arXiv:2006.00702v1 [cond-mat.stat-mech] 1 Jun 2020

Article

Interacting particle solutions of Fokker–Planck equations through gradient–log–density estimation

Dimitra Maoutsa [1], Sebastian Reich [2] and Manfred Opper [1]

[1] Artificial Intelligence Group, Technische Universität Berlin, Marchstraße 23, Berlin 10587, Germany
[2] Institute of Mathematics, University of Potsdam, Karl-Liebknecht-Str. 24/25, 14476 Potsdam, Germany
* Correspondence: dimitra.maoutsa@tu-berlin.de; manfred.opper@tu-berlin.de

Received: date; Accepted: date; Published: date

Abstract: Fokker–Planck equations are extensively employed in various scientific fields as they characterise the behaviour of stochastic systems at the level of probability density functions. Although broadly used, they allow for analytical treatment only in limited settings, and often is inevitable to resort to numerical solutions. Here, we develop a computational approach for simulating the time evolution of Fokker—Planck solutions in terms of a mean field limit of an interacting particle system. The interactions between particles are determined by the gradient of the logarithm of the particle density, approximated here by a novel statistical estimator. The performance of our method shows promising results, with more accurate and less fluctuating statistics compared to direct stochastic simulations of comparable particle number. Taken together, our framework allows for effortless and reliable particle-based simulations of Fokker–Planck equations in low and moderate dimensions. The proposed gradient–log–density estimator is also of independent interest, for example, in the context of optimal control.

Keywords: stochastic systems; Fokker-Planck equation; interacting particles; multiplicative noise; gradient flow; Stochastic differential equations

1. Introduction

The Fokker–Planck equation (FPE) describes the evolution of the probability density function (PDF) for the state variables of dynamical systems modelled by stochastic differential equations (SDE). Fokker–Planck equations are widely used for modelling stochastic phenomena in various fields, such as, for example, in physics, finance, biology, neuroscience, traffic flow [1]. Yet, explicit closed-form solutions of FPE are rarely available [2], especially in settings where the underlying dynamics is nonlinear. In particular, exact analytic solutions may be obtained only for a restricted class of systems following linear dynamics perturbed by white Gaussian noise, and for some nonlinear Hamiltonian systems [3,4].

Existing numerical approaches for computing Fokker–Planck solutions may be grouped into three broad categories: grid based, semi-analytical, and sample based methods. The first category, comprises mainly finite difference and finite element methods [5,6]. These frameworks, based on integration of FPE employing numerical solvers for partial differential equations, entail computationally demanding calculations with inherent finite spatial resolution [7].

Conversely, semi-analytical approaches try to reduce the number of required computations by assuming conditional Gaussian structures [8], or by employing cumulant neglect closures [9], statistical linearisation [10,11], or stochastic averaging [12]. Although efficient for the settings they are devised for, their applicability is limited, since, the resulting solutions are imprecise or unstable in certain settings.

On the other hand, in the sample based category, Monte Carlo methods resort to stochastic integration of a large number of *independent* stochastic trajectories that as an ensemble represent the probability density [13,14]. These methods are appropriate for computing *unbiased* estimates of exact expectations from empirical averages. Nevertheless, as we show in the following, cumulants of resulting distributions exhibit strong temporal fluctuations, when the number of simulated trajectories is not sufficiently large.

Surprisingly, there is an alternative sample based approach built on *deterministic* particle dynamics. In this setting, the particles are not independent, but they rather *interact* via an (approximated) probability density, and the FPE describes the mean field limit, when their number grows to infinity. This approach introduces a bias in the approximated expectations, but significantly reduces the variance for a given particle number.

Recent research, see e.g. [15–18], has focused on particle methods for models of thermal equilibrium, where the stationary density is known analytically. For these models, interacting particle methods have found interesting new applications in the field of probabilistic Bayesian inference: by treating the Bayesian posterior probability density as the stationary density of a FPE, the particle dynamics provides posterior samples in the long time limit. For this approach, the particle dynamics is constructed by exploiting the gradient structure of the probability flow of the FPE. This involves the relative entropy distance to the equilibrium density as a Lyapunov function. Unfortunately, this structure does not apply to general FPEs in *non–equilibrium* settings, where the stationary density is usually unknown.

In this article, we introduce a framework for interacting particle systems that may be applied to general types of Fokker–Planck equations. Our approach is based on the fact that the instantaneous effective force on a particle due to diffusion is proportional to the *gradient* of the *logarithm* of the exact probability *density* (GLD). Rather than computing a differentiable estimate of this density (say by a kernel density estimator), we estimate the GLD directly without requiring knowledge of a stationary density. Thereby, we introduce an approximation to the effective force acting on each particle, which becomes exact in the large particle number limit given the consistency of the estimator.

Our approach is motivated by recent developments in the field of machine learning, where GLD estimators have been studied independently and are used to fit probabilistic models to data. An application of these techniques to particle approximations for FPE is, to our knowledge, new. [1] Furthermore, our method provides also straightforward approximations of entropy production rates, which are of primary importance in non–equilibrium statistical physics [20].

This article is organised as follows: Section 2 describes the deterministic particle formulation of the Fokker–Planck equation. Section 3 shows how a gradient of the logarithm of a density may be represented as the solution of a variational problem, while in Section 4 we discuss an empirical approximation of the gradient-log-density. In Section 5, we introduce function classes for which the variational problem may be solved explicitly, while in Section 6 we compare the temporal derivative of empirical expectations based on the particle dynamics with exact results derived from the Fokker–Planck equation. Section 7 is devoted to the class of equilibrium Fokker–Planck equations, where we discuss relations to Stein Variational Gradient Descent and other particle approximations of Fokker–Planck solutions. In Section 8, we show how our method may be extended to general diffusion processes with state dependent diffusion, while Section 9 discusses how our framework may be employed to simulate second order Langevin dynamics. In Section 10 we demonstrate various aspects of our method by simulating Fokker–Planck solutions for different dynamical models. Finally, we conclude with a discussion and an outlook in Section 11.

[1] The approach in [19] uses a GLD estimator different from ours for particle dynamics but with a probability flow towards equilibrium which is not given by a standard FPE.

2. Deterministic particle dynamics for Fokker–Planck equations

We consider Fokker–Planck equations of the type

$$\frac{\partial p_t(x)}{\partial t} = -\nabla \cdot \left[f(x) p_t(x) - \frac{\sigma^2}{2} \nabla p_t(x)) \right] . \tag{1}$$

Given an initial condition $p_0(x)$, Eq. (1) describes the temporal development of the density $p_t(x)$ for the random variable $X(t) \in R^d$ following the stochastic differential equation

$$dX(t) = f(X(t))dt + \sigma dB(t) . \tag{2}$$

In Eq. (2), $f(x) \in R^d$ denotes the drift function characterising the deterministic part of the driving force, while $dB(t) \in R^d$ represents the differential of a vector of independent Wiener processes capturing stochastic, Gaussian white noise excitations. For the moment, we restrict ourselves to state independent and diagonal *diffusion matrices*, i.e. diffusion matrices independent of $X(t)$ (additive noise) with diagonal elements σ^2 characterising the noise amplitude in each dimension. Extensions to more general settings are deferred to Section 8.

We may rewrite the FPE Eq. (1) in the form of a *Liouville* equation

$$\frac{\partial p_t(x)}{\partial t} = -\nabla \cdot [g(x,t) \; p_t(x)] \tag{3}$$

for the *deterministic* dynamical system

$$\frac{dX}{dt} = g(X,t) \; , \qquad X(0) \sim p_0(x), \tag{4}$$

(dropping the time argument in $X(t)$ for simplicity) with velocity field

$$g(x,t) = f(x) - \frac{\sigma^2}{2} \nabla \ln p_t(x) . \tag{5}$$

Hence, by evolving an ensemble of N *independent* realisations of Eq. (4) (to be called 'particles' in the following) according to

$$\frac{dX_i}{dt} = g(X_i,t) \; , \qquad i = 1, \dots, N \qquad X_i(0) \sim p_0(x), \tag{6}$$

we obtain an empirical approximation to the density $p_t(x)$.

Since the only source of randomness in Eq. (4) can be attributed to the initial conditions $X_i(0)$, averages computed from the particle approximation (Eq. (6)) are expected to have smaller variance compared to N independent simulations of the SDE (Eq. (2)). Unfortunately, this approach requires perfect knowledge of the unknown instantaneous density $p_t(x)$ (c.f. Eq. (5)), that is actually the quantity we want to compute.

Here, we circumvent this issue by introducing *statistical estimators* for the term $\nabla \ln p_t(x)$, computed from the entire ensemble $(X_1(t)), \dots, X_N(t))$ of particles at time t. Although this additional approximation introduces interactions among the particles via the estimator, for sufficiently large particle number N, fluctuations of the estimator are expected to be negligible and the limiting dynamics should converge to its mean field limit (Eq. (4)) provided the estimator is asymptotically consistent. Thus, rather than computing a differentiable approximation to $p_t(x)$ from the particles, e.g. by a kernel density estimator, we show in the following section, how the function $\nabla \ln p_t(x)$ may be directly estimated from samples of $p_t(x)$.

3. Variational representation of gradient–log–densities

To construct a gradient–log–density (GLD) estimator we rely on a variational representation introduced by *Hyvärinen* in his *score–matching* approach for the estimation of non–normalised statistical models [21]. We favoured this approach over other estimators [22,23] due to its flexibility to adapt to different function classes chosen to approximate the GLD.

Here, we use a slightly more general representation compared to [21] allowing for an extra arbitrary reference function $r(x) = (r^{(1)}(x), \dots, r^{(d)}(x))$ such that the component α of the gradient is represented as

$$\partial_\alpha \ln p(x) = r^{(\alpha)}(x) + \arg\min_{\phi} \; \mathcal{L}^r_\alpha[\phi, p], \tag{7}$$

where $\partial_\alpha \doteq \frac{\partial}{\partial x^{(\alpha)}}$ stands for the partial derivative with respect to coordinate α of the vector $x \equiv (x^{(1)}, \dots x^{(d)})$.

The cost function is defined as an expectation with respect to the density $p(x)$ by

$$\mathcal{L}^r_\alpha[\phi, p] = \int p(x) \left(\phi^2(x) + 2r^{(\alpha)}(x)\phi(x) + 2\partial_\alpha \phi(x) \right) dx \,, \tag{8}$$

with dx representing the volume element in R^d. To obtain this relation, we use integration by parts (assuming appropriate behaviour of densities and ϕ at boundaries), and get

$$\begin{aligned} \mathcal{L}^r_\alpha[\phi, p] \;&=\; \int p(x) \left(\phi(x) + r^{(\alpha)}(x) - \partial_\alpha \ln p(x) \right)^2 dx \\ &-\; \int p(x) \left(\partial_\alpha \ln p(x) - r^{(\alpha)}(x) \right)^2 dx. \end{aligned} \tag{9}$$

Minimisation with respect to ϕ yields Eq. (7).

4. Gradient–log–density Estimator

To transform the variational formulation into a GLD estimator based on N sample points $(X_1, \dots, X_N)$, we replace the density $p(x)$ in Eq. (8) by the empirical distribution $\hat{p}_t(x) = \frac{1}{N} \sum_{i=1}^N \delta(x - X_i(t))$, i.e.

$$\mathcal{L}^r_\alpha[\phi, p_t] \approx \mathcal{L}^r_\alpha[\phi, \hat{p}_t] = \frac{1}{N} \sum_{i=1}^N \left(\phi^2(X_i) + 2r^{(\alpha)}(X_i)\phi(X_i) + 2\partial_\alpha \phi(X_i) \right) \,, \tag{10}$$

and

$$\partial_\alpha \ln p_t(x) \approx r^{(\alpha)} + \arg\min_{\phi \in \mathcal{F}} \; \mathcal{L}^r_\alpha[\phi, \hat{p}_t] \,, \tag{11}$$

where $\mathcal{F}$ is an appropriately chosen family of functions with controllable complexity. By introducing the estimator of Eq. (11) in Eq. (6), we obtain a particle representation for the Fokker–Planck equation

$$\frac{dX_i^{(\alpha)}}{dt} = f^{(\alpha)}(X_i) - \frac{\sigma^2}{2} \left(r^{(\alpha)}(X_i) + \arg\min_{\phi \in \mathcal{F}} \; \mathcal{L}^r_\alpha[\phi, \hat{p}_t] \right) \,, \tag{12}$$

for $i = 1, \dots, N$ and $\alpha = 1, \dots, d$, with

$$\hat{p}_t(x) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - X_i)\,, \qquad X_i(0) \sim p_0(x).$$

Although, in this article, we use $r \equiv 0$ for all simulated examples, the choice $r(x) = \frac{2}{\sigma^2} f(x)$, which cancels the first two terms in Eq. (12), leads to interesting relations with other particle approaches for simulating Fokker–Planck solutions for equilibrium systems (c.f. Section 7).

4.1. Estimating the entropy rate

Interestingly, the variational approach provides us with a simple, built in method for computing the entropy rate (temporal change of entropy) of the stochastic process (Eq. (2)).

Using the FPE (1) and integration by parts one can derive the well known relation, see e.g. [24],

$$-\frac{d}{dt}\int p_t(x)\ln p_t(x)dx = \frac{\sigma^2}{2}\sum_{\alpha=1}^{d}\int p_t(x)\left(\partial_\alpha \ln p_t(x)\right)^2 dx + \int p_t(x)\nabla\cdot f(x)dx\,. \tag{13}$$

The first term on the right hand side is usually called entropy production, whereas the second term corresponds to the entropy flux. In the stationary state, the total entropy rate vanishes. For equilibrium dynamics, both terms vanish individually at stationarity. This should be compared to the minimum of the cost function (Eq. (9)) which for $r \equiv 0$ equals

$$\min_{\phi} \mathcal{L}^0_\alpha[\phi, p_t] = -\int p_t(x)\left(\partial_\alpha \ln p_t(x)\right)^2 dx\,. \tag{14}$$

Thus we obtain the estimator

$$-\frac{d}{dt}\int p_t(x)\ln p_t(x)dx \approx -\frac{\sigma^2}{2}\sum_{\alpha=1}^{d}\min_{\phi}\mathcal{L}^0_\alpha[\phi,\hat{p}_t] + \frac{1}{N}\sum_{i=1}^{N}\nabla\cdot f(X_i)\,. \tag{15}$$

We will later see for the case of equilibrium dynamics that a similar method may be employed to approximate the relative entropy distance to the equilibrium density.

5. Function classes

In the following, we discuss choices for families of functions $\mathcal{F}$ leading to explicit, closed form solutions for estimators.

