta_K]\, \exp\Big[-\frac{1}{4}\int \mathrm{d}^d x\, \mathrm{d}^d y\, \eta_K(x,\tau)K(x,y)\eta_K(y,\tau)\Big]. \tag{40}$$

Equivalently, $\eta(x,\tau)$ itself is kept as the Gaussian noise and the (square-root of) kernel $K(x,y)$ could be multiplied to the SDE in Eq. (31). If we use this modified noise, the SDE should be altered as

$$\dot{\phi}_{\eta_K}(x,\tau) = -\int \mathrm{d}^d y\, K(x,y)\, \frac{\delta S[\phi_{\eta_K}]}{\delta \phi_{\eta_K}(x,\tau)} + \eta_K(x,\tau)\,. \tag{41}$$

This modification is sometimes required; the naïve application of the method to fermions breaks down and it is convenient to *bosonize* the SDE with an appropriate kernel.

Let us turn back to the discussions about Fig. 5. From the final distribution of the stochastic trajectories, we can compute the average values of ϕ^{2n}. Here, we specifically evaluate:

$$\overline{\phi^2}_{\rm c} := \frac{1}{N_{\rm traj}}\sum_{i=1}^{N_{\rm traj}} \phi_{i,T}^2\,, \tag{42}$$

$$\overline{\phi^4}_{\rm c} := \frac{1}{N_{\rm traj}}\sum_{i=1}^{N_{\rm traj}} \phi_{i,T}^4 - \big(\overline{\phi^2}_{\rm c}\big)^2\,, \tag{43}$$

using randomly sampled $N_{\rm traj}$ trajectories. We then compare $\overline{\phi^2}_{\rm c}$ and $\overline{\phi^4}_{\rm c}$ with the exact values in Eq. (30). We expect that these average values should converge to the exact answer as $N_{\rm traj}$ increases. We quantify this tendency by making plots of $\overline{\phi^{2n}}_{\rm c}$ as functions of $N_{\rm traj}$ as shown in Figs. 6 and 7.

From these results in Figs. 6 and 7, as expected, we can conclude that the average values over the trajectories certainly approach the exact answers as $N_{\rm traj}$ increases for sufficiently large $N_{\rm step}$. Here, we fixed T and decreased $\Delta\tau$ for larger $N_{\rm step}$. We can alternatively increase T with fixed $\Delta\tau$.

It is rather perplexing that the convergence appears not so fast; even in this simplest toy model, the average values with $N_{\rm step} = 2^{10}$ and $N_{\rm traj} = 10^4$ are not satisfactorily close to the

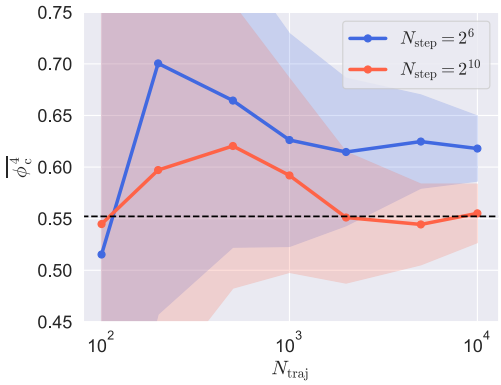

Fig. 7. Ensemble average of the connected part of ϕ^4 over randomly sampled $N_{\rm traj}$ trajectories with different time steps. The 1σ band is estimated from $N_{\rm traj}$ trajectories. The dashed black line represents the exact answer.

exact answers. To quantify this, we estimated the error bar from $\sigma = \sqrt{\overline{\phi^4}_{\rm c}/(N_{\rm traj}-1)}$ as displayed by the band in Fig. 6. For $N_{\rm step} = 2^{10}$ and $N_{\rm traj} = 10^4$, we find that $\overline{\phi^2}_{\rm c} = 1.8387 \pm 0.0075$, while the exact answer is ≈ 1.83534, which actually shows good agreement. In the same way, we estimated the error bar for $\overline{\phi^4}_{\rm c}$ to conclude $\overline{\phi^4}_{\rm c} = 0.5552 \pm 0.0283$ for the exact answer ≈ 0.55221. In this case, the numerical result happens to be close to the exact answer, but the error bar is still large and the numerical agreement seems to be accidental.

