COMMUN. MATH. SCI.
Vol. 15, No. 7, pp. 2087–2097

FAST COMMUNICATION

GAUSSIAN APPROXIMATIONS OF SMALL NOISE DIFFUSIONS IN KULLBACK–LEIBLER DIVERGENCE*

DANIEL SANZ-ALONSO† AND ANDREW M. STUART‡

Abstract. We study Gaussian approximations to the distribution of a diffusion. The approximations are easy to compute: they are defined by two simple ordinary differential equations for the mean and the covariance. Time correlations can also be computed via solution of a linear stochastic differential equation. We show, using the Kullback–Leibler divergence, that the approximations are accurate in the small noise regime. An analogous discrete time setting is also studied. The results provide both theoretical support for the use of Gaussian processes in the approximation of diffusions, and methodological guidance in the construction of Gaussian approximations in applications.

Keywords. Gaussian approximations; diffusion processes; small noise; Kullback–Leibler divergence.

AMS subject classifications. 28C20; 60H10; 65L05; 60G15.

1. Introduction

Consider the stochastic differential equation (SDE)

$$dv^\epsilon(t) = f\big(v^\epsilon(t)\big)dt + \sqrt{\epsilon\Sigma}\,dW(t), \quad t\in[0,T], \quad v^\epsilon(0)\sim\mu_0^\epsilon. \tag{1.1}$$

Here, and throughout, the time-horizon $T>0$ is finite and fixed, the drift $f:\mathbb{R}^D\to\mathbb{R}^D$ is nonlinear in the case of interest, Σ is a positive-definite matrix for simplicity assumed to be constant, and dW is a standard D-dimensional Wiener process. Assume that the paths v^ϵ are continuous, and let μ^ϵ be the law of v^ϵ in $C([0,T],\mathbb{R}^D)$. We show that in the small ϵ regime —corresponding to small diffusion coefficient and little uncertainty in the initial condition— μ^ϵ can be accurately approximated by a Gaussian measure ν^ϵ in $C([0,T],\mathbb{R}^D)$, whose mean function and marginal covariances satisfy simple ordinary differential equations (ODEs). Precisely, we show that the Kullback–Leibler divergence $D_{\mathrm{KL}}(\nu^\epsilon\|\mu^\epsilon)$ is of order ϵ. We investigate the effect that numerically approximating these ODEs has on the approximation of μ^ϵ. Finally, we construct —and show the accuracy of— Gaussian approximations to Euler–Maruyama discretizations of equation (1.1).

We aim to provide a rigorous justification of Gaussian process approximations of small noise diffusions. The use of Gaussian processes is now pervasive in applications [15]. However, Gaussian assumptions are often used for algorithmic and mathematical convenience without much theoretical support. An example of this is in the field of data assimilation, where many algorithms invoke Gaussian approximations in order to apply Kalman formulae [8]. The paper [1] suggests a variational approach to computing Gaussian process approximations to posterior measures arising from discretely observed diffusions. The algorithm proposed aims to minimize the Kullback–Leibler divergence

*Received: May 19, 2016; accepted (in revised form): September 30, 3016. Communicated by David Anderson.
DS-A is supported by EPSRC as part of the MASDOC DTC at the University of Warwick with grant No. EP/H023364/1. The work of AMS is supported by DARPA, EPSRC and ONR.

†Mathematics Institute, University of Warwick, Coventry CV4 7AL, United Kingdom (d.sanz-alonso@warwick.ac.uk).
‡Mathematics Institute, University of Warwick, Coventry CV4 7AL, United Kingdom (a.m.stuart@warwick.ac.uk).

between the Gaussian process and the posterior of interest. In [14], the authors study the well-posedness of the abstract problem of finding the best (in the Kullback–Leibler sense) Gaussian approximation ν to a measure μ. They use a nonconstructive calculus of variation approach. In the complementary paper [13], the authors propose an algorithm that allows numerical computation of such best approximation. The paper [11] seeks the best Kullback–Leibler Gaussian approximation to the distribution of conditioned diffusions arising in molecular dynamics. Here our interest is not in finding the best Gaussian approximation, but in constructing Gaussians that can be proved to approximate accurately the distribution of a small noise diffusion. The Gaussian approximations are built using certain linearizations of the drift. Crucially, the time-marginals of our approximations can be found by solving (perhaps numerically) two simple ODEs, which can be advantageous over numerically solving the full Fokker–Planck equation. Our computable Gaussian approximations and error bounds are of potential interest in filtering signals arising from small noise diffusions, for instance within the methodology in [10].