5.1. Linear models

A simple possibility is to choose linearly parametrised functions of the form

$$\phi(x) = \sum_{l=1}^{m} a_k \phi_k(x)\,, \tag{16}$$

where the $\phi_k(x)$ are appropriate basis functions, e.g. polynomials, radial basis functions or trigonometric functions. For this linear parametrisation, the empirical cost (Eq. (10)) is quadratic in the parameters a_k and can be minimised explicitly. A straightforward computation shows that

$$\frac{dX_i}{dt} = f(X_i) - \frac{\sigma^2}{2}r(X_i) + \frac{\sigma^2}{2}\sum_{k,j=1}^{m}(C^{-1})_{kj}\phi_k(X_i)\sum_{l=1}^{N}\left\{\nabla\phi_j(X_l) + \phi_j(X_l)r(X_l)\right\}\,, \tag{17}$$

with $C_{kl} = \sum_{i=1}^{N} \phi_k(X_i)\phi_l(X_i)$.

Obviously, we require the number of samples to be greater than the number of employed basis functions, i.e. $N \geq m+1$, to have a non–singular matrix C. This restriction can be lifted by introducing an additional penalty for regularisation. Eq. (17) is independent of the reference function r, when r belongs to the linear span of the selected basis functions. However, this model class with a finite parameter number has limited complexity. Thus, even when the sample number N grows large, we do not expect, in general, convergence to the mean field limit.

5.2. Kernel approaches

Here, we consider a family $\mathcal{F}$ of functions for which the effective number of parameters to be computed is not fixed beforehand, but rather increases with the sample number N: a *reproducing kernel Hilbert space* (RKHS) of functions defined by a positive definite (Mercer) kernel $K(\cdot,\cdot)$. Statistical models based on such function spaces have played a prominent role in the field of machine learning in recent years [25].

A common, kernel based approach to regularise the minimisation of empirical cost functions is via penalisation using the RKHS norm $\|\cdot\|_{\text{RKHS}}$ of functions in $\mathcal{F}$. This can also be understood as penalised version of a linear model (16) with infinitely many feature functions ϕ_k. For so called universal kernels [26] this unbounded complexity suggests that we could expect asymptotic convergence of the GLD estimator (see [27] for related results) and a corresponding convergence of the particle model to its mean field limit. However, a rigorous proof may not be trivial, since particles in our setting are not independent.

The explicit form of the kernel based approximation is given by

$$\partial_\alpha \ln p(x) \approx r^{(\alpha)}(x) + \arg\min_{\phi\in\mathcal{F}} \left\{ \mathcal{L}^r_\alpha[\phi,\hat{p}] + \frac{\lambda}{N}\|\phi\|^2_{\text{RKHS}} \right\}, \tag{18}$$

where the parameter λ controls the strength of the penalisation. Again, this optimisation problem can be solved in closed form in terms of matrix inverses. One can prove a *representer theorem* which states that the minimiser $\phi(x)$ in Eq. (18) is a linear combination of kernel functions evaluated at the sample points X_i, i.e.,

$$\phi(x) = \sum_{i=1}^{N} a_i K(x, X_i)\,. \tag{19}$$

For such functions, the RKHS norm is given by

$$\|\phi\|^2_{\text{RKHS}} = \sum_{i,j=1}^{N} a_i a_j K(X_i, X_j)\,. \tag{20}$$

Hence, this representation leads again to a quadratic form in the N coefficients.

A short computation yields

$$a_j = -\sum_{k=1}^{N} \left((K^2 + \lambda K)^{-1} \right)_{jk} \sum_{l=1}^{N} \left\{ \partial_{\alpha_l} K(X_l, X_k) + K(X_l, X_k) r^{(\alpha)}(X_l) \right\}, \tag{21}$$

where $K_{ij} \doteq K(X_i, X_j)$. Similar approaches for kernel based GLD estimators have been discussed in [22,23]. For $r = 0$, Eq. (21) agrees with the GLD estimator of [22] derived by inverting *Stein's* equation, or by minimising the *Kernelised Stein discrepancy*.

The resulting particle dynamics is given by

$$\frac{dX_i}{dt} = f(X_i) - \frac{\sigma^2}{2} r(X_i) + \frac{\sigma^2}{2} \sum_{k=1}^{N} \left((K + \lambda I)^{-1} \right)_{ik} \sum_{l=1}^{N} \{ \nabla_l K(X_l, X_k) + K(X_l, X_k) r(X_l) \} \,. \tag{22}$$

Note that here also the inverse matrix depends on the particles X_k. In the limit of small λ, the right hand side becomes independent of the reference function r.

In the present article, we employ Gaussian radial basis function (RBF) kernels given by

$$K(x, x') = \exp\left[-\frac{1}{2\, l^2} \| x - x' \|^2 \right] \,, \tag{23}$$

with a length scale l. A different possibility would be given by kernels with a *finite dimensional* feature representation

$$K(x, x') = \sum_{j=1}^{m} \phi_j(x) \phi_j(x') \,, \tag{24}$$

which may also be interpreted as a *linear model* as in Eq. 16 with a L_2 penalty on the unknown coefficients.

5.3. A sparse kernel approximation

The inversions of the $N \times N$ matrices in Eq. (22) have to be performed at each step of a time discretised ODE system (Eq. (22)). For large N, the cubic complexity could become too time consuming. Hence, here, we resort to a well established approximation in machine learning to overcome this issue, by applying a sparse approximation to the optimisation problem of Eq. (18), see e.g. [28]. In particular, we introduce a smaller set of $M \ll N$ *inducing points* $\{z_k\}_{k=1}^{M}$, that need not necessarily be a subset of the N particles. We then minimise the penalised cost function (Eq. 18) in the finite dimensional family of functions

$$\phi(x) = \sum_{i=1}^{M} a_i K(x, z_i) \,. \tag{25}$$

This may also be understood as a special linear parametric approximation. To keep matrices well conditioned, in practice we add a small 'jitter' term to Eq. (18), i.e., we use

$$\lambda \|\phi\|^2_{\text{RKHS}} + \epsilon \|\phi\|_2^2 \,, \tag{26}$$

as the total penalty. In the limit $\lambda, \epsilon \to 0$, this representation reduces to an approximation of the form of Eq. (16) with M basis functions $K(\cdot, z_l)$ for $l = 1, \ldots, M$.

By introducing the matrices

$$K^{zz}_{kl} \doteq K(z_k, z_l) + \epsilon \delta_{kl} \qquad K^{xz}_{ij} \doteq K(X_i, z_j) \,, \tag{27}$$

and

$$A \doteq K^{xz} \left[(\lambda + \epsilon) I + (K^{zz})^{-1} (K^{xz})^\top (K^{xz}) \right]^{-1} (K^{zz})^{-1} \,, \tag{28}$$

we replace the particle dynamics of Eq. (22) by

$$\frac{dX_i}{dt} = f(X_i) - \frac{\sigma^2}{2} r(X_i) + \frac{\sigma^2}{2} \sum_k A_{ik} \sum_l \{ \nabla_l K(X_l, z_k) + K(X_l, z_k) r(X_l) \} \,. \tag{29}$$

Hence, for this approximation we have to invert only $M \times M$ matrices. For fixed M, the complexity of the GLD estimator is limited. Results for log–density–estimators in machine learning (obtained for independent data) indicate that for a moderate growth of the number of inducing points M with the number of particles N, similar approximation rates may be obtained as for full kernel approaches.

6. A note on expectations

In this section we present a preliminary discussion of the quality of the particle method to approximate expectations of scalar functions h of the random variable $X(t)$. We concentrate on the temporal development of $h(X(t))$. While it would be important to obtain an estimate of the approximation error over time, we will defer such an analysis to future publications and only concentrate on a result for the first time derivative of expectations, i.e. the evolution over infinitesimal times.

Using the FPE (Eq. (1)) and integrations by part one derives the exact result

$$\frac{d\langle h(X)\rangle}{dt} = \langle L_x h(X)\rangle\,, \tag{30}$$

where $\langle\cdot\rangle$ denotes the expectation with respect to $p_t(x)$ and the operator L_x equals the *generator* of the process, i.e.,

$$L_x \doteq f(X)\cdot\nabla + \frac{\sigma^2}{2}\nabla^2\,. \tag{31}$$

To obtain a related result for empirical expectations based on particles, we employ the relation

$$\frac{dh(X_i)}{dt} = \nabla h(X_i)\cdot\frac{dX_i}{dt} \tag{32}$$

and a direct computation using the dynamics of Eq. (17) and Eq. (22) yields the result

$$\frac{d\langle h(X)\rangle_{\hat{p}_t}}{dt} = \langle L_x h(X)\rangle_{\hat{p}_t} + \Delta\,, \tag{33}$$

where $\langle\cdot\rangle_{\hat{p}_t}$ denotes expectation with respect to the empirical distribution $\hat{p}_t$ of the particles. Hence, if the remainder Δ is small, the change of empirical particle averages should not deviate much from the corresponding exact ones. This remainder term is given by

$$\Delta = \frac{\sigma^2}{2}\langle(r(X)+\nabla)\cdot\left(\hat{\nabla}h(x)-\nabla h(x)\right)\rangle_{\hat{p}_t}\,, \tag{34}$$

where $\hat{\nabla}h(x)$ stands for the approximation of the vectorial function $\nabla h(x)$ based on the 'data' $\nabla h(X_l)$ using regression with a linear combination of basis functions $\phi_h(x)$ or by regularised kernel regression. The explicit formulas for the two cases are

$$\hat{\nabla}h(x) = \sum_{j,k=1}^{M}\phi_j(x)\left(C^{-1}\right)_{jk}\sum_l \phi_k(x_l)\nabla h(X_l) \tag{35}$$

and

$$\hat{\nabla}h(x) = \sum_{j,k=1}^{N}K(x,X_j)\left((K+\lambda I)^{-1}\right)_{jk}\nabla h(X_k)\,, \tag{36}$$

respectively. If $\nabla h(x)$ is well approximated by basis functions, the remainder Δ is small. If indeed $\nabla h(x) = \sum_{n=1}^{M} c_n\phi_n(x)$, for some $c_n \in R^d$, the remainder term vanishes, $\Delta = 0$. By its similarity to the finite basis function model, this result should also be valid for the sparse kernel dynamics of Eq. (29), when the penalty λ is small. One might conjecture that the temporal development of expectations for reasonably smooth functions might be faithfully represented by the particle dynamics. This conjecture is supported by our numerical results.

7. Equilibrium dynamics

An important class of stochastic dynamical systems describe *thermal equilibrium*, for which the drift function f is the negative gradient of a potential U, while the limiting equilibrium density p_∞ is explicitly given by a Gibbs distribution:

$$f(x) = -\nabla U(x) \tag{37}$$

$$\nabla \ln p_\infty(x) = \frac{2}{\sigma^2} f(x). \tag{38}$$

For this class of models, our method provides a simple and built in estimator for the relative entropy between the instantaneous, p_t, and the equilibrium density, p_∞. As we discuss here, our framework may also be related to two other particle approaches, that converge to the (approximate) equilibrium density.

7.1. Relative entropy

The relative entropy or *Kullback–Leibler divergence* is defined as

$$D(p_t|p_\infty) \doteq \int p_t(x) \ln \frac{p_t(x)}{p_\infty(x)} dx\,. \tag{39}$$

Following a similar calculation that led to Eq. (13), we obtain

$$\begin{aligned} \frac{d}{dt} D(p_t|p_\infty) &= -\frac{\sigma^2}{2} \int p_t(x) \left\| \nabla \ln p_t(x) - \nabla \ln p_\infty(x) \right\|^2 dx \\ &= -\frac{2}{\sigma^2} \int p_t(x) \|g(x,t)\|^2 dx\,, \end{aligned} \tag{40}$$

where $g(x,t)$ indicates the velocity field of the particle system defined in Eq. (4). The first equality holds for arbitrary drift functions. To obtain the second equality, we have inserted the explicit result for p_∞.

Hence, we may compute the relative entropy at any time T as a time integral

$$D(p_T|p_\infty) = D(p_0|p_\infty) - \frac{2}{\sigma^2} \int_0^T \left\{ \int p_t(x) \|g(x,t)\|^2 dx \right\} dt\,, \tag{41}$$

where the inner expectation is easily approximated by our particle algorithm. This result shows that the exact velocity field $g(x,t)$ converges to 0 for $t \to \infty$ and one expects particles to also converge to fixed points. For other, non–equilibrium systems asymptotic fixed points are, however, the exception.

7.2. Relation to Stein Variational Gradient Descent

Recently, *Stein variational gradient descent* (SVGD), a kernel based particle algorithm, has attracted considerable attention in the machine learning community [29,30]. The algorithm is designed to provide approximate samples from a given density p_∞ as the asymptotic fixed points of a deterministic particle system. Setting $-\ln p_\infty(x) = U(x) + \text{const}$, SVGD is based on the dynamics

$$\frac{dX_i}{dt} = \sum_l \{-K(X_i, X_l)\nabla U(X_l) + \nabla_l K(X_i, X_l)\} \ . \tag{42}$$

This can be compared to our approximate FPE dynamics (Eq. (22)) for the equilibrium case by setting $\sigma^2 = 2$ and $r(x) = f(x) = -\nabla U(x)$. For this setting, both algorithms have in fact, the same conditions

$$\sum_l \{-K(X_i, X_l)\nabla U(X_l) + \nabla_l K(X_i, X_l)\} = 0\,, \tag{43}$$

for the 'equilibrium' fixed points. See [18] for a discussion of these fixed points for different kernel functions. However, both dynamics differ for finite times t, where a single time step of SVGD is computationally simpler, being free of the matrix inversion required by our framework. The mean field limit $N \to \infty$ of Eq. (42) differs from the FPE, and the resulting partial differential equation is nonlinear [31]. Nevertheless, it is possible to interpolate between the two particle dynamics. In fact, in the limit of a large regularisation parameter $\lambda \to \infty$, the inverse matrix in Eq. (22) becomes diagonal, i.e. $(K + \lambda I)^{-1} \simeq \frac{1}{\lambda} I$, and we recover SVGD (Eq. (42)) by introducing a rescaled time $\tau \doteq t/\lambda$. This result could be of practical importance when the goal is to approximate the stationary distribution, irrespectively of the finite time dynamics. The SVGD combines faster matrix operations with slower relaxation times to equilibrium compared to the FPE dynamics. It would be interesting to see, if an optimal computational speed of a particle algorithm might be achieved at some intermediate regularisation parameter λ.