To improve the convergence problem, the common strategy is to replace the *ensemble average* over trajectories with the *time average* of stochastic evolution. In principle, if the time extent is large enough, even a single trajectory should reproduce the correct answer in this way. Let us check how this strategy works. We make a plot similar to Fig. 6 to demonstrate the convergence properties of the time-averaged value. We calculated not only $\overline{\phi^2}_{\rm c}$ but also $\overline{\phi^4}_{\rm c}$, but it is sufficient to show the comparison of $\overline{\phi^2}_{\rm c}$ for the present demonstration. Figure 8 presents the results for $N_{\rm traj} = 1$ and $N_{\rm traj} = 10$ with the error bar estimated from fluctuations with 100 indepen-

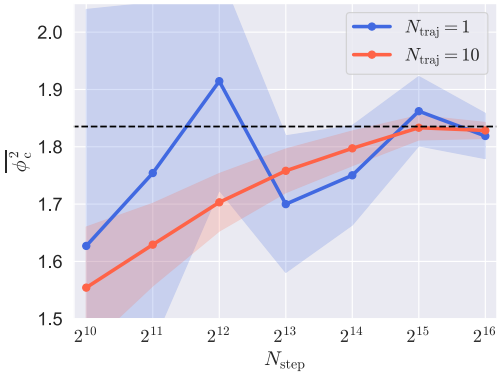

Fig. 8. Time average of the connected part of ϕ^2 over $N_{\rm step}$ time steps and $N_{\rm traj}$ trajectories. The 1σ band is estimated from 100 runs. The dashed black line represents the exact answer.

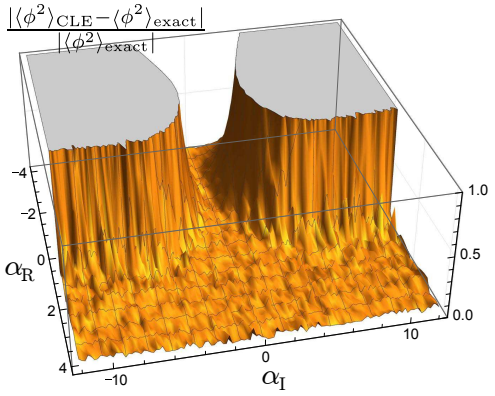

Fig. 9. Comparison to the exact answer in complex parameter space. In the present notation, the quadratic coefficients are related as $a_{\rm R/I} = a_{\rm R/I}/2$ and the quartic coefficient is fixed as $\beta = b/4 = 1$. Figure is adapted from the paper.[24)]

dent runs. From these figures, we see that the time average of even the single trajectory ($N_{\rm traj} = 1$) can approach the exact answer if $N_{\rm step}$ is large enough. Of course, the agreement with the exact answer is more guaranteed by further taking the ensemble average over $N_{\rm traj} > 1$ trajectories.

4. Sign Problem

The reinterpretation of the stochastic quantization and the diffusion model has been considered[9)] which aims to improve the efficiency to sample quantum lattice field configurations. The applications of two seemingly different but essentially equivalent formulations have just started recently, and further developments should await to be revealed.

Among various possibilities, one direction more exciting than improving the efficiency is an attempt to evade the sign problem. As long as $e^{-S[\phi]}$ is positive definite, we can give it a meaning as the probability, $p_\tau[\phi] \sim e^{-S[\phi]}$. However, in many interesting physical systems, $e^{-S[\phi]}$ is not necessarily positive definite. For example, the real-time evolution requires the Minkowskian formulation in which the functional integral involves $e^{iS[\phi]}$. Because of this complex nature, $e^{iS[\phi]}$ is an oscillatory function and the Monte-Carlo integration algorithm breaks down. Another example is found in fermionic systems at finite chemical potential; see reviews.[14–17)]