The paper is organized as follows. Subsection 1.1 sets the mathematical framework, and introduces the ODEs that will be used to define the means and covariances of the time-marginals of our approximations. Subsection 1.2 gives some background on the Kullback–Leibler divergence. Subsection 1.3 motivates the choice of the ODEs. Section 2 studies the approximation of the SDE using a Gaussian whose time-marginals mean and covariance satisfy the ODEs given in Subsection 1.1. In Section 3 we investigate the effect that numerically solving the ODEs has on the quality of the resulting Gaussian approximation. Finally, Section 4 studies the approximation of Euler discretizations of equation (1.1) via Gaussians with means and covariances defined through discretization of the ODEs.

Notation. We denote by $|\cdot|$ the standard inner product in $\mathbb{R}^D$. For positive-definite $P \in \mathbb{R}^{D\times D}$ we denote $|\cdot|_P := |P^{-1/2}\cdot|$. Subscripts in path measures will denote time-marginals. Thus, for a measure μ in $C([0,T],\mathbb{R}^D)$ and $0 \le t \le T$, we denote by μ_t the push-forward measure in $\mathbb{R}^D$ via the evaluation map $E_t(h) := h(t)$, $h \in C([0,T],\mathbb{R}^D)$. In the discrete setting, given a path measure $\mu_{0:k}$ in $\mathbb{R}^{D(k+1)}$ and $0 \le j \le k$, μ_j will denote the push-forward measure in $\mathbb{R}^D$ via the evaluation map $E_j : (v_d)_{d=1}^{D(k+1)} \in \mathbb{R}^{D(k+1)} \mapsto (v_{j+d})_{d=1}^{D} \in \mathbb{R}^D$. A subscript d in a function or vector denotes, as usual, component and coordinate. Superscripts highlight parameters of interest in the analysis.

1.1. Mathematical framework. ODEs for mean and covariance. We work under the following assumption (recalling that T is fixed and finite):

ASSUMPTION 1.1. *$f \in C^2(\mathbb{R}^D,\mathbb{R}^D)$ and there exist $c>0$ and $s \in \mathbb{N}$ such that, for $1 \le d \le D$, and every multi-index α with $|\alpha| = 2$, $|\partial^\alpha f_d(u)| \le c(1+|u|^s)$. Moreover, for any $m_0 \in \mathbb{R}^D$, the ODE*

$$\frac{dm}{dt} = f(m), \quad t \in [0,T], \quad m(0) = m_0 \in \mathbb{R}^D, \tag{1.2}$$

together with the SDE (1.1), have a unique solution with no explosions.

Equation (1.2), or discretizations thereof, will be used to define the mean of our Gaussian approximations. The covariance of the time-marginals will be given by $C^\epsilon = \epsilon C$ where

$$\frac{dC}{dt} = Df(m)C + CDf(m)^T + \Sigma, \quad t \in [0,T], \quad C(0) = C_0 \in \mathbb{R}^{D\times D}, \tag{1.3}$$

or discretizations thereof, and where C_0 is a given positive semi-definite matrix —see Remark 1.1. This equation has a unique solution on $[0,T]$ under Assumption 1.1.

It is intuitively clear that, for small ϵ, the solution $v^\epsilon = \{v^\epsilon(t)\}_{0\le t\le T}$ of equation (1.1) is close to the deterministic solution $m = \{m(t)\}_{0\le t\le T}$ of (1.2). In Section 2 we show that, for small ϵ, the law of $v^\epsilon(t)$ is approximately Gaussian with mean $m(t)$ and covariance $C^\epsilon(t) = \epsilon C(t)$. More precisely, the error in this approximation, measured in the Kullback–Leibler divergence, is of order ϵ. Indeed we show more: the law μ^ϵ of the paths v^ϵ can be approximated by a Gaussian measure ν^ϵ in $C([0,T],\mathbb{R}^D)$. Subsection 1.3 shows that the choice of m as the mean of the Gaussian approximation is crucial. If any other function is chosen as mean, the Kullback–Leibler error remains of order 1.