7.3. Relation to geometric formulation of FPE flow

Following Otto [32] and Villani [33], the FPE for the equilibrium case can be viewed as a gradient flow on the manifold of probability densities with respect to the Wasserstein metric. This formulation can be used to define an implicit Euler time discretisation method for the dynamics of the density p_t. For small times δt (and $\sigma^2 = 2$) this is given by the variational problem

$$p_{t+\delta t} = \arg\inf_p \left(W_2^2(p, p_t) + \delta t D(p \| p_\infty) \right) \tag{44}$$

in terms of the Kullback–Leibler divergence and the L_2 *Wasserstein distance* $\mathcal{W}_2$. The latter gives the minimum of $\langle \|X - X(t)\|^2 \rangle$ for two random variables $X(t)$ and X where the expectation is over the joint distribution with fixed marginals p_t and p. Using the dual formulation for a regularised Wasserstein distance, approximate numerical algorithms for solving Eq. (44) have been developed by [34] and by [35] with applications to simulations of FPE.

We show in the following that Eq. (44) may be cast into a form closely related to our variational formulation (Eq. (7)) for $r(x) = f(x)$. Assuming that X and $X(t)$ are related through a deterministic (transport) mapping of the form

$$X = X(t) + \delta t \nabla \psi(X(t))\ , \tag{45}$$

we may represent the Wasserstein distance in terms of ψ and the variational problem may be rewritten as

$$p_{t+\delta t}(x) = p_t(x) - \delta t \nabla \left(p_t(x) \nabla \psi^*(x) \right), \tag{46}$$

where

$$\psi^* = \arg\min_{\nabla\psi} \frac{\delta t^2}{2} \int \|\nabla\psi(x)\|^2 p_t(x) dx + \delta t D(p_{t+dt} \| p_\infty) \,. \tag{47}$$

To proceed, we expand the relative entropy to first order in δt, inserting the representation Eq. (46) for $p_{t+\delta t}(x)$, obtaining thereby

$$\begin{aligned} \frac{\delta t}{2} \int \|\nabla\psi(x)\|^2 p_t(x) dx + D(p_{t+\delta t} \| p_\infty) = D(p_t \| p_\infty) + \\ + \frac{\delta t}{2} \left(\int p_t(x) \left\{ \|\nabla\psi(x)\|^2 - 2\nabla^2\psi(x) + 2\nabla U(x) \cdot \nabla\psi(x) \right\} dx \right) + \\ + O(\delta t^2). \end{aligned} \tag{48}$$

Minimisation ignoring the $O(\delta t^2)$ terms (employing integration by parts) yields

$$\nabla\psi^*(x) = -\nabla U(x) - \nabla \ln p_t(x), \tag{49}$$

which is closely related to our cost function Eq. (8), if we identify $\phi(x) = -\nabla\psi(x)$. By replacing p_t by samples, the empirical cost function may be regularised with a RKHS norm penalty resulting in a nonparametric estimator for unnormalised log–density $\psi^*(x) = -\ln p_t(x) - U(x) + \text{const}$ as shown in [36]. One could use this estimator as an alternative to our approach. This would lead to a simultaneous estimate of all components of the GLD. In our approach, each of the d components of the gradient is computed individually. In this way, we avoid additional second derivatives of kernels, which would increase the dimensionality of the resulting matrices.

8. Extension to general diffusion processes

The Fokker–Planck equations for an SDE with arbitrary drift $f(x)$ and general, state dependent diffusion matrix $D(x)$ is given by

$$\frac{\partial p_t(x)}{\partial t} = \nabla \cdot \left[-f(x) p_t(x) + \frac{1}{2} \nabla \cdot (D(x)\; p_t(x)) \right]. \tag{50}$$

This may again be written in the form of a Liouville equation (Eq. (3)) where the effective force term equals

$$g(x,t) = f(x) - \frac{1}{2}\nabla \cdot D(x) - \frac{1}{2} D(x) \nabla \ln p_t(x). \tag{51}$$

9. Second order Langevin dynamics (Kramer's equation)

For second order Langevin equations, the system state comprises positions $X \in R^d$ and velocities $V \in R^d$ following the coupled SDE

$$dX = V dt \tag{52}$$

$$dV = (-\gamma V + f(X))\, dt + \sigma dB_t \,. \tag{53}$$

In Eq. (52), the dynamics describe the effect of a friction force, γV, an external force, $f(X)$, and a fluctuating force, where γ denotes the dissipation constant. In this setting, the effective *deterministic* ODE system is given by

$$\begin{aligned}\frac{dX}{dt} &= V \\ \frac{dV}{dt} &= -\gamma V + f(\mathrm{X}) - \frac{\sigma^2}{2}\nabla_v \ln p_t(\mathrm{X}, V)\ .\end{aligned} \tag{54}$$

Considering here the equilibrium case, we set $f(x) = -\nabla U(x)$ for which the stationary density equals

$$\ln p_\infty(X, V) = -\beta\left(\frac{\|V\|^2}{2} + U(X)\right) \equiv -\beta H(X, V), \tag{55}$$

where $\beta = \frac{2\gamma}{\sigma^2}$ and $H(x, v) = \frac{\|V\|^2}{2} + U(x)$ denotes the *Hamiltonian* function. Inserting p_∞ into Eq. (54), we find that for $t \to \infty$, the damping and the density dependent part of the force cancel and we are left with pure Hamiltonian dynamics

$$\begin{aligned}\frac{dX}{dt} &= V \\ \frac{dV}{dt} &= -\nabla U(X),\end{aligned} \tag{56}$$

for which all particles become completely *decoupled*, with each one conserving energy separately. Of course, this result also precludes fixed point solutions to the particle dynamics.

The asymptotic behaviour is also reflected in the expression for the change of the relative entropy for Kramer's equation. Similar to Eq. (40) we obtain

$$\begin{aligned}\frac{d}{dt}D(p_t|p_\infty) = -\frac{\sigma^2}{2}\int p_t(x, v)\,\|\nabla_v \ln p_t(x, v) - \nabla_v \ln p_\infty(x, v)\|^2\, dxdv \\ = -\frac{2}{\sigma^2}\int p_t(x, v)\|\gamma v + \frac{\sigma^2}{2}\nabla_v \ln p(x, v)\|^2 dxdv.\end{aligned} \tag{57}$$

When the system approaches equilibrium, both terms in the norm cancel out and the entropy production rate converges to 0.

10. Simulating accurate Fokker–Planck solutions for model systems

To demonstrate the accuracy of our approach, we simulated solutions of FPEs for a range of model systems and compared the results with those obtained from direct stochastic simulations (Monte Carlo sampling) of same particle number, and analytic solutions, where relevant. We tested our framework on systems with diverse degrees of nonlinearity and dimensionality, as well as with various types of noise (additive/multiplicative). We quantified the accuracy of transient and steady state solutions resulting from our method in terms of 1-Wasserstein distance [33] and Kullback Leibler (KL) divergence (Appendix C and D), along with squared error of distances between distribution cumulants. For evaluating particle solutions for nonlinear processes, where analytical solutions of the Fokker–Planck equation are intractable, we simulated a very large number (N^{∞}) of stochastic trajectories that we considered as ground truth Fokker–Planck solutions. We employed an Euler–Maruyama and forward Euler integration scheme of constant step size $dt = 10^{-3}$ for stochastic and deterministic simulations respectively.

10.1. Linear conservative system with additive noise

For a two dimensional Ornstein-Uhlenbeck process (Appendix A.1) transient and stationary densities evolved through deterministic particle simulations (D) consistently outperformed their stochastic counterparts (S) comprising same number of particles in terms accuracy in approximating the underlying density (Fig. 2). In particular, comparing the 1-Wasserstein distance between samples from analytically derived densities (P_t^A) (Appendix B) - considered here to reflect the ground truth - and the deterministically (D) or stochastically (S) evolved densities (P_t^N), $\mathcal{W}_1(P_t^A, P_t^N)$, we observed smaller Wasserstein distances to ground truth for densities evolved according to our deterministic particle dynamics, both for transient (Fig. 2(a.)) and stationary (Fig. 2(c.)) solutions. Specifically, we quantified the transient deviation of simulated densities from ground truth by the average temporal 1-Wasserstein distance, $\langle \mathcal{W}_1(P_t^A, P_t^N) \rangle_t$. For small particle number, deterministically evolved interacting particle trajectories represented more reliably the evolution of the true probability density compared to independent stochastic ones, as portrayed by smaller average Wasserstein distances. For increasing particle number the accuracy of the simulated solutions with the two approaches converged. Yet, although for $N = 2500$ particles the stochastically evolved densities suggest *on average* (over trials) comparable approximation precision with their deterministic counterparts, the deterministically evolved densities delivered more reliably densities of a certain accuracy, as proclaimed by the smaller dispersion of Wasserstein distances among different realisations (Fig. 2(a., c.)).

Likewise, we observed similar results when comparing only the stationary distributions, $\mathcal{W}_1(P_\infty^A, P_\infty^N)$ (Fig. 2(c.)). While for small particle number, the interacting particle system more accurately captured the underlying limiting distribution, for increasing particle number the accuracy of both approaches converged, with our method delivering consistently more reliable approximations among individual repetitions.

Moreover, densities evolved with our deterministic framework exhibited less fluctuating cumulant trajectories in time, compared to their stochastic counterparts (Fig. 1(c.)). In particular, even for limited particle number cumulants calculated over deterministically evolved particles progressed smoothly in time, while substantially more particles for the stochastic simulations were required for the same temporal cumulant smoothness. To quantify further the transient accuracy of Fokker–Planck solutions computed with our method, we compared the average transient discrepancy between the first two analytic cumulants (m_t and C_t) to those estimated from the particles ($\hat{m}_t$ and $\hat{C}_t$), $\langle \|\hat{m}_t - m_t\|_2 \rangle_t$ (Fig. 1(b.)) and $\langle \|\hat{C}_t - C_t\|_F \rangle_t$ (Fig. 1(d.)). In line with our previous results, our deterministic framework delivered considerably more accurate transient cumulants, when compared to stochastic simulations, with more

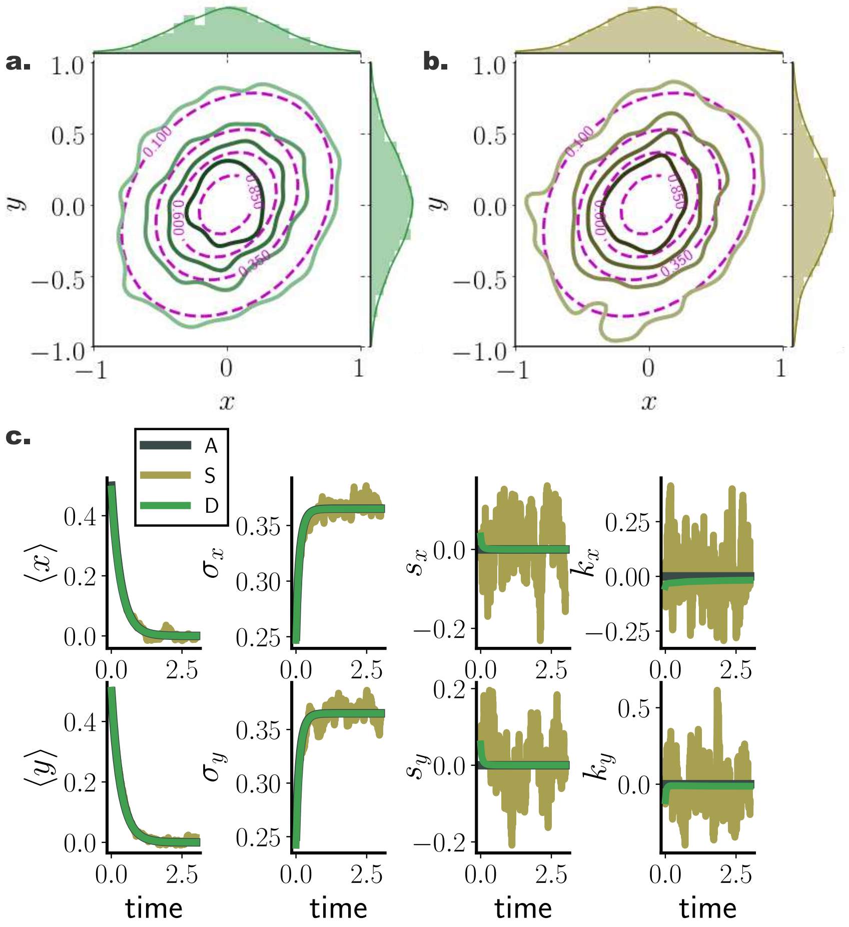

Figure 1. Stationary and transient Fokker–Planck solutions computed with deterministic (green) and stochastic (brown) particle dynamics for a two dimensional Ornstein Uhlenbeck process.(a.,b.) Estimated stationary PDFs arising from deterministic ($N = 1000$) (green), and stochastic ($N = 1000$) (brown) particle dynamics. Purple contours denote analytically calculated stationary distributions, while top and side histograms display marginal distributions for each dimension. **(c.)** Temporal evolution of marginal statistics, mean $\langle x \rangle$, standard deviation σ_x, skewness s_x, and kurtosis k_x, for analytic solution (A), and for stochastic (S) and deterministic (D) particle systems comprising $N = 1000$, with initial state distribution $\mathcal{N}\left(\begin{bmatrix}0.5\\0.5\end{bmatrix}, \begin{bmatrix}0.05^2 & 0\\0 & 0.05^2\end{bmatrix}\right)$, for $M = 100$ randomly selected inducing points employed in the gradient–log–density estimation. Deterministic particle simulations deliver smooth cumulant trajectories, as opposed to highly fluctuating stochastic particle cumulants.(Further parameter values: regularisation constant $\lambda = 0.001$, and RBF kernel length scale l estimated at every time point as two times the standard deviation of the state vector. Inducing point locations were selected randomly at each time step from a uniform distribution spanning the state space volume covered by the state vector.)