Although the probability interpretation loses its meaning, the stochastic quantization scheme still looks feasible. The only extension is that the trajectories may spread over the complex plane; in other words, ϕ_η may become a complex-valued function. This generalized stochastic quantization method is referred to as the Complex Langevin Equation (CLE) approach. Actually, the CLE application to the Strong Interaction has a long history traced back to 1985 by a pioneering work.[18)] Since then, the interest in nuclear physics was revived around 2010; see, e.g., a discussion on the prospect of gauge-cooling technique.[19)] Continued and latest applications include the quark-flavor number dependence of finite-density matter[20)] and full real-time simulations of strongly interacting system.[21)]

The problem in the CLE method is that the results may not converge or even the converged results may not be correct. The convergence criterion has been known; see some discussions.[22,23)]

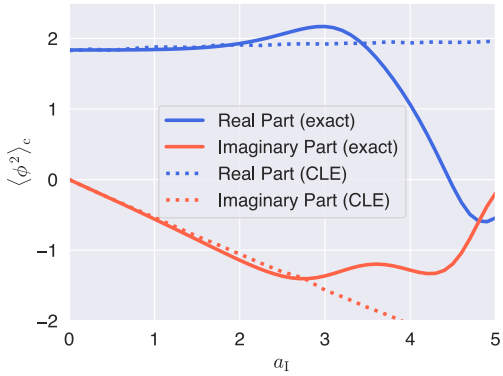

Fig. 10. CLE simulations compared to the exact numerical values as a function of $a_{\rm I}$ with $a_{\rm R} = -4$ and $b = 1$ fixed.

We can introduce the sign problem into our simple toy model. It is an intriguing question what would happen if we generalize the model parameters on the complex plane. Here, let us take $a = a_{\rm R} + ia_{\rm I}$ with $a_{\rm R} = -4$ fixed and we change $a_{\rm I}$ to control the degree of complexification. Actually, this model has been carefully studied in the paper[24)] with a slightly different notation; $\alpha = a/2$ and $\beta = b/4$ in the potential coefficients and $\beta = 1$ was chosen there.[24)] Then, the breakdown of the CLE method has been quantified as shown in Fig. 9 as a function of the real and imaginary parts of α.

As long as $\alpha_{\rm R} > 0$, the convergence property is good and regardless of complexification with $\alpha_{\rm I} \neq 0$, the CLE can converge to the correct answer of $\langle\phi^2\rangle_c$. In the region with $\alpha_{\rm R} < 0$, the sign problem occurs and the CLE breaks down except for the region with $|\alpha_{\rm I}| \ll |\alpha_{\rm R}|$ where the method is reduced to the real stochastic quantization. In this way, as intuitively expected, we understand that the sign problem turns out to be severe in the region with $\alpha_{\rm R} < 0$ and $|\alpha_{\rm I}| > |\alpha_{\rm R}|$. Strictly speaking, this conclusion is valid for $\langle\phi^2\rangle_c$ and higher-order operator expectations might be more sensitive to the sign problem.

Now, let us carry out the direct CLE simulation using the

present setup of the toy model. We fix $a_{\rm R} = -4$ and $b = 1$ (which corresponds to $\alpha_{\rm R} = -8$ and $\beta = 4$ in the previous convention[24]), and calculate the real and imaginary parts of $\langle\phi^2\rangle_{\rm c}$. The results are summarized in Fig. 10. For this simulation, we chose $N_{\rm traj} = 10$ and $N_{\rm step} = 2^{16}$ and took the time-average. Also, we used a discretization scheme with some resummation to avoid run-away behavior of trajectories. For technical details, see relevant discussions in the anharmonic oscillator simulation.[25]

In view of Fig. 10, the sign problem certainly gets worse for $a_{\rm I} \gtrsim |a_{\rm R}|$. Of course, the best would be providing the solution of the sign problem, but the sign problem is known to be NP-hard,[26] and simple solutions within reasonable time scale would be unattainable. Then, the second best would be the identification of the difficulty in the formalism. In this context, it is shown that the analysis of the Lefschetz thimble can make it clear where the CLE may fail.[27]