REMARK 1.1. Our results will cover two possible initializations of the SDE (1.1). First, deterministic initial condition $v^\epsilon(0) = v_0$, corresponding to $\mu_0^\epsilon = \delta_{v_0}$. In this case equations (1.2) and (1.3) should be initialized with $m_0 = v_0$ and $C_0 = 0$, corresponding to the same Dirac measure. Second, where μ_0^ϵ has positive Lebesgue density in $\mathbb{R}^D$, and $D_{\mathrm{KL}}\left(N(m_0,\epsilon C_0)\|\mu_0^\epsilon\right)$ is of order ϵ.

REMARK 1.2. In Sections 3 and 4 we discretize equations (1.2) and (1.3) with step-size Δt. Assumption 1.1 and the convergence of the methods employed imply that, for all sufficiently small Δt, the discretized means and covariances are uniformly bounded. That is, there is $M > 0$ such that, for all sufficiently small Δt, $|m_k^{\Delta t}|, |C_k^{\Delta t}| \le M$, provided that $0 \le k\Delta t \le T$.

1.2. Background on Kullback–Leibler divergence. Let ν and μ be two probability measures on a measurable space $(X,\mathcal{X})$. The Kullback–Leibler divergence (also known as relative entropy) of ν with respect to μ is given by

$$D_{\mathrm{KL}}(\nu\|\mu) := \mathbb{E}^\nu \log\left(\frac{d\nu}{d\mu}\right)$$

if ν is absolutely continuous with respect to μ, denoted $\nu \ll \mu$, and $D_{\mathrm{KL}}(\nu\|\mu) = \infty$ otherwise. The Kullback–Leibler divergence satisfies $D_{\mathrm{KL}}(\nu\|\mu) \ge 0$, but it is not a metric on the space of probability measures since it may not be finite, it is not symmetric, and it does not satisfy the triangle inequality. However, it does quantify the proximity of the measures ν and μ. For instance, it provides an upper bound on the total variation distance

$$d_{\mathrm{TV}}(\nu,\mu) := \sup\left\{|\nu(A) - \mu(A)| : A \in \mathcal{X}\right\}$$

via Pinsker's inequality [12]

$$d_{\mathrm{TV}}(\nu,\mu) \le D_{\mathrm{KL}}(\nu\|\mu)^{1/2}. \tag{1.4}$$

We will use the chain rule for Kullback–Leibler divergence [3], which is well known in information theory. We recall it in the next lemma. The result underlies the proof of Lemma 4.1, which is analogous to the continuous time result Lemma 2.1.

LEMMA 1.1. *Let X and Y be Polish spaces and ν and μ be probability measures on the measurable space $(X\times Y,\mathcal{F})$. Denote by ν_x and μ_x the first marginals of ν and μ and let $\nu(dy|x)$ and $\mu(dy|x)$ be stochastic kernels on Y given X for which, for $A\times B \in \mathcal{F}$,*

$$\nu(A\times B) = \int_A \nu(B|x)\nu_x(dx), \quad \mu(A\times B) = \int_A \mu(B|x)\mu_x(dx).$$

Then,

$$D_{\mathrm{KL}}(\nu\|\mu) = D_{\mathrm{KL}}(\nu_x\|\mu_x) + \mathbb{E}^{\nu_x} D_{\mathrm{KL}}\big(\nu(\cdot|x)\|\mu(\cdot|x)\big). \tag{1.5}$$

The non-negativity of the Kullback–Leibler divergence and equation (1.5) imply that any two corresponding marginals are closer in Kullback–Leibler than the full measures, i.e.

$$D_{\mathrm{KL}}(\nu_x\|\mu_x) \le D_{\mathrm{KL}}(\nu\|\mu). \tag{1.6}$$

The chain rule is powerful when studying the approximation, by a Gaussian measure ν, of a non-Gaussian measure μ for which $\mu(\cdot|x)$ is Gaussian: the structure in equation (1.5) allows exploitation of the Gaussianity of the kernels $\nu(\cdot|x)$ and $\mu(\cdot|x)$. This will become apparent in Section 4, Lemma 4.1.