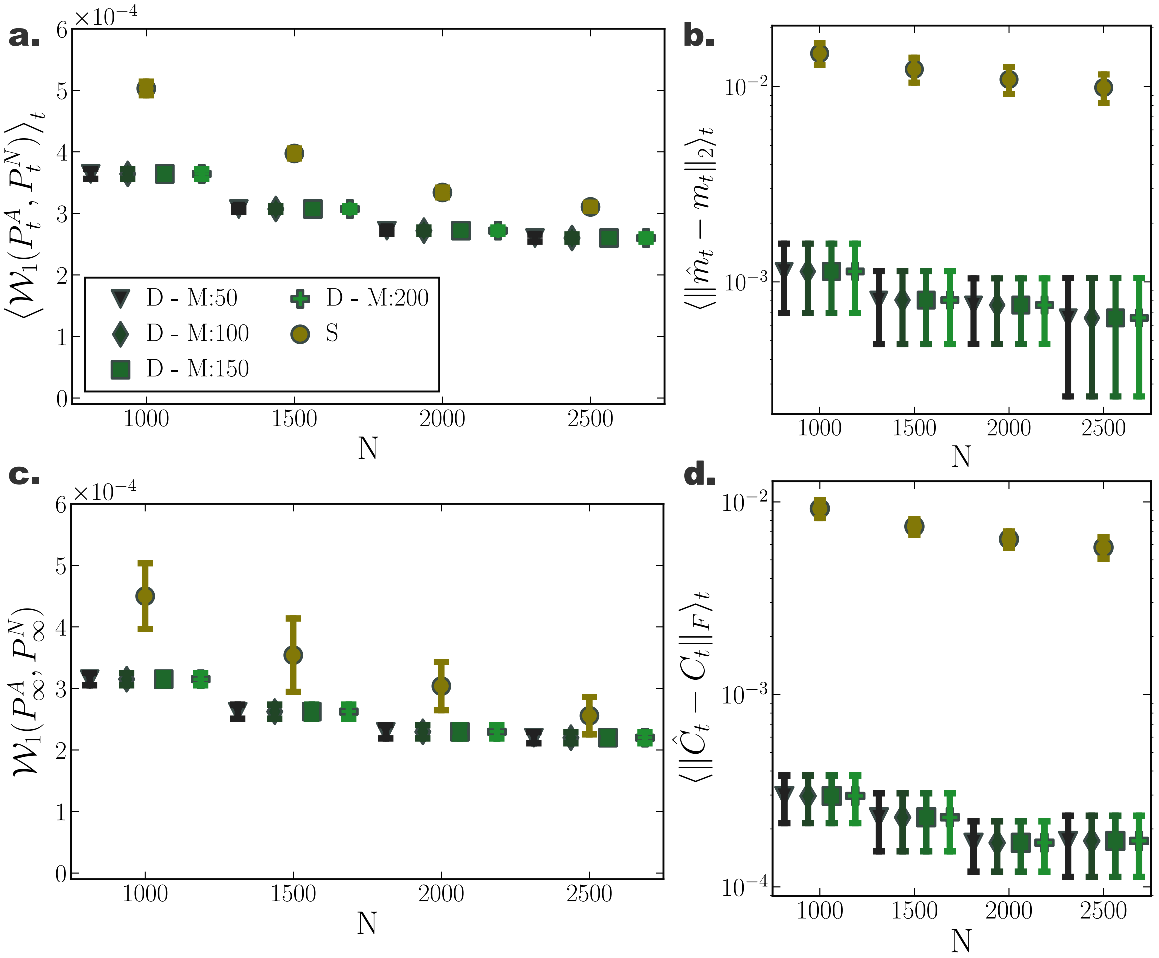

Figure 2. Accuracy of Fokker–Planck solutions for two dimensional Ornstein Uhlenbeck process. (a.) Mean, $\langle \mathcal{W}_1(P_t^A, P_t^N) \rangle_t$, and **(c.)** stationary $\mathcal{W}_1(P_\infty^A, P_\infty^N)$, 1-Wasserstein distance, between analytic solution and deterministic(D)/stochastic(S) simulations of N particles (for different inducing point number M). **(b.)** Average temporal deviations from analytic mean m_t and **(d.)** covariance matrix C_t for deterministic and stochastic system for increasing particle number N. Deterministic particle simulations consistently outperformed stochastic ones in approximating the temporal evolution of the mean and covariance of the distribution for all examined particle number settings. (Further parameter values: regularisation constant $\lambda = 0.001$, Euler integration time step $dt = 10^{-3}$, and RBF kernel length scale l estimated at every time point as two times the standard deviation of the state vector. Inducing point locations were selected randomly at each time step from a uniform distribution spanning the state space volume covered by the state vector.)

consistent results among individual realisations, denoted by smaller dispersion of average cumulant differences. (Notice the logarithmic y-axis scale in Fig. 1(b., d.). Error bars for the stochastic solutions were in fact larger than those for the deterministic solutions on a linear scale.)

Interestingly, the number of sparse points M employed in the gradient–log–density estimation had only minor influence on the quality of the solution (Fig. 2(a., c.)). This hints to substantially low computational demands for obtaining accurate Fokker–Planck solutions, since our method is computationally limited by the inversion of the $M \times M$ matrix in Eq. (28).

10.2. Bi-stable nonlinear system with additive noise

For nonlinear processes, since the transient solution of the FPE is analytically intractable, we compared the transient and stationary densities estimated by our method with those returned from stochastic simulations of $N^{\infty} = 2650$ particles, and contrasted them against stochastic simulations with same particle number.

For a system with bi-modal stationary distribution (Appendix A.2), the resulting particle densities from our deterministic framework closely agreed with those arising from the stochastic system with $N^{\infty} = 26000$ particles (Fig. 3(a.)). In particular, deterministically evolved distributions respected the symmetry of the underlying double–well potential, while the stochastic system failed to accurately capture the potential symmetric structure Fig. 3(a.iii.).

Systematic comparisons of the 1-Wasserstein distance between deterministic and stochastic N particle simulations with the "N^{∞}" stochastic simulation comprising 2650 particles, revealed that our approach efficiently captured the underlying PDF already with $N = 500$ particles (Fig. 3(c.,d.)). For increasing particle number, the two systems converged to the "N^{∞}" one. However, we observed a systematically increasing approximation accuracy delivered from the deterministic simulations compared to their stochastic counterparts.

It is noteworthy, that on average deterministic simulations of $N = 500$ particles conveyed a better approximation of the underlying transient PDF compared to stochastic simulations of $N = 2500$ particles (Fig. 3(c.)).

Interestingly, for small particle number, the number of employed inducing points M did not to influence significantly the accuracy of the approximated solution. However for increasing particle number, enlarging the set of inducing points contributed to more accurate approximation of Fokker–Planck equation solutions, with the trade off of additional computational cost.

Similar to the Ornstein Uhlenbeck process (Section 10.1), comparing cumulant trajectories computed from both the deterministic and stochastic particle systems revealed less fluctuating cumulant evolution for densities evolved with our deterministic framework also in this nonlinear setting (Fig. 3(b.)).

10.3. Nonlinear system perturbed by multiplicative noise

To asses the accuracy of our framework on general diffusion processes perturbed by state dependent (multiplicative) noise, we simulated a bi-stable system with dynamics governed by Eq. (A3) with diffusion function $D(x) = sin^2(x)$ according to Eq. (51). Also in this setting, deterministic particle distributions delivered a closer approximation of the underlying density, when compared to direct stochastic simulations. In particular, we found that in this setting, deterministically evolved distributions captured more accurately the tails of the underlying distribution, mediated here by stochastic simulations of $N^{\infty} = 35000$ particles (Fig. 4(a.,b.)).

Similar to the previously examined settings, the deterministic framework delivered more reliable and smooth trajectories for the marginal statistics of the underlying distribution (Fig. 4(c.)).

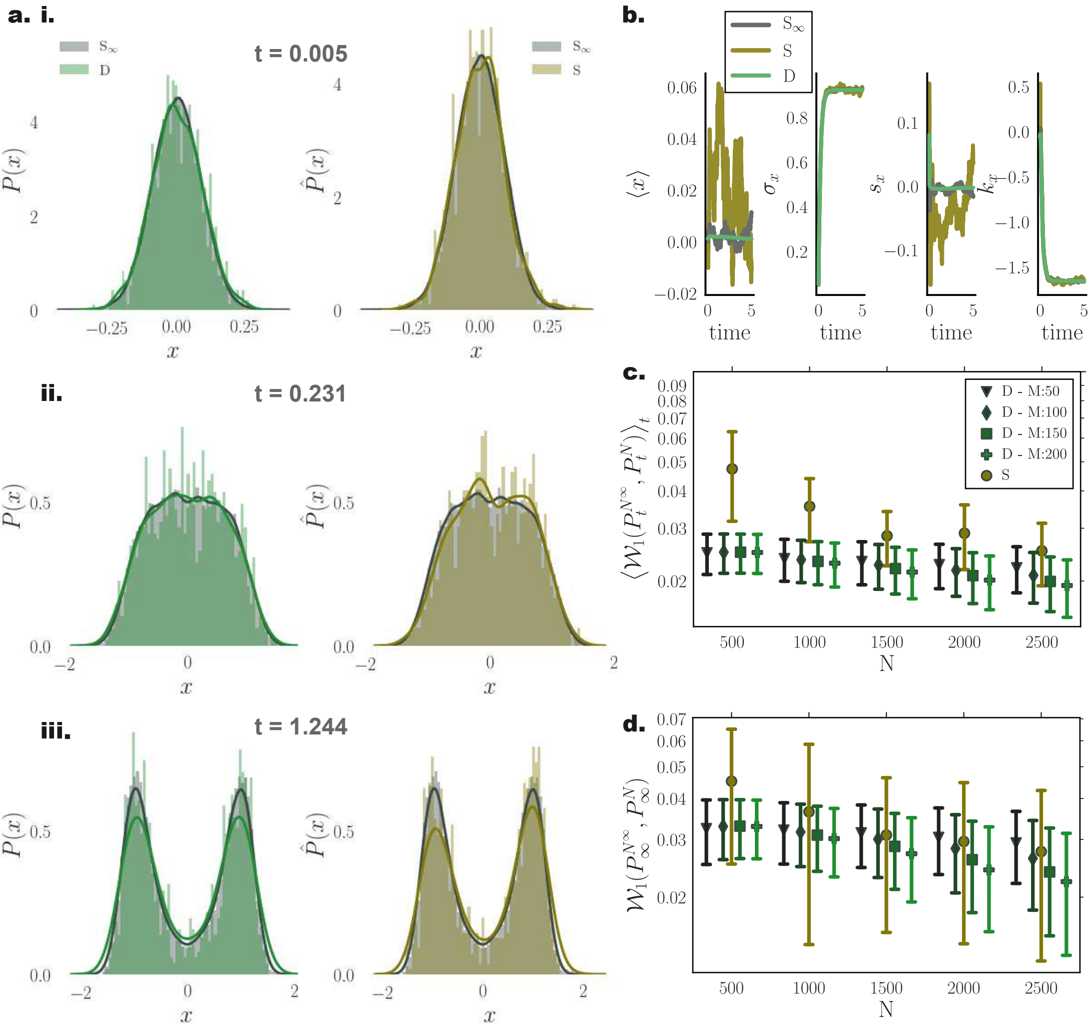

Figure 3. Performance of deterministic (green) and stochastic (brown) N particle solutions compared to N^∞ (grey) stochastic particle densities for a nonlinear bi-stable process. (a.) Instances of estimated pdfs arising from **(left)** stochastic ($N^\infty = 26000$) (grey) and deterministic ($N = 1000$) (green), and **(right)** stochastic ($N^\infty = 2650$) (grey) and stochastic ($N = 1000$) (brown) particle dynamics at times **(i.)** $t = 0.005$, **(ii.)** $t = 0.231$, and **(iii.)** $t = 1.244$. **(b.)** Temporal evolution of first four distribution cumulants, mean $\langle x \rangle$, standard deviation σ_x, skewness s_x, and kurtosis k_x, for stochastic (S^∞ and S) and deterministic (D) systems comprising $N^\infty = 26000$, $N = 1000$, with initial state distribution $\mathcal{N}(0, 0.05^2)$, by employing $M = 150$ inducing points in the gradient–log–density estimation. **(c.)** Mean, $\langle \mathcal{W}_1(P_t^{N^\infty}, P_t^N) \rangle_t$, and stationary, $\mathcal{W}_1(P_\infty^A, P_\infty^N)$, 1-Wasserstein distance, between $N^\infty = 2650$ stochastic, and deterministic (D)/stochastic (S) simulations of N particles (for different inducing point number M). (Further parameter values: regularisation constant $\lambda = 0.001$, Euler integration time step $dt = 10^{-3}$, and RBF kernel length scale $l = 0.5$. Inducing point locations were selected randomly at each time step from a uniform distribution spanning the state space volume covered by the state vector.)

Comparing the temporal average and stationary 1-Wasserstein distance (Fig. 4(d.,f.)) between the optimal stochastic distributions and the deterministic and stochastic particle distributions of size N, we found that the deterministic system delivered consistently more accurate approximations, as portrayed by smaller 1-Wasserstein distances.

Interestingly, we found that for deterministic particle simulations, the number of employed sparse points in the gradient–log–density estimation mediated a moderate approximation improvement for small system sizes, while for systems comprising more than $N = 2000$ particles, the number of sparse points had minimal or no influence on the accuracy of the resulting distribution (Fig. 4(e.,g.)).

10.4. Performance in higher dimensions

To quantify the scaling and performance of the proposed framework for increasing system dimension, we systematically compared simulated densities with analytically calculated ones for Ornstein–Uhlenbeck processes of dimension $D = \{2, 3, 4, 5\}$ following the dynamics of Eq. (A4). To evaluate simulated Fokker–Planck solutions we calculated Kullback–Leibler divergence between analytically evolved densities (Appendix B) and particle densities. We employed the closed form equation for estimating KL divergence between two Gaussian distributions (Appendix C) for empirically estimated mean, $\hat{m}_t$, and covariance, $\hat{C}_t$, for particle distributions.

For all dimensionalities, the deterministic particle solutions approximated transient and stationary densities remarkably accurately with Kullback–Leibler divergence between the simulated and analytically derived densities below 10^{-2} for all dimensions, both for transient and stationary solutions (Fig. 5(a.,d.). In fact, the deterministic particle solutions delivered more precise approximations of the underlying densities compared to direct stochastic simulations of the same particle number. Remarkably, even for processes of dimension $D = 5$ deterministically evolved solutions mediated through $N = 500$ particles resulted in approximately same KL divergence of stochastic particle solutions of $N = 6500$ particles.

Our deterministic particle method delivered consistently better approximations of the mean of the underlying densities compared to stochastic particle simulations (Fig. 5(b.,e.). Specifically, estimations of the stationary mean of the underlying distributions were more than two orders of magnitude accurate that their stochastically approximated counterparts already for small particle number (Fig. 5(e.).

Yet, the accuracy of our deterministic framework deteriorated for increasing dimension (Fig. 5(a.,d.). More precisely, although for low dimensionalities the covariance matrices of the underlying densities were accurately captured by deterministically evolved particles, for increasing system dimension approximations of covariance matrices became progressively worse. Yet, even for systems of dimension $D = 5$, covariance matrices computed from deterministically simulated solutions of $N = 500$ particles were at the same order of magnitude as accurate as covariances delivered by stochastic particle simulations of size $N = 6500$.