The idea of complexifying the path integral using the Lefschetz thimble was demonstrated in the seminal work,[28] which was successfully implemented for numerical simulations in quantum field theories.[29] Since this is nothing but the higher-dimensional extension of the complex analysis for the one-dimensional integral, we do not have to consider thimble structures for the present setup of the one-dimensional problem. Let us introduce a complex variable; $z = {\rm Re}\phi + i{\rm Im}\phi$. We can deal with the complexified action, $S[\phi] \to S[z]$, then. We should find the critical points, z_i, by solving the saddle-point condition,

$$S'[z]\Big|_{z=z_\sigma} = 0\,. \tag{44}$$

In our present case with $S[z] = az^2 + bz^4$, we should get three critical points at $z = 0$, $z = \pm\sqrt{-a/(2b)}$. The steepest descent cycles are defined with a time-like variable, τ, as

$$\mathcal{I}_\sigma := \left\{ z(\tau) \,\middle|\, \frac{{\rm d}z}{{\rm d}\tau} = \overline{\frac{\partial S}{\partial z}}\,,\ z(\tau \to -\infty) = z_\sigma \right\}. \tag{45}$$

The nicest feature about the steepest descent cycles is that the phase oscillation is suppressed on it, which is almost obvious from $({\rm d}/{\rm d}\tau){\rm Im}S \propto ({\rm d}/{\rm d}\tau)(S - \overline{S}) = S'({\rm d}z/{\rm d}\tau) - \overline{S}'({\rm d}\overline{z}/{\rm d}\tau) = S'\overline{S'} - \overline{S}'S' = 0$. The original integral can be safely deformed to the sum of integrals on the steepest descent cycles or the Lefschetz thimbles. Once this rewriting is complete, it is only the residual sign problem which causes obstacle in the numerical simulation.[30]

Figures 11 and 12 show the structures of the steepest descents by the purple lines and the steepest ascents by the green lines as well as the slope, $\overline{S'}$, indicated by small arrows (the length of the arrows is square-root proportional to the modulus of the slope). When the sign problem is minor with $a_{\rm I} = 1$, three critical points are almost along the real axis as seen in Fig. 11. The original integration path should be deformed into three pieces attached to three critical points. For the calculation of $\langle\phi^2\rangle_{\rm c}$, the contribution from $z = 0$ is suppressed, and the saddle-point approximation at $z = \pm\sqrt{-a/(2b)}$ leads to

$$\langle\phi^2\rangle_{\rm c} \simeq \frac{-(a/b){\rm e}^{a^2/(4b^2)}}{1 + 2{\rm e}^{a^2/(4b^2)}}\,. \tag{46}$$

If $|2{\rm e}^{a^2/(4b^2)}| \gg 1$, then we can further simplify the above estimate as $\langle\phi^2\rangle_{\rm c} \simeq -a/2 = 2 - ia_{\rm I}/2$. This formula nicely explains the small-$a_{\rm I}$ behavior of Fig. 10.

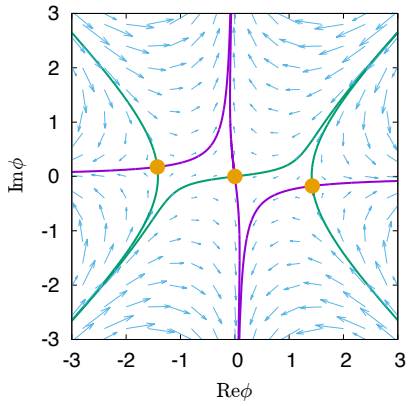

Fig. 11. Structure of the steepest descents (purple) and the steepest ascents (green) with the saddle points (orange dots) for $a_{\rm I} = 1$ with $a_{\rm R} = -4$ and $b = 1$ fixed. The slope, $\overline{S'}$, is indicated by the arrows.

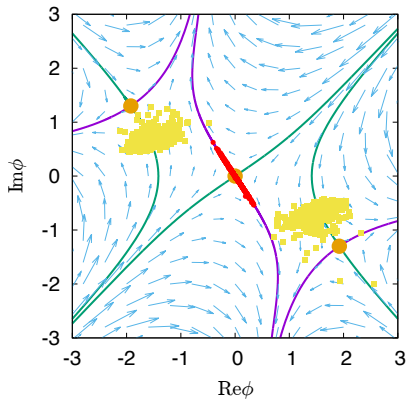

Fig. 12. Same as Fig. 11 for $a_{\rm I} = 10$ with $a_{\rm R} = -4$ and $b = 1$ fixed. Scattering data represent the CLE sampled points without the modified kernel (yellow) and with the modified kernel (red).