1.3. Large deviations and the choice of mean. Small noise diffusions have been extensively studied using large deviations. An early and fundamental result in the theory can be found in the first edition of [5], where it was shown that the collection $\{v^\epsilon, \epsilon \in (0,1)\}$ defined by equation (1.1) with initial condition $v^\epsilon(0) = v_0$ satisfies a large deviation principle on $C([0,T],\mathbb{R}^D)$ with rate function

$$I(\varphi) := \inf_{u\in U_\varphi} \left\{ \frac{1}{2}\int_0^T |u(t)|^2\,dt \right\}, \tag{1.7}$$

where

$$U_\varphi := \left\{ u \in L^2([0,T],\mathbb{R}^D) : \varphi(t) = v_0 + \int_0^T f\big(\varphi(s)\big)\,ds + \sqrt{\Sigma}\int_0^t u(s)\,ds \right\},$$

for absolutely continuous φ with $\varphi(0) = v_0$, $U_\varphi = \emptyset$ for all other $\varphi \in C([0,T],\mathbb{R}^D)$, and the infimum over the empty set in definition (1.7) is taken to be ∞.

It follows from the definition of the rate function I in definition (1.7) that $I(\varphi) = 0$ iff the zero function $u \equiv 0$ is in U_φ. This holds iff φ is the solution to the ODE (1.2) with initial condition $\varphi(0) = v_0$. In a very rough sense, this implies that the probability of v^ϵ lying in a small tube centered around any function other than the solution to the ODE decays at least exponentially as $\epsilon \to 0$. That is, for small $\delta, \epsilon > 0$, we have, as $\epsilon \to 0$,

$$\mathbb{P}\big(v^\epsilon \in B_\delta(\varphi)\big) \approx \exp\left(-\frac{I(\varphi)}{\epsilon}\right)$$

where $B_\delta(\varphi)$ denotes the ball (in the supremum norm) of radius $\delta > 0$ and center $\varphi \in C([0,T],\mathbb{R}^D)$.

Combining the large deviation result and the Markov inequality we obtain the following:

LEMMA 1.2. *Let μ^ϵ be the law in $C([0,T],\mathbb{R}^D)$ of v^ϵ given by equation (1.1) with initial condition $v^\epsilon(0) = v_0$. Consider a Gaussian measure ν^ϵ in $C([0,T],\mathbb{R}^D)$ with mean $m \in C([0,T],\mathbb{R}^D)$ and time-marginal covariances $C^\epsilon(t) = \epsilon C(t)$. Then, unless m is the solution to equation (1.2) with initial condition $m(0) = v_0$ we have, as $\epsilon \to 0$,*

$$d_{\mathrm{TV}}(\nu^\epsilon, \mu^\epsilon) \to 1.$$

Note that this lemma and Pinsker's inequality (1.4) imply that the Kullback–Leibler divergence remains at least order one in the small ϵ limit, unless m solves equation (1.2).

The analysis in [5] also shows that, *asymptotically*, the rescaled fluctuations of the process v^ϵ around its mean follow a Gaussian process. These asymptotic normality is fundamental in the study of small noise diffusions, e.g. [9]. We will give *explicit bounds* that further justify the use of Gaussian process approximations for sufficiently small ϵ, and we will extend the analysis to discrete settings.

2. Gaussian approximation of SDEs via ODEs

Let ν^ϵ be the law of l^ϵ defined via the linear SDE

$$dl^\epsilon(t) = \Big(f\big(m(t)\big) + Df\big(m(t)\big)\big(l^\epsilon(t) - m(t)\big)\Big)dt + \sqrt{\epsilon\Sigma}\,dW(t), \quad l^\epsilon(0) \sim N\big(m_0, \epsilon C_0\big). \tag{2.1}$$

The main result of this section is Theorem 2.1 below. It shows that, for small ϵ, the Gaussian measure ν^ϵ accurately approximates μ^ϵ, the law of v^ϵ given by equation (1.1). The proof is based on two observations. First, that ν^ϵ has time-marginals $\nu_t^\epsilon = N\big(m(t), C^\epsilon(t)\big)$, where m solves the ODE (1.2), and $C^\epsilon = \epsilon C$ with C solving equation (1.3). Second, the following lemma:

LEMMA 2.1. *Let $g_t(\cdot): \mathbb{R}^D \to \mathbb{R}^D$ be an affine function defined by*

$$g_t(l) = a(t) + B(t)l$$

for some continuous functions $a: [0,T] \mapsto \mathbb{R}^D$ and $B: [0,T] \mapsto \mathbb{R}^{D\times D}$. Let $\nu_0^\epsilon \ll \mu_0^\epsilon$, and let μ^ϵ and ν^ϵ be the laws in $C([0,T],\mathbb{R})$ of v^ϵ and l^ϵ given by

$$\begin{aligned} dv^\epsilon &= f(v^\epsilon)dt + \sqrt{\epsilon\Sigma}\,dW, \quad v(0) \sim \mu_0^\epsilon, \\ dl^\epsilon &= g_t(l^\epsilon)dt + \sqrt{\epsilon\Sigma}\,dW, \quad l^\epsilon(0) \sim \nu_0^\epsilon. \end{aligned}$$

Then

$$D_{\mathrm{KL}}(\nu^\epsilon \| \mu^\epsilon) = D_{\mathrm{KL}}(\nu_0^\epsilon \| \mu_0^\epsilon) + \frac{1}{2\epsilon}\mathbb{E}^{\nu^\epsilon}\int_0^T |f(v) - g_t(v)|_\Sigma^2\, dt.$$

Proof. We apply Girsanov's theorem in the form of Theorem 11A of [4], similarly as in Lemma 5.2 of [6]. Because equation (1.1) is assumed to have no explosions in the time-interval $[0,T]$ the density of μ^ϵ with respect to Wiener measure exists. Indeed the two measures are equivalent, and the density is given by the usual Girsanov formula. Furthermore ν^ϵ defines a Gaussian measure, equivalent to Wiener measure. Dividing the densities of ν^ϵ and μ^ϵ with respect to Wiener measure gives

$$\frac{d\nu^\epsilon}{d\mu^\epsilon}(v) = \frac{d\nu_0^\epsilon}{d\mu_0^\epsilon}\big(v(0)\big)\exp\Big(\frac{1}{\epsilon}\int_0^T \langle g_t(v) - f(v), dv\rangle_\Sigma - \frac{1}{2\epsilon}\int_0^T |g_t(v)|_\Sigma^2 - |f(v)|_\Sigma^2\, dt\Big).$$

Therefore,

$$D_{\mathrm{KL}}(\nu^\epsilon \| \mu^\epsilon) = \mathbb{E}^{\nu^\epsilon}\log\Big(\frac{d\nu^\epsilon}{d\mu^\epsilon}\Big) = D_{\mathrm{KL}}(\nu_0^\epsilon \| \mu_0^\epsilon) + I,$$

where using the martingale property of Ito's integral

$$I = \frac{1}{\epsilon}\mathbb{E}^{\nu^\epsilon}\Big(\int_0^T \langle g_t(v) - f(v), dv\rangle_\Sigma - \frac{1}{2}\int_0^T |g_t(v)|_\Sigma^2 - |f(v)|_\Sigma^2\, dt\Big)$$

$$= \frac{1}{\epsilon}\mathbb{E}^{\nu^\epsilon}\left(\int_0^T \langle g_t(v) - f(v), \sqrt{\epsilon\Sigma}\, dW\rangle_\Sigma + \frac{1}{2}\int_0^T |f(v)-g_t(v)|^2_\Sigma\, dt\right)$$
$$= \frac{1}{2\epsilon}\mathbb{E}^{\nu^\epsilon}\int_0^T |f(v)-g_t(v)|^2_\Sigma\, dt.$$

□

THEOREM 2.1. *Suppose that Assumption 1.1 holds. Let μ^ϵ and ν^ϵ be the laws in $C([0,T],\mathbb{R}^D)$ of, respectively, v^ϵ given by equation (1.1) and l^ϵ given by equation (2.1). Then there is $c>0$ such that, for all ϵ sufficiently small,*

$$D_{\mathrm{KL}}(\nu^\epsilon\|\mu^\epsilon) \le D_{\mathrm{KL}}(\nu_0^\epsilon\|\mu_0^\epsilon) + c\epsilon.$$

Moreover, for $t\in[0,T]$,

$$D_{\mathrm{KL}}(\nu_t^\epsilon\|\mu_t^\epsilon) \le D_{\mathrm{KL}}(\nu_0^\epsilon\|\mu_0^\epsilon) + c\epsilon,$$

where ν_t^ϵ and μ_t^ϵ denote the time-marginals of ν^ϵ and μ^ϵ at time t.