10.5. Second order Langevin systems

To demonstrate the performance of our framework for simulating solutions of the FPEs for second order Langevin systems as described in Section 9, we incorporated our method in a symplectic Verlet integrator (Eq.(A10- A12)) simulating the second order dynamics captured by Eq. (54) for a linear $f(x) = -4\,x$ and a nonlinear, $f(x) = -4\,x^3 + 4\,x$, drift function (Eq. (A10)), and compared the results with stochastic simulations integrated by a semi-symplectic framework [37]. In agreement with previous results, cumulant trajectories evolved smoother in time for deterministic particle simulations when compared to their stochastic counterparts (Fig. 6(a.) and Fig. 7(c.)). Stationary densities closely matched analytically derived ones (see Eq. (A7)) (purple contour lines in Fig. 6(b.) and Fig. 7(b.)), while transient

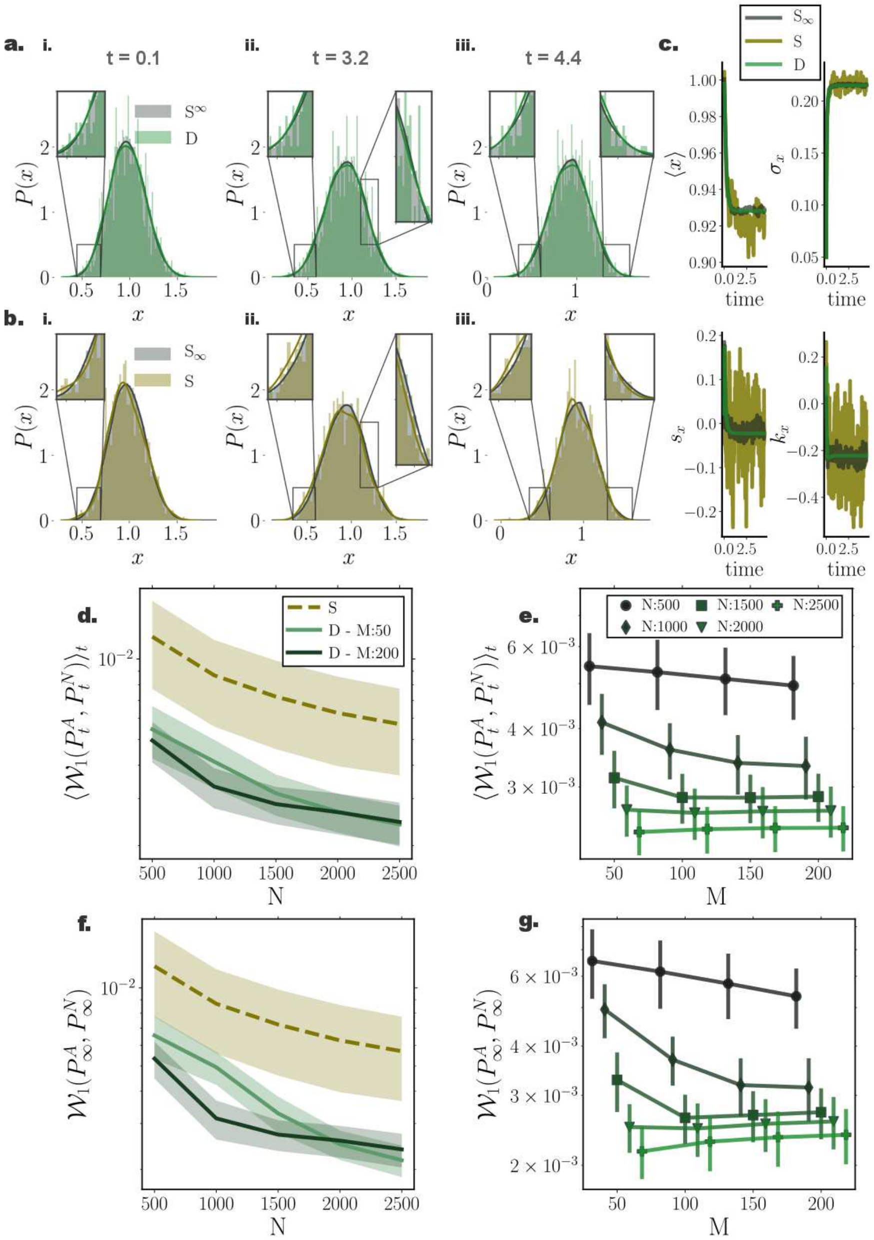

Figure 4. Accuracy of Fokker–Planck solutions for a nonlinear system perturbed with state dependent noise. (a.) Instances of $N = 1000$ particle distributions resulting from deterministic (green) and **(b.)** stochastic (brown) simulations against stochastic particle distributions comprising $N^{\infty} = 35000$ particles (grey) for **(i.)** $t = 0.1$, **(ii.)** $t = 3.2$, and **(iii.)** $t = 4.4$. Insets provide a closer view of details of distribution for visual clarity. Distributions resulting from deterministic particle simulations closer agree with underlying distribution for all three instances. **(c.)** Temporal evolution of first four cumulants for the three particle systems (grey: S_{∞} - stochastic with $N^{\infty} = 35000$ particles, brown: S - stochastic with $N = 1000$ particles, and green: D - deterministic with $N = 1000$ particles). Deterministically evolved distributions result in smooth cumulant trajectories. **(d., e.)** Temporal average and **(f., g.)** stationary 1-Wasserstein distance between distributions mediated through stochastic simulations of $N^{\infty} = 35000$, and through deterministic (green) and stochastic (brown) simulations of N particles against particle number N. Shaded regions and error bars denote one standard deviation among 20 independent repetitions. Different green hues designate different inducing point number M employed in the gradient–log–density estimation. (Further parameter values: regularisation constant $\lambda = 0.001$, Euler integration time step $dt = 10^{-3}$, and RBF kernel length scale $l = 0.25$. Inducing points were arranged on a regular grid spanning the instantaneous state space volume captured by the state vector.)

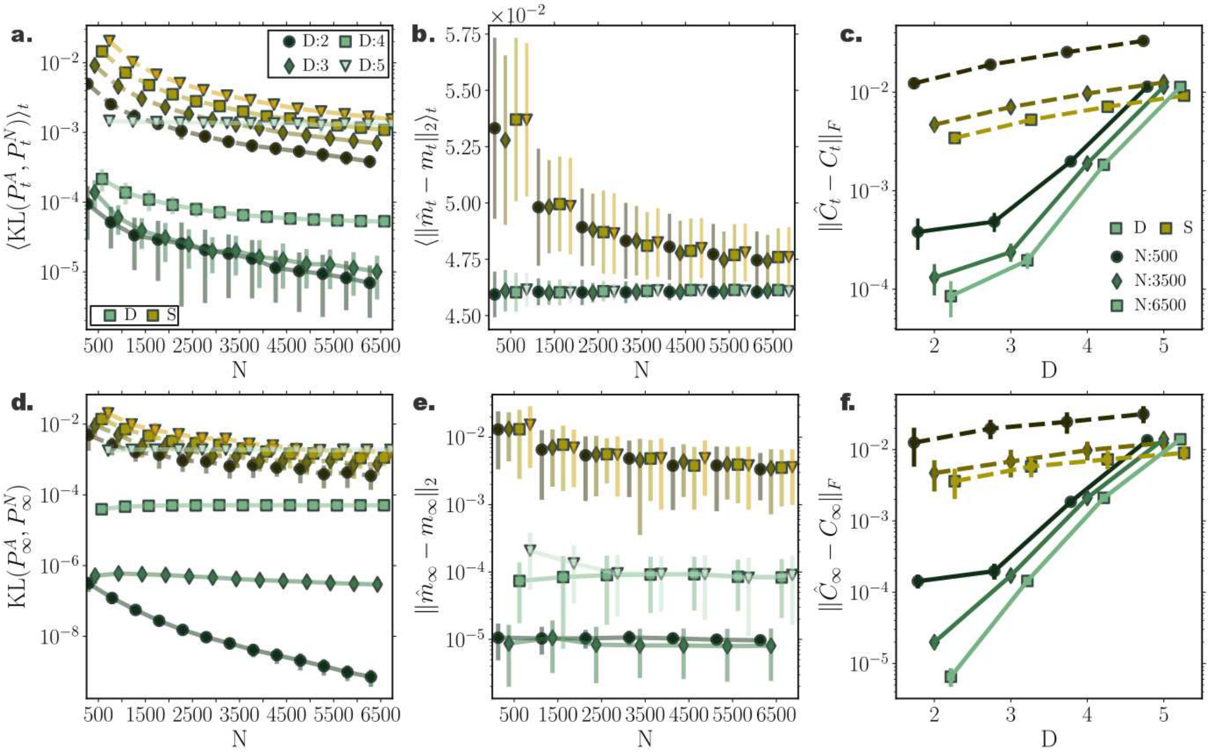

Figure 5. Accuracy of Fokker-Planck solutions for multi-dimensional Ornstein–Uhlenbeck processes. Comparison of deterministic particle Fokker–Planck solutions with stochastic particle systems and analytic solutions for multi-dimensional Ornstein–Uhlenbeck process of D={2,3,4,5} dimensions. **(a.)** Time averaged and **(d.)** stationary Kullback–Leibler (KL) divergence between simulated particle solutions (black/green: deterministic, brown/yellow: stochastic) and analytic solutions for different dimensions. Deterministic particle simulations outperform stochastic particle solutions even for increasing system dimensionality. **(b.)** Time averaged and **(e.)** stationary error between analytic, m_t, and sample mean, $\hat{m}_t$, for increasing particle number. **(c.)** Time averaged and **(f.)** stationary discrepancy between simulated, $\hat{C}_t$, and analytic covariances, C_t, as captured by the Frobenius norm of the relevant covariance matrices difference. The accuracy of the estimated covariance decreases for increasing dimensionality. (Further parameter values: number of inducing points $M = 100$, regularisation constant $\lambda = 0.001$, Euler integration time step $dt = 10^{-3}$, and adaptive RBF kernel length scale l calculated at every time step as two times the standard deviation of the state vector. Inducing point locations were selected randomly at each time step from a uniform distribution spanning the state space volume covered by the state vector.)

densities captured the fine details of simulated stochastic particle densities comprising $N^{\infty} = 20000$ (Fig. 7(a.)).

Furthermore, the symplectic integration contributed to the preservation of energy levels for each particle, after the system reached equilibrium (Fig. 6(e.) and Fig. 7(f.)), which was also evident when observing individual particle trajectories in the state space (Fig. 6(c., d.) and Fig. 7(d., e.)).

As already conveyed in Section 9, the velocity term and the gradient–log–density term canceled out in the long time limit (Fig. 6(f.) and Fig. 7(g.)) for each particle individually, while the average kinetic energy in equilibrium exactly resorted to the value dictated by the fluctuation–dissipation relation and the equipartition of energy property, i.e. $\langle \mathcal{K}^{(i)} \rangle_N = \frac{\sigma^2}{2\gamma}$ (Fig. 6(g.) and Fig. 7(h.)).

10.6. Nonconservative chaotic system with additive noise (Lorenz63)

As a final assessment of our framework for simulating accurate solutions of Fokker–Planck equations, we employed a Lorenz63 model with parameters rendering the dynamics chaotic, perturbed by moderate additive Gaussian noise (Eq. (A13)). By comparing stochastic simulations of $N^{\infty} = 150000$ particles and deterministic and stochastic simulations of $N = 4000$ particles (Fig. 8), we observed that the deterministic framework captured more precisely finer details of the underlying distribution (Fig. 8(a.)), represented here by the N^{∞} stochastic simulation. While both stochastic and deterministic simulations capture the overall butterfly profile of the Lorenz attractor, the deterministic system delivered indeed a closer match to the underlying distribution.

Similar to the previously examined models, cumulant trajectories computed from deterministically evolved particles show closer agreement with those computed from the N^{∞} stochastic system, compared to the stochastic system comprising N particles (Fig. 8(b.)). In particular, cumulants for the x and y states exhibited high temporal fluctuations when computed from stochastically evolved distributions, while our framework conveyed more accurate cumulant trajectories, closer to those delivered by the N^{∞} stochastic system.

11. Discussion and Outlook

We presented a particle method for simulating solutions of FPEs governing the temporal evolution of the probability density for stochastic dynamical systems of the diffusion type. By reframing the FPE in a Liouville form, we obtained an effective dynamics in terms of independent deterministic particle trajectories. Unfortunately, this formulation requires the knowledge of the gradient of the logarithm of the instantaneous probability density of the system state, which is the quantity we try to compute. We circumvented this complication by introducing statistical estimators for the gradient–log–density based on a variational formulation. To combine high flexibility of estimators with computational efficiency, we employed kernel based estimation together with an additional sparse approximation. For the case of equilibrium systems, we related our framework to Stein Variational Gradient Descent, a particle based dynamics to approximate the stationary density, and to a geometric formulation of Fokker–Planck dynamics. We further discussed extensions of our method to settings with multiplicative noise and to second order Langevin dynamics.