The CLE method clearly breaks down for $a_{\rm I} \gtrsim 4$ as quantified in Fig. 10, and the configuration of the steepest descents/ascents for $a_{\rm I} = 10$ is shown in Fig. 12. At first glance, we see no qualitative difference from Fig. 11. The saddle point approximation for $a_{\rm I} = 4$ gives ${\rm Re}\langle\phi^2\rangle_{\rm c} \approx 0.78$ and ${\rm Im}\langle\phi^2\rangle_{\rm c} \approx -2.57$. So, the saddle-point approximation is not good any more. As a matter of fact, the breakdown of the CLE method and the deviation from the leading-order saddle-point approximation are somehow related. The weight of the saddle-point contributions at $z = \pm\sqrt{-a/(2b)}$ is $|{\rm e}^{a^2/(4b^2)}| = {\rm e}^{4-a_{\rm I}^2/4}$, so that the exponent changes its sign at $|a_{\rm I}| = 4$. Thus, around $a_{\rm I} \sim 4$, the saddle-point approximation is no longer effective with such a small exponent, and the relative weight between 1 from $z = 0$ and ${\rm e}^{a^2/(4b^2)}$ from $z = \pm\sqrt{-a/(2b)}$ is flipped.

In view of Fig. 10, the CLE results can be understood if 1 in the denominator in Eq. (46) is dropped, though it should not be dropped in reality. Indeed, this observation can be confirmed by the distribution of the CLE sampled points overlaid by yellow dots on Fig. 12. It is obvious that the CLE method fails to collect the contribution near $z = 0$. A more comprehensive analysis is found in the literature.[31] If we make a

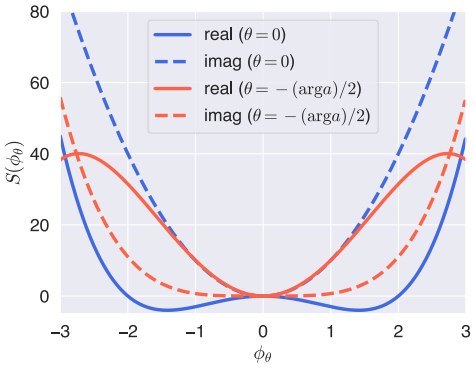

Fig. 13. Action or potential energy for $a_{\rm I} = 10$ along $\mathrm{Im}\phi = 0$, i.e., $\theta = 0$ indicated by the blue solid (real part) and dashed (imaginary part) lines. With the kernel, the red solid (real part) and dashed (imaginary part) lines represent the action along $\arg(a\phi^2) = 0$, i.e., $\theta = -(\arg a)/2$.

similar scattering plot on Fig. 11, it is also the case that the $z = 0$ point repels sampled data. Of course, such repulsion around $z = 0$ is perfectly reasonable for $a_{\rm I} \approx 0$, for $e^{-S[z]}$ has maxima at $z = \pm\sqrt{-a/(2b)}$, while the $z = 0$ point corresponds to the minimum, which is least favored naturally. In terms of the Lefschetz thimble, the repulsion simply means a small relative weight. As $a_{\rm I}$ grows up, the saddle points are aligned along the path with the phase angle by $-(\arg a)/2$. It has been shown that the CLE method with the rotated variable from ϕ to ϕ_θ by $-(\arg a)/2$, that is,