To demonstrate the performance of our framework, we provided detailed tests and comparisons with stochastic simulations and analytic solutions (when possible). We demonstrated the accuracy of our method on conservative and non-conservative model systems with different dimensionalities. In particular, we found, that our framework outperforms stochastic simulations both in linear and nonlinear settings, by delivering more accurate densities for small particle number when the dimensionality is small enough. For increasing particle number, the accuracy of both approaches converges. Yet, our deterministic framework delivered *consistently* results with smaller variance among individual

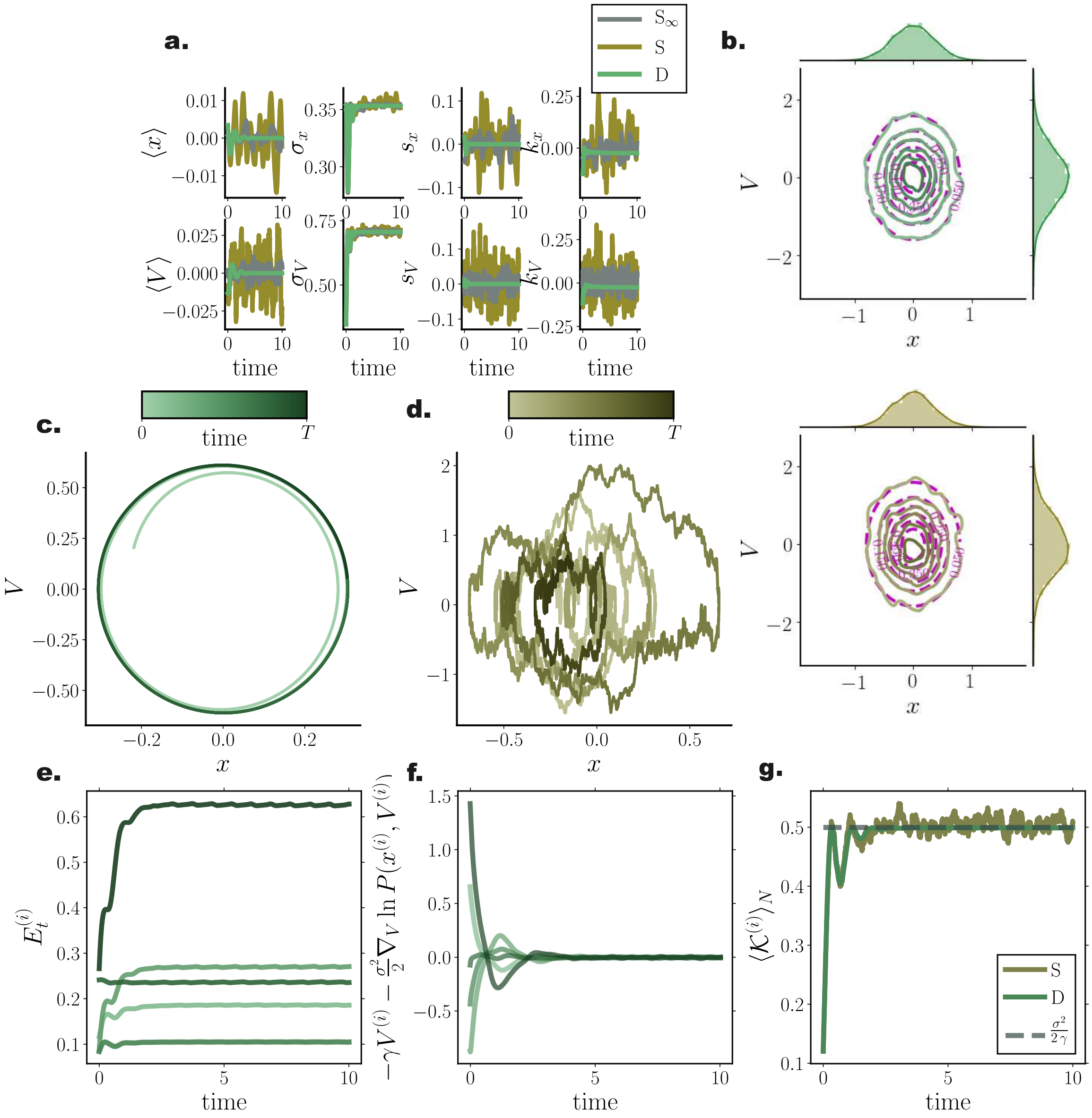

Figure 6. Energy preservation for second order Langevin dynamics in a quadratic potential. Comparison of deterministic particle Fokker–Planck solutions with stochastic particle systems for a harmonic oscillator **(a.)** First four cumulant temporal evolution for deterministic (green) and stochastic (brown) system. **(b.)** Stationary joint and marginal distributions for deterministic (green) and stochastic (brown) systems. Purple lines denote analytically derived stationary distributions. **(c., d.)** State space trajectory of a single particle for deterministic (green) and stochastic (brown) system. Color gradients denote time. **(e.)** Temporal evolution of individual particle energy $E_t^{(i)}$ for deterministic system for 5 particles. **(f.)** Difference between velocity and gradient–log–density term for individual particles. After the system reaches stationary state the particle velocity and GLD term cancel out. **(g.)** Ensemble average kinetic energy through time resorts to $\frac{\sigma^2}{2\gamma}$ (grey dashed line) after equilibrium is reached. (Further parameter values: regularisation constant $\lambda = 0.001$, integration time step $dt = 2 \cdot 10^{-3}$, and adaptive RBF kernel length scale l calculated at every time step as two times the standard deviation of the state vector. Number of inducing points $M = 300$. Inducing point locations were selected randomly at each time step from a uniform distribution spanning the state space volume covered by the state vector.)

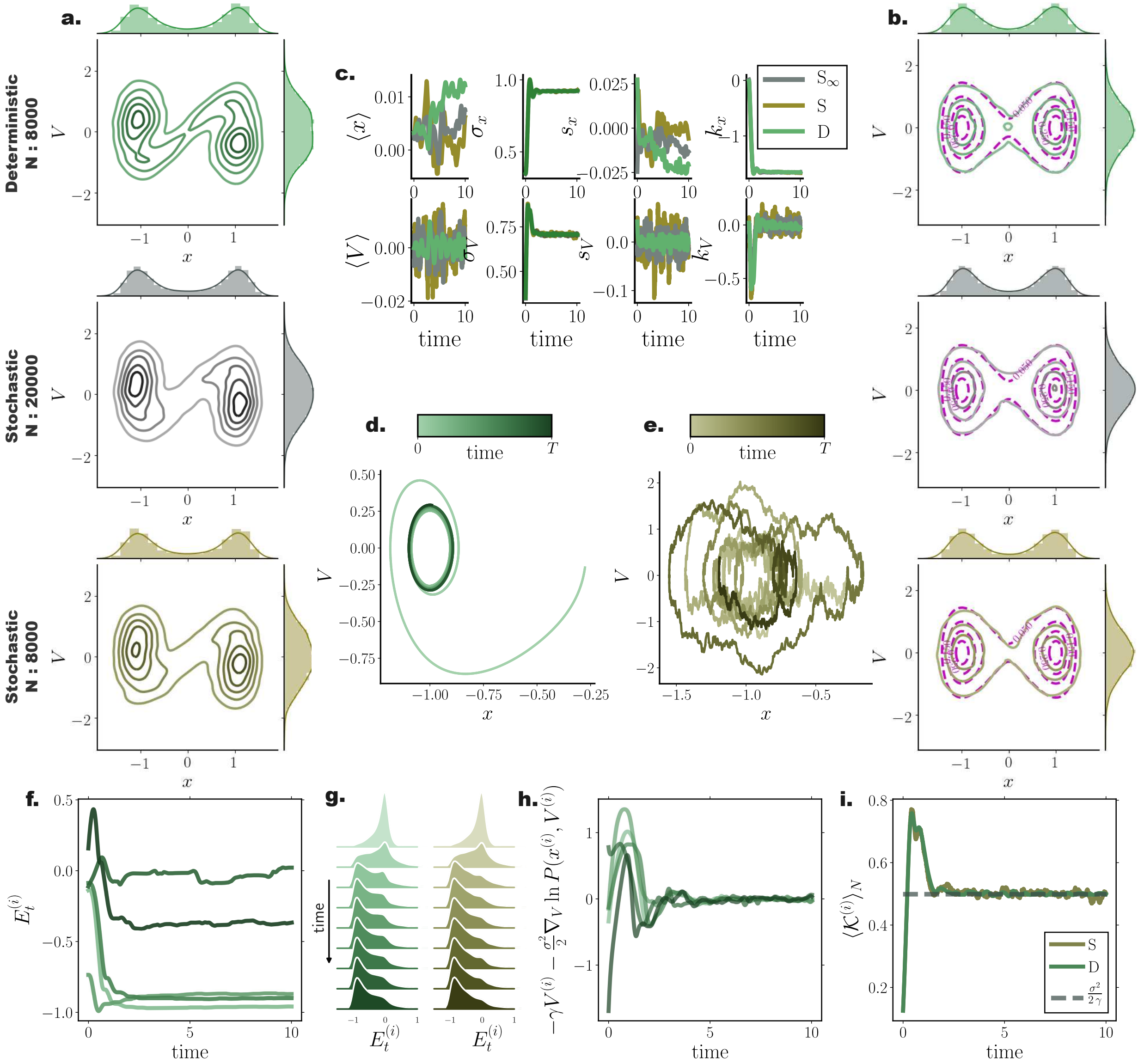

Figure 7. Energy preservation for second order Langevin dynamics in a double well potential. Comparison of deterministic particle Fokker–Planck solutions with stochastic particle systems for a bistable process. **(a., b.)** Joint and marginal distributions of system states mediated by $N = 8000$ particles evolved with our framework (green) and with direct stochastic simulations comprising $N^\infty = 20000$ (grey) and $N = 8000$ (brown) particles at **(a.)** $t = 0.6$, and **(b.)** $t = 10$. Purple lines denote the analytically derived stationary density. **(c.)** First four cumulant temporal evolution for deterministic (green) and stochastic (brown) system. **(d.)** State space trajectory of a single particle for deterministic and **(e.)** stochastic system. Color gradients denote time. **(f.)** Temporal evolution of individual particle energy $E_t^{(i)}$ for deterministic system for 5 particles. **(g.)** Temporal evolution of distribution of particle energies $E_t^{(i)}$ for deterministic (green) and stochastic (brown) system. **(h.)** Difference between velocity and gradient log density term for individual particles. **(i.)** Ensemble average kinetic energy through time resorts to $\frac{\sigma^2}{2\gamma}$ (grey dashed line) after equilibrium is reached. (Further parameter values: regularisation constant $\lambda = 0.001$, integration time step $dt = 2 \cdot 10^{-3}$ and adaptive RBF kernel length scale l calculated at every time step as two times the standard deviation of the state vector. Number of inducing points $M = 300$. Inducing point locations were selected randomly at each time step from a uniform distribution spanning the state space volume covered by the state vector.)

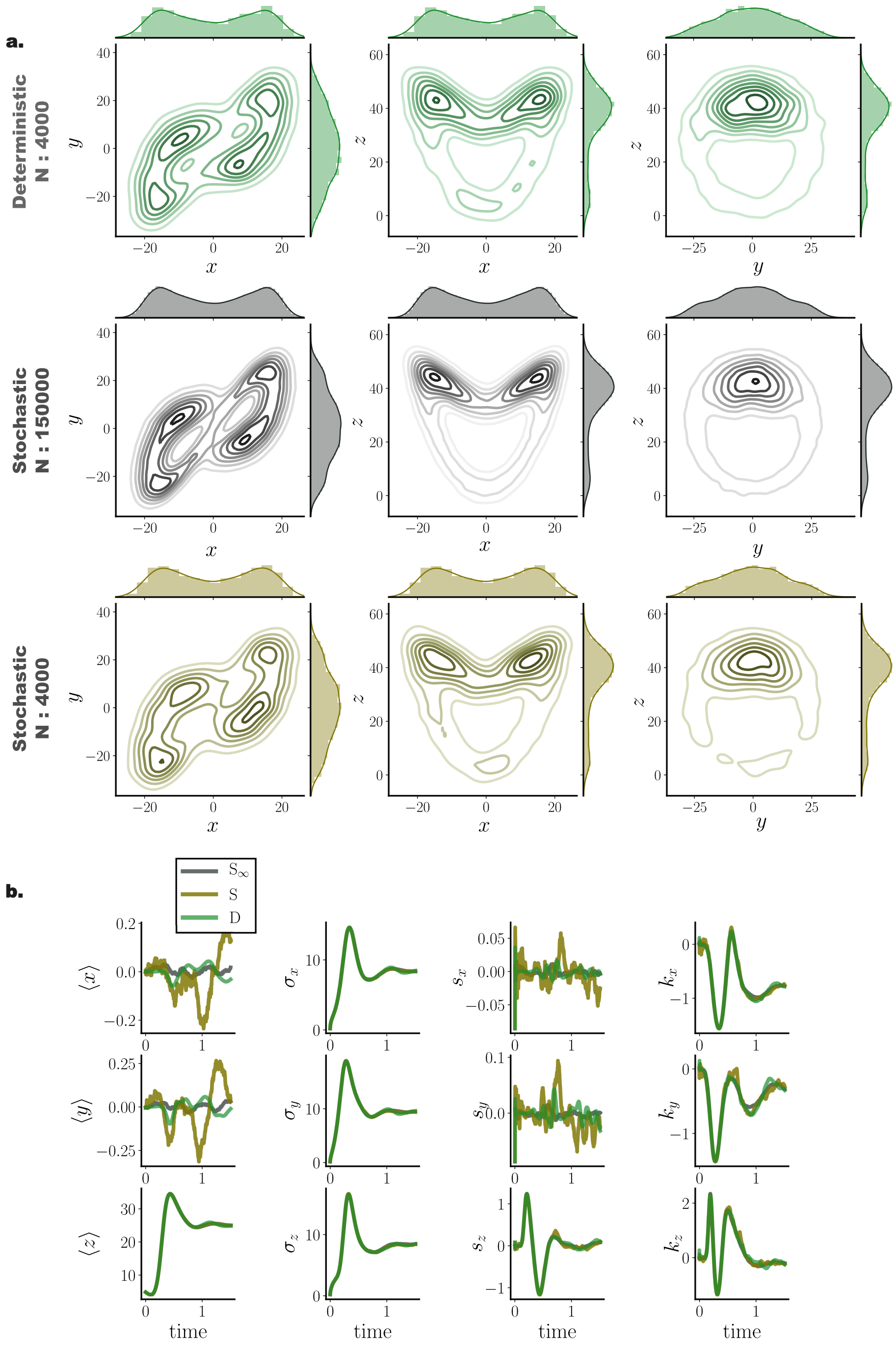

Figure 8. Deterministic (green) and stochastic (brown) Fokker–Planck particle solutions for a three dimensional Lorenz63 system in the chaotic regime perturbed by additive Gaussian noise. (a.) Joint and marginal distributions of system states mediated by $N = 4000$ particles evolved with our framework (green) and with direct stochastic simulations comprising $N = 150000$ (grey) and $N = 4000$ (brown) particles at $t = 0.4$. **(b.)** Cumulant trajectories for the three particle systems. Cumulants derived from deterministic particle simulations (green) closer match cumulant evolution of the underlying distribution (grey) compared to stochastic simulations (brown). (Further parameter values: regularisation constant $\lambda = 0.001$, Euler integration time step $dt = 10^{-3}$, adaptive RBF kernel length scale l calculated at every time step as two times the standard deviation of the state vector. Number of inducing points: $M = 200$. Inducing point locations were selected randomly at each time step from a uniform distribution spanning the state space volume covered by the state vector.)

repetitions. Furthermore, we showed that our method, even for small particle numbers, exhibits low order cumulant trajectories with significantly less temporal fluctuations when compared against to stochastic simulations of the same particle number.

We envisage several ways to improve and extend our method. There is room for improvement by optimising hyper parameters of our algorithm such as inducing point position and kernel length scale. Current grid based and uniform random selection of inducing point position may contribute to the deterioration of solution accuracy in higher dimensions. Other methods, such as subsampling or clustering of particle positions may lead to further improvements. On the other hand, a hyper parameter update may not be at all necessary at each time step in certain settings, such that a further speedup of our algorithm could be achieved.

The implementation of our method depends on the function class chosen to represent the estimator. In this paper we have focused on linear representations, leading to simple closed form expressions. It would be interesting to see if other, nonlinear parametric models, such as neural networks, (see e.g. [38]) could be employed to represent estimators. While, in this setting, there would be no closed form solutions, the small changes in estimates between successive time steps, suggest that only a few updates of numerical optimisation may be necessary at each step. Moreover, the ability of neural networks to automatically learn relevant features from data might help to improve performance for higher dimensional problems when particle motion is typically restricted on lower dimensional submanifolds.

From a theoretical point of view, rigorous results on the accuracy of the particle approximation would be important. These would depend on the speed of convergence of estimators towards exact gradients of log–densities. However, to obtain such results may not be easy. While rates of convergence for kernel based estimators have been studied in the literature, the methods for proofs usually rely on the independence of samples and would not necessarily apply to the case of interacting particles.

We have so far addressed only the forward simulation of FPEs. However, preliminary results indicate that related techniques may be applied to particle based simulations for smoothing (forward–backward) and related control problems for diffusion processes [39]. Such problems involve computations of an effective, controlled drift function in terms of gradient–log–densities. We defer further details and discussions on subsequent publications on the topic.

Taken together, the main advantage of our framework is its minimal requirement in simulated particle trajectories for attaining reliable Fokker–Planck solutions with smoothly evolving transient statistics. Moreover, our proposed method is nearly effortless to set up when compared to classical grid based FPE solvers, while it delivers more reliable results than direct stochastic simulations.

Author Contributions: Conceptualization, S.R. and M.O.; methodology, D.M. and M.O.; software, D.M.; validation, D.M. and M.O.; formal analysis, D.M. and M.O.; investigation, D.M.; resources, M.O.; data curation, D.M.; writing–original draft preparation, D.M. and M.O.; writing–review and editing, D.M., S.R and M.O.; visualization, D.M.; supervision, M.O.; project administration, M.O.; funding acquisition, S.R. and M.O.

Funding: This research has been partially funded by Deutsche Forschungsgemeinschaft (DFG) - SFB1294/1 - 318763901.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Schadschneider, A.; Chowdhury, D.; Nishinari, K. *Stochastic transport in complex systems: From molecules to vehicles*; Elsevier, 2010.

2. Kumar, P.; Narayanan, S. Solution of Fokker-Planck equation by finite element and finite difference methods for nonlinear systems. *Sadhana* **2006**, *31*, 445–461.
3. Risken, H. Fokker-Planck equation. In *The Fokker-Planck Equation*; Springer-Verlag, 1996; pp. 63–95.
4. Brics, M.; Kaupuzs, J.; Mahnke, R. How to solve Fokker-Planck equation treating mixed eigenvalue spectrum? *arXiv preprint arXiv:1303.5211* **2013**.
5. Chang, J.; Cooper, G. A practical difference scheme for Fokker-Planck equations. *Journal of Computational Physics* **1970**, *6*, 1–16.
6. Pichler, L.; Masud, A.; Bergman, L.A. Numerical solution of the Fokker–Planck equation by finite difference and finite element methods—a comparative study. In *Computational Methods in Stochastic Dynamics*; Springer-Verlag, 2013; pp. 69–85.
7. Leimkuhler, B.; Reich, S. *Simulating Hamiltonian dynamics*; Vol. 14, Cambridge University Press, 2004.
8. Chen, N.; Majda, A.J. Efficient statistically accurate algorithms for the Fokker–Planck equation in large dimensions. *Journal of Computational Physics* **2018**, *354*, 242–268.
9. Lin, Y.; Cai, G. Probabilistic structural dynamics: Advanced theory and applications. *New York: McGraw-Hill* **1995**.
10. Roberts, J.B.; Spanos, P.D. *Random vibration and statistical linearization*; Courier Corporation, 2003.
11. Proppe, C.; Pradlwarter, H.; Schuëller, G. Equivalent linearization and Monte Carlo simulation in stochastic dynamics. *Probabilistic Engineering Mechanics* **2003**, *18*, 1–15.
12. Grigoriu, M. *Stochastic calculus: Applications in science and engineering*; Springer Science & Business Media, 2013.
13. Øksendal, B. *Stochastic differential equations*; Springer-Verlag, 2003.
14. Kroese, D.P.; Taimre, T.; Botev, Z.I. *Handbook of Monte Carlo methods*; Vol. 706, John Wiley & Sons, 2013.
15. Carrillo, J.; Craig, K.; Patacchini, F. A blob method for diffusion. *Calculus of Variations and Partial Differential Equations* **2019**, *58*, 1–53.
16. Pathiraja, S.; Reich, S. Discrete gradients for computational Bayesian inference. *J. Comp. Dynamics* **2019**, *6*, 236–251.
17. Reich, S.; Weissmann, S. Fokker-Planck particle systems for Bayesian inference: Computational approaches. *arXiv preprint arXiv:1911.10832* **2019**.
18. Liu, Q.; Lee, J.; Jordan, M. A kernelized Stein discrepancy for goodness-of-fit tests. International conference on machine learning, 2016, pp. 276–284.
19. Taghvaei, A.; Mehta, P.G. Accelerated flow for probability distributions. *arXiv preprint arXiv:1901.03317* **2019**.
20. Velasco, R.M.; Scherer García-Colín, L.; Uribe, F.J. Entropy production: Its role in non-equilibrium thermodynamics. *Entropy* **2011**, *13*, 82–116.
21. Hyvärinen, A. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research* **2005**, *6*, 695–709.
22. Li, Y.; Turner, R.E. Gradient estimators for implicit models. *arXiv preprint arXiv:1705.07107* **2017**.
23. Shi, J.; Sun, S.; Zhu, J. A spectral approach to gradient estimation for implicit distributions. *arXiv preprint arXiv:1806.02925* **2018**.
24. Tomé, T.; De Oliveira, M.J. *Stochastic dynamics and irreversibility*; Springer, 2015.
25. Shawe-Taylor, J.; Cristianini, N.; et al.. *Kernel methods for pattern analysis*; Cambridge University Press, 2004.
26. Scholkopf, B.; Smola, A.J. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*; MIT Press, 2001.
27. Sutherland, D.J.; Strathmann, H.; Arbel, M.; Gretton, A. Efficient and principled score estimation with Nyström kernel exponentia *arXiv preprint arXiv:1705.08360* **2017**.
28. Rasmussen, C.E. Gaussian processes in machine learning. Summer School on Machine Learning. Springer-Verlag, 2003, pp. 63–71.
29. Liu, Q.; Wang, D. Stein variational gradient descent: A general purpose Bayesian inference algorithm. Advances in neural information processing systems, 2016, pp. 2378–2386.

30. Liu, Q. Stein variational gradient descent as gradient flow. Advances in neural information processing systems, 2017, pp. 3115–3123.
31. Garbuno-Inigo, A.; Nüsken, N.; Reich, S. Affine invariant interacting Langevin dynamics for Bayesian inference. Technical Report arXiv:1912.02859, *SIAM J. Dyn. Syst.* in press, 2019.
32. Otto, F. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations* **2001**, *26*, 101–174.
33. Villani, C. *Optimal transport: Old and new*; Springer Science & Business Media, 2008.
34. Frogner, C.; Poggio, T. Approximate inference with Wasserstein gradient flows. *arXiv preprint arXiv:1806.04542* **2018**.
35. Caluya, K.; Halder, A. Gradient flow algorithms for density propagation in stochastic systems. *IEEE Transactions on Automatic Control* **2019**. doi:doi: 10.1109/TAC.2019.2951348.
36. Batz, P.; Ruttor, A.; Opper, M. Variational estimation of the drift for stochastic differential equations from the empirical density. *Journal of Statistical Mechanics: Theory and Experiment* **2016**, *2016*, 083404.
37. Milstein, G.; Tretyakov, M. Computing ergodic limits for Langevin equations. *Physica D: Nonlinear Phenomena* **2007**, *229*, 81–95.
38. Saremi, S.; Mehrjou, A.; Schölkopf, B.; Hyvärinen, A. Deep energy estimator networks. *arXiv preprint arXiv:1805.08306* **2018**.
39. Reich, S.; Cotter, C. *Probabilistic forecasting and Bayesian data assimilation*; Cambridge University Press, 2015.
40. Lorenz, E.N. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences* **1963**, *20*, 130–141.

Appendix A. Simulated systems

Appendix A.1. Two dimensional Ornstein-Uhlenbeck process

For comparing Fokker-Planck solutions computed with our approach with solutions derived from stochastic simulations, we considered the two dimensional Ornstein-Uhlenbeck process captured by the following equations

$$dX_t = (-4X_t + Y_t)\,dt + \sigma\,dB_1 \tag{A1}$$
$$dY_t = (-4Y_t + X_t)\,dt + \sigma\,dB_2, \tag{A2}$$

where the related potential is $U(x,y) = 2x^2 - x\,y + 2y^2$. Simulation time was set to $T = 3$ with Euler–Maruyama integration step $dt = 10^{-3}$. For estimating the instantaneous gradient log density we employed $M = \{50, 100, 150, 200\}$ inducing points, randomly selected at every time point from a uniform distribution spanning the state space volume covered by the particles at the current time point.

Appendix A.2. Bistable nonlinear system

For testing our framework on nonlinear settings, we simulated

$$dX_t = \left(-4{X_t}^3 + 4X_t\right) dt + D(X_t)^{\frac{1}{2}} dB_t, \tag{A3}$$

with $D(x) = 1$ for evaluating solutions with additive Gaussian noise, and with $D(x) = sin^2(x)$ for multiplicative noise FP solutions. The associated potential reads $U(x) = x^4 - 2x^2$.

Appendix A.3. Multi-dimensional Ornstein-Uhlenbeck processes

For quantifying the scaling of our method for increasing system dimension, we simulated systems of dimensionality $D = \{2, 3, 4, 5\}$ according to the following equation

$$dX_{(i)_t} = \left(-4X_{(i)_t} + \sum_{j=1,i\neq j}^{d} \frac{1}{2} X_{(j)_t} \right) dt + dB_{(i)}, \tag{A4}$$

for $d \in D$. Simulation time was determined by the time required for analytic mean m_t to converge to its stationary solution within $\tilde{\epsilon}$ precision $\tilde{\epsilon} = 10^{-5}$, while the integration step was set to $dt = 10^{-3}$.

Appendix A.4. Second order Langevin dynamics

For demonstrating the energy preservation properties of our method for second order Langevin dynamics, we incorporated our framework into a Verlet symplectic integration scheme (Eq. (A10)), and compared the results with stochastic simulations integrated according to a semi-symplectic scheme [37].

We consider a system with dynamics for positions X and velocities V captured by

$$dX = V dt \tag{A5}$$

$$dV = (-\gamma V + f(X))\, dt + \sigma dB_t\ , \tag{A6}$$

where the velocity change (acceleration) is the sum of a deterministic drift f, a velocity dependent damping $-\gamma V$, and a stochastic noise term σdB_t.

In conservative settings the drift comes as the gradient of a potential $f(x) = -\nabla U(x)$. Here we used a quadratic (harmonic) potential $U(x) = 2x^2$ and a double-well potential $U(x) = x^4 - 2x^2$.

In equilibrium, the Fokker–Planck solution is the Maxwell–Boltzmann distribution, i.e.

$$p_\infty(X, V) = \frac{1}{Z} e^{-\beta H(X,V)} = \frac{1}{Z} e^{-\beta\left(\frac{\|V\|^2}{2} + U(X)\right)}, \tag{A7}$$

with partition function $Z = \int e^{-\beta\left(\frac{\|V\|^2}{2} + U(X)\right)} dx dv$.

We may compute the energy of each particle at each time point as the sum of its kinetic and potential energies

$$E_t^{(i)} = \frac{1}{2} V^{(i)^2} + U(X^{(i)}). \tag{A8}$$

Here the superscripts denote individual particles. After the system has reached equilibrium, energy levels per particle are expected to remain constant.

From the equipartition of energy and the fluctuation–dissipation relation, in the long time limit the average kinetic energy of the system is expected to resort to

$$\lim_{t\to\infty} \langle \mathcal{K} \rangle = \lim_{t\to\infty} \frac{1}{2} \langle V^2 \rangle = \frac{\sigma^2}{2\,\gamma}. \tag{A9}$$

Symplectic integration [7] of Eq. (54) follows the equations

$$V_{n+\frac{1}{2}} = V_n + \frac{dt}{2} \left(-\gamma V_n + f(X_n) - \frac{\sigma^2}{2} \nabla_v \ln p_t(X_n, V_n) \right) \tag{A10}$$

$$X_{n+1} = X_n + dt\, V_{n+\frac{1}{2}} \tag{A11}$$

$$V_{n+1} = V_{n+\frac{1}{2}} + \frac{dt}{2} \left(-\gamma V_{n+\frac{1}{2}} + f(X_{n+1}) - \frac{\sigma^2}{2} \nabla_v \ln p_t(X_{n+1}, V_{n+\frac{1}{2}}) \right), \tag{A12}$$

where n denotes a single integration step.

Appendix A.5. Lorenz63

For simulating trajectories of the noisy Lorenz63 system we employed the following equations

$$dx_t = \sigma(y - x)dt + \sigma dW_x \tag{A13}$$

$$dy_t = (x(\rho - z) - y)\,dt + \sigma dB_y \tag{A14}$$

$$dz_t = (x\,y - \beta z)\,dt + \sigma dB_z, \tag{A15}$$

with parameters $\sigma = 10$, $\rho = 28$, and $\beta = \frac{8}{3}$, that render the deterministic dynamics chaotic [40], employing moderate additive Gaussian noise.

Appendix B. Computing central moment trajectories for linear processes

For a linear process

$$dX_t = A\,X_t dt + \sigma dB, \tag{A16}$$

the joint density of the state vector X remains Gaussian for all times when the initial density is Gaussian. The mean vector m and covariance matrix C may be computed by solving the ODE system

$$\frac{dm}{dt} = Am \tag{A17}$$

$$\frac{dC}{dt} = AC + CA^\top + \sigma^2 I. \tag{A18}$$

Appendix C. Kullback–Leibler divergence for Gaussian distributions

We calculated the KL divergence between the theoretical and simulated distributions with

$$KL\,(P_1||P_2) = \frac{1}{2}\left(\log(|\Sigma_2|/|\Sigma_1|) - d + Tr(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T\Sigma_2^{-1}(\mu_2 - \mu_1)\right), \tag{A19}$$

where $P_x \sim \mathcal{N}\,(\mu_x, \Sigma_x)$.