$$\phi =: e^{i\theta}\phi_\theta\,, \qquad \theta = -(\arg a)/2\,, \tag{47}$$

can improve the convergence to the correct answer.[24] Then, we see $a\phi^2 = |a|\phi_\theta^2$. In fact, in the vicinity of $z = 0$, we draw the action $S[z]$ (or it could be called the potential) as a function of this rotated variable ϕ_θ in Fig. 13. Along this direction of ϕ_θ with $\theta = -(\arg a)/2$, the imaginary part of the action is flat and the real part shows a minimum. This makes a sharp contrast to the $\theta = 0$ case that the real part has a double-well shape with a maximum at $\phi = 0$. Accordingly, for the case with $\theta = -(\arg a)/2$, the sampled points are certainly localized around $z \simeq 0$ as seen from the red dots in Fig. 12. The important point is that the SDE (31) is not invariant under the variable change like Eq. (47). In fact, as we already mentioned, the SDE could be modified as Eq. (41) with a kernel and the phase transformation in Eq. (47) corresponds to the choice of the kernel as $K(x, y) = e^{2i\theta}\delta(x - y)$. Surprisingly, the choice of the variable or the kernel would affect the final numerical output. In the present case with $a_{\rm I} = 10$, for example, the correct answer is $\langle\phi^2\rangle \approx -0.01051 - 0.04931\mathrm{i}$. The naïve application of the CLE method gives a totally wrong result, $\langle\phi^2\rangle_{\rm CLE} \approx (1.98058 \pm 0.02333) + (-5.06101 \pm 0.00082)\mathrm{i}$, while the modified CLE method with the kernel results in a much better value, $\langle\phi_\theta^2\rangle_{\rm CLE} \approx (-0.01901 \pm 0.00079) + (-0.04503 \pm 0.00184)\mathrm{i}$.

5. Speculative Prospects

Now, we have seen a quite suggestive analogy between the stochastic quantization and the diffusion model. In the most direct application of the analogy, the diffusion models can be utilized to generate the lattice configurations efficiently. Here, we have put more emphasis on the dark side of the numerical simulation, namely, the sign problem. Actually, for the diffusion models too, the difficulty encountered with complex variables is not necessarily an academic exercise. Nowadays, "quantum" is such a fashionable keyword, and the crucial difference between quantum and classical lies in the information of the complex phase from where the interference effect emerges. Therefore, if one would like to *quantumize* the generative models, the first step would be to complexify the formalism. Although the complexification may not cause troubles in some parameter space, the empirical rule we know from physics realms is that the problem gets harder in the regimes with more interesting contents. We speculate that our knowledge about the CLE method should be useful for future studies along these lines. In particular, we have established good understanding of the breakdown of the method in terms of the Lefschetz thimble and demonstrated a potential resolution by means of the optimized kernel.

A more ambitious direction of research is to solve or tame the sign problem by means of the machine-learning techniques. There are some preceding works to optimize the integration path,[32,33] for example. The analogy to the diffusion models may pave a novel passage to tackle the sign problem. In the CLE, the convergence problem occurs due to run-away trajectories. In the present simulation, we used a half-implicit method to regularize such trajectories, but in general, there are always unstable directions in complex plane. This problem may be cured by the ODE-type evolution without any fluctuations. In gauge theories, the gauge cooling is a technology to evade the run-away trajectories, and if the ODE reformulation turns out to be effective, the algorithm may be improved. An even more radical speculation is the possibility of accessing the analytical structures of the trained score function. The score function provides us with a mapping between the normal distribution and the probability distribution of our interest, which is nothing but the procedure to *solve* the theory in physics. The mapping is also translated into the change from the original variables to the optimal variables in the integral, and if the optimization imposes the condition to suppress the phase oscillation in a way as described below Eq. (45), the flow to de-complexify the theory may naturally find the paths along the Lefschetz thimbles. Alternatively, the optimized kernel could be found in the machine-learning assisted algorithm.

What is speculated here may sound too optimistic, but anything is not mature yet. There are many fascinating attempts in physics, and the Lefschetz thimble is just one of them. We did not mention here, but the stochastic quantization has a useful mathematical structure of supersymmetry (see discussions in the literature[34]). We are making efforts to export the physics wisdom to the diffusion models. We would like to invite interested physicists to this aspiring project and we hope that our present review serves as an interpreter of two descriptions in physics and machine learning.

We would like to thank Yuji Hirono, Jan Pawlowski, Franz Sattler, Akinori Tanaka, Lingxiao Wang, and Kai Zhou for useful conversations. This work is partially supported by the JSPS KAKENHI Grant No. 22H05118.

Syo Kamata is a project assistant professor at Department of Physics, The University of Tokyo. The research subjects include the Lefschetz thimble in low-dimensional quantum field theory, the exact WKB, and the machine-

learning analysis of the astrophysical observations.