Appendix D. Wasserstein distance

We employed the 1-Wasserstein distance [33] as a distance metric for comparing pairs of empirical distributions.

For two distributions P and Q, we denote with $\mathcal{J}(P,Q)$ all joint distributions J for a pair of random variables (X,Y) with marginals P and Q. Then the *Wasserstein distance* between these distributions reads

$$\mathcal{W}_p(P,Q) = \left(\inf_{J\in\mathcal{J}(P,Q)} \int \|x - y\|^p dJ(x,y)\right)^{\frac{1}{p}}, \tag{A20}$$

where for the 1-Wasserstein distances (used in the present manuscript) $p = 1$.

Interestingly, the Wasserstein distance between two one dimensional distributions P and Q obtains a closed form solution

$$\mathcal{W}_p(P,Q) = \left(\int_0^1 \|F_P^{-1}(\tau) - F_Q^{-1}(\tau)\|^p d\tau\right)^{1/p}, \tag{A21}$$

with F_P and F_Q indicating the cumulative distribution functions of P and Q.

Moreover, for one dimensional empirical distributions P and Q with samples of same size $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, the Wasserstein distance simplifies into computation of differences of order statistics

$$\mathcal{W}_p(P, Q) = \left(\sum_{i=1}^{n} \| X_{(i)} - Y_{(i)} \|^p \right)^{\frac{1}{p}}, \tag{A22}$$

where $X_{(i)}$ and $Y_{(i)}$ indicates the i-th order statistic of the sample $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, i.e. $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ and $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$.

Appendix E. Influence of hyperparameter values on the performance of the Gradient–Log–Density estimator

To determine the influence of the hyperparameter values on the performance of the gradient–log–density estimator, we systematically evaluated the approximation error of our estimator for $N = 1000$ samples of a one dimensional log–normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 0.5$ for 20 independent realisations.

We quantified the approximation error as the average error between the analytically calculated and predicted gradient-log-density on each sample, i.e.

$$\text{Approximation error} = \frac{1}{N} \sum_{i=1}^{N} \| \nabla \ln p(x_i) - (\nabla \widehat{\ln p(x_i)}) \|, \tag{A23}$$

where the analytically calculated gradient-log-density was determined as $\nabla \ln p(x) = \frac{\mu - \sigma^2 - \ln(x)}{\sigma^2 x}$.

By systematically varying the regularisation parameter λ, the kernel length scale l, and the inducing point number M we observed the following:

- The hyperparameter that strongly influences the approximation accuracy is the kernel length scale l (Fig. A1).
- Underestimation of kernel length scale l has stronger impact on approximation accuracy, than overestimation (Fig. A1).
- For increasing regularisation parameter value λ, underestimation of l has less impact on the approximation accuracy (Fig. A1 and Fig. A2).
- For overestimation of the kernel length scale l, regularisation parameter λ and inducing point number M have nearly no effect on the resulting approximation error (Fig. A1).
- For underestimation of kernel length scale l, increasing the number of inducing points M in the estimator results in larger approximation errors (Fig. A2 (upper left)).

Appendix F. Required number of particles for accurate Fokker–Planck solutions

To compare the computational demands of the deterministic and stochastic particle systems we determined the required particle number each system needed to attain a specified accuracy to ground truth transient solutions. In particular, for a two dimensional Ornstein–Uhlenbeck process we identified the minimal number of particles N^*_{KL} both systems required to achieve a certain time averaged Kullback–Leibler distance to ground truth transient solutions, $\langle \mathrm{KL}\left(P_t^A, P_t^N\right)\rangle_t$. As already indicated in the previous sections, the stochastic system required considerably larger particle number to achieve the same time averaged KL distances to ground truth when compared to our proposed framework. In fact, for the entire range of examined KL distances, our method consistently required at least one order of magnitude less particles compared to the its stochastic counterpart.

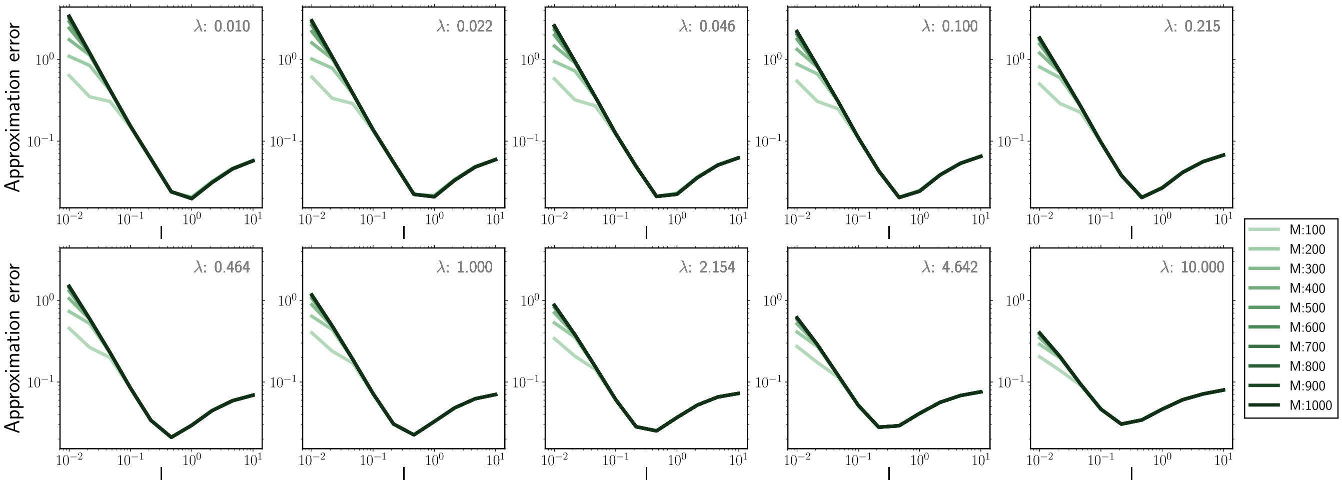

Figure A1. **Approximation error for increasing kernel length scale l for different regularisation parameter values λ and inducing point number M.**

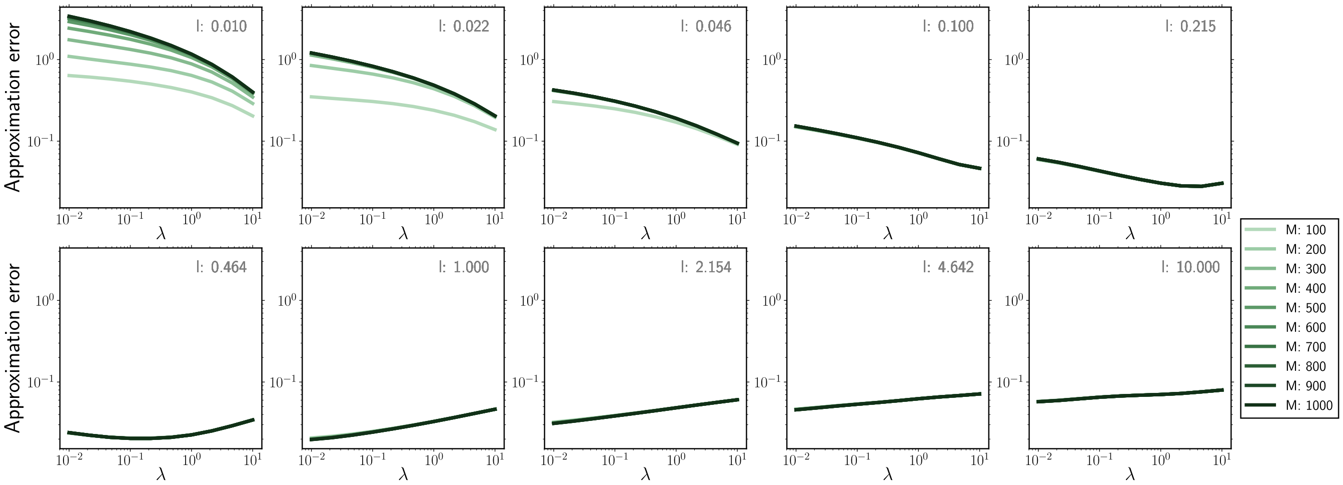

Figure A2. **Approximation error for increasing regularisation parameter value λ for different kernel length scale l and inducing point number M.**

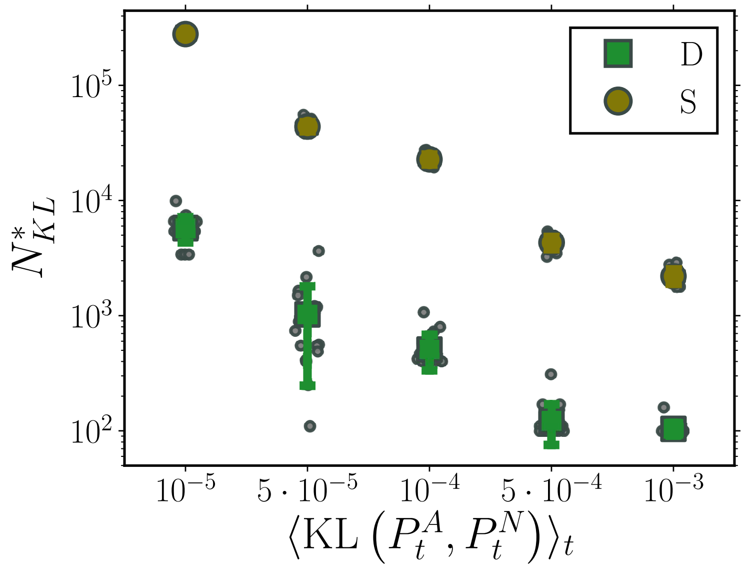

Figure A3. Required particle number, N^*_{KL}, to attain time averaged Kullback–Leibler divergence to ground truth, $\langle \mathrm{KL}\left(P_t^A, P_t^N\right)\rangle_t$, for deterministic (green) and stochastic (brown) particle systems for a two dimensional Ornstein-Uhlenbeck process. Markers indicate mean required particle number, while error bars denote one standard deviation over 20 independent realisations. Grey circles indicate required particle number for each individual realisation. Deterministic particle system consistently required at least one order of magnitude less particles compared to its stochastic counterpart. (Further parameters values: regularisation constant $\lambda = 0.001$, inducing point number $M = 100$, and RBF kernel length scale l estimated at every time point as two times the standard deviation of the state vector. Inducing point locations were selected randomly at each time step from a uniform distribution spanning the state space volume covered by the state vector.)

Appendix G. Algorithm for simulating deterministic particle system

Here we provide the algorithm for simulating deterministic particle trajectories according to our proposed framework. In the comments, we denote the computational complexity of each operation in the gradient–log–density estimation in terms of big-$\mathcal{O}$ notation. Since the inducing point number M employed in the gradient–log–density estimation is considerably smaller than sample number N, i.e. $M \ll N$, the overall computational complexity of a *single* gradient-log-density evaluation amounts to $\mathcal{O}\left(N\,M^2\right)$.

Algorithm 1: Gradient Log Density Estimator

```
Input:  $X$: $N \times D$ state vector
        $Z$: $M \times D$ inducing points vector
        $d$: dimension for gradient
        $l$: RBF Kernel length scale
Output: $G$: $N \times 1$ vector for gradient-log-density at each position $X$ in $d$ dimension
$K^{xz} \longleftarrow K(X, Z; l)$                                   // $N \times M$  $\mathcal{O}(N\,M)$
$K^{zz} \longleftarrow K(Z, Z; l)$                                   // $M \times M$  $\mathcal{O}(M^2)$
$I\_K^{zz} \longleftarrow \left(K^{zz} + 10^{-3}\, I\right)^{-1}$    // $M \times M$  $\mathcal{O}(M^3)$
$grad\_K \longleftarrow \nabla_{X^{(d)}} K(X, Z; l)$                 // $N \times M$  $\mathcal{O}(N\,M)$
$sgrad\_K \longleftarrow \sum_{X_i} grad\_K$                         // $1 \times M$
$G \longleftarrow K^{xz}\ \left(\lambda\, I + I\_K^{zz}\ (K^{xz})^\intercal\ K^{xz} + 10^{-3}\, I\right)^{-1}\ I\_K^{zz}\, sgrad\_K^\intercal$   // $N \times 1$
                                                                     // $\mathcal{O}(N\,M^2) + \mathcal{O}(M^3)$
```

Algorithm 2: Deterministic Particle Simulation

```
Input: $x_0$: $1 \times D$ initial condition
       $s_0$: variance of initial condition
       $N$: particle number
       $M$: inducing point number
       $T$: duration of simulation
       $dt$: integration time step
       $f(\cdot)$: drift function
       $D(\cdot)$: diffusion function
       $l\_0$: $1 \times D$ Kernel length scale or False        // FALSE for adaptive length scale selection
       $random\_M$: boolean variable     // TRUE for selecting inducing points from random uniform
                                         // distribution; FALSE for arranging them on a regular grid
Output: $\{X_t\}_{t=0}^{T/dt}$: $N \times D \times \left\lceil \frac{T}{dt} \right\rceil$ particle trajectories

Initialization: $X_0 \leftarrow$ Draw $N$ samples from Gaussian $\mathcal{N}(x_0, s_0)$
for $t \leftarrow 1$ to $\frac{T}{dt}$ do
    if $random\_M$ is TRUE then
                                                                  // Select inducing points
        for $d \leftarrow 1$ to $D$ do
            $Z^{(d)} \longleftarrow$ Draw $M$ samples from Uniform(min($X_{t-1}^{(d)}$), max($X_{t-1}^{(d)}$))
        end
    else
        for $d \leftarrow 1$ to $D$ do
            $Z^{(d)} \longleftarrow$ CreateRegularGrid(min($X_{t-1}^{(d)}$), max($X_{t-1}^{(d)}$), $M$)
        end
    end
    if $l\_0$ is FALSE then
                                                                  // Set length scale
        $l \longleftarrow 2\,std(X_{t-1})$
    else
        $l \longleftarrow l\_0$
    end
    for $d \leftarrow 1$ to $D$ do
        $G^{(d)} \longleftarrow$ GradientLogDensityEstimation($X_{t-1}$,$Z$,$d$,$l$)
    end
    $X_t \longleftarrow X_{t-1} + \Big(f(X_{t-1}) - \frac{1}{2}D(X_{t-1}) \circ G - \frac{1}{2}\nabla D(X_{t-1})\Big)\,dt$
end
```