Kenji Fukushima is a professor at Department of Physics, The University of Tokyo. The research subjects include the high-energy nuclear physics in extreme environments such as high temperature, high baryon density, strong magnetic field, etc.

1) J. Ho, A. Jain, and P. Abbeel: Advances in neural information processing systems **33** (2020) 6840.
2) J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli: International conference on machine learning, 2015, pp. 2256–2265.
3) Y. Song and S. Ermon: Advances in neural information processing systems **32** (2019).
4) Z. Ramzi, B. Remy, F. Lanusse, J.-L. Starck, and P. Ciuciu: arXiv preprint arXiv:2011.08698 (2020).
5) Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole: arXiv preprint arXiv:2011.13456 (2020).
6) Y. Hirono, A. Tanaka, and K. Fukushima: (2024).
7) J. Song, C. Meng, and S. Ermon: arXiv preprint arXiv:2010.02502 (2020).
8) P. H. Damgaard and H. Huffel: Phys. Rept. **152** (1987) 227.
9) L. Wang, G. Aarts, and K. Zhou: JHEP **05** (2024) 060.
10) E. Nelson: Phys. Rev. **150** (1966) 1079.
11) Y. V. Kovchegov and E. Levin: *Quantum Chromodynamics at High Energy* (Oxford University Press, 2013), Vol. 33.
12) J.-P. Blaizot, E. Iancu, and H. Weigert: Nucl. Phys. A **713** (2003) 441.
13) K. Rummukainen and H. Weigert: Nucl. Phys. A **739** (2004) 183.
14) S. Muroya, A. Nakamura, C. Nonaka, and T. Takaishi: Prog. Theor. Phys. **110** (2003) 615.
15) K. Fukushima and T. Hatsuda: Rept. Prog. Phys. **74** (2011) 014001.
16) G. Aarts: J. Phys. Conf. Ser. **706** (2016) 022004.
17) K. Nagata: Prog. Part. Nucl. Phys. **127** (2022) 103991.
18) F. Karsch and H. W. Wyld: Phys. Rev. Lett. **55** (1985) 2242.
19) G. Aarts, L. Bongiovanni, E. Seiler, D. Sexty, and I.-O. Stamatescu: Eur. Phys. J. A **49** (2013) 89.
20) Y. Namekawa, Y. Asano, Y. Ito, T. Kaneko, H. Matsufuru, J. Nishimura, A. Tsuchiya, S. Tsutsui, and T. Yokota: PoS **LATTICE2021** (2022) 623.
21) K. Boguslavski, P. Hotzy, and D. I. Müller: Phys. Rev. D **109** (2024) 094518.
22) G. Aarts, P. Giudice, and E. Seiler: Annals Phys. **337** (2013) 238.
23) S. Shimasaki, K. Nagata, and J. Nishimura: PoS **LATTICE2016** (2016) 071.
24) Y. Abe and K. Fukushima: Phys. Rev. D **94** (2016) 094506.
25) R. Anzaki, K. Fukushima, Y. Hidaka, and T. Oka: Annals Phys. **353** (2015) 107.
26) M. Troyer and U.-J. Wiese: Phys. Rev. Lett. **94** (2005) 170201.
27) T. Hayata, Y. Hidaka, and Y. Tanizaki: Nucl. Phys. B **911** (2016) 94.
28) E. Witten: (2010).
29) M. Cristoforetti, F. Di Renzo, and L. Scorzato: Phys. Rev. D **86** (2012) 074506.
30) H. Fujii, D. Honda, M. Kato, Y. Kikukawa, S. Komatsu, and T. Sano: JHEP **10** (2013) 147.
31) K. Nagata, J. Nishimura, and S. Shimasaki: Phys. Rev. D **94** (2016) 114515.
32) Y. Mori, K. Kashiwa, and A. Ohnishi: Phys. Rev. D **96** (2017) 111501.
33) Y. Mori, K. Kashiwa, and A. Ohnishi: PTEP **2018** (2018) 023B04.
34) K. Fukushima and Y. Tanizaki: PTEP **2015** (2015) 111A01.