Proof. If ν_0^ϵ is not absolutely continuous with respect to μ_0^ϵ, then $D_{\mathrm{KL}}(\nu_0^\epsilon\|\mu_0^\epsilon)=\infty$ and the result is trivial. Thus we assume that absolute continuity holds. Throughout the proof c denotes a positive constant which is independent of all sufficiently small ϵ, and may vary from line to line. As noted before, ν^ϵ has time-marginals $\nu_t = N\big(m(t),C^\epsilon(t)\big)$ with $C^\epsilon=\epsilon C$. This, combined with Lemma 2.1 with the choice $g_t(\cdot) := f\big(m(t)\big) + Df\big(m(t)\big)\big(\cdot - m(t)\big)$, gives

$$D_{\mathrm{KL}}(\nu^\epsilon\|\mu^\epsilon) = D_{\mathrm{KL}}(\nu_0^\epsilon\|\mu_0^\epsilon) + \frac{1}{2\epsilon}\int_0^T \mathbb{E}^{N(m(t),C^\epsilon(t))}|f(v)-g_t(v)|^2_\Sigma dt. \tag{2.2}$$

Recall Taylor's formula with reminder

$$\big(f(u)-g_t(u)\big)_d = 2\sum_{|\alpha|=2}\frac{(u-m)^\alpha}{\alpha!}\int_0^1 (1-t)^2\partial^\alpha f_d\big(m+t(u-m)\big)\,dt, \quad 1\le d\le D,$$

to deduce, using Assumption 1.1, that

$$|\big(f(u)-g_t(u)\big)| \le c|u-m|^2\int_0^1 1+|m+t(u-m)|^s\,dt$$
$$\le c\sum_{r=0}^s |u-m|^{r+2}.$$

Thus,

$$|f(u)-g_t(u)|^2 \le c\sum_{r=0}^s |u-m|^{2r+4}. \tag{2.3}$$

Combining equation (2.2) with inequality (2.3) and using that $C^\epsilon=\epsilon C$ yields, for all ϵ sufficiently small,

$$\begin{aligned} D_{\mathrm{KL}}(\nu^\epsilon\|\mu^\epsilon) &= D_{\mathrm{KL}}(\nu_0^\epsilon\|\mu_0^\epsilon) + \frac{1}{2\epsilon}\int_0^T \mathbb{E}^{N(m(t),C^\epsilon(t))}|f(v)-g_t(v)|^2_\Sigma dt \\ &\le D_{\mathrm{KL}}(\nu_0^\epsilon\|\mu_0^\epsilon) + \frac{c}{\epsilon}\int_0^T \mathbb{E}^{N(m(t),C^\epsilon(t))}\sum_{r=0}^s |v(t)-m(t)|^{2r+4} dt \\ &\le D_{\mathrm{KL}}(\nu_0^\epsilon\|\mu_0^\epsilon) + \frac{c}{\epsilon}\max_{0\le r\le s}\int_0^T \left(\mathbb{E}^{N(m(t),C^\epsilon(t))}|v(t)-m(t)|^2\right)^{r+2} dt. \\ &\le D_{\mathrm{KL}}(\nu_0^\epsilon\|\mu_0^\epsilon) + c\epsilon, \end{aligned}$$

which completes the proof of the first claim. The bound for the marginals then follows from inequality (1.6). ∎

REMARK 2.1. It is not difficult to see that if f is linear then

$$D_{\mathrm{KL}}(\nu^\epsilon \| \mu^\epsilon) = D_{\mathrm{KL}}(\nu_0^\epsilon \| \mu_0^\epsilon).$$

This is the well-known Kullback–Leibler stability of the Fokker–Planck equation with respect to initial conditions, of which a more general version can be found in [2].

3. Gaussian approximations of SDEs via discretized ODEs

In this section we study the approximation of μ^ϵ, the law in $C([0,T],\mathbb{R}^D)$ of v^ϵ given by equation (1.1), by a Gaussian $\nu^{\epsilon,\Delta t}$. The path measure $\nu^{\epsilon,\Delta t}$ will be constructed so that the means and covariances of the time-marginals are given by numerical approximations with step size Δt to the ODEs (1.2) and (1.3), respectively. For simplicity, we study the effect of discretizing the ODEs with the Euler method, but the results extend with no effort to other Runge-Kutta methods and numerical schemes. Theorem 3.1 below bounds $D_{\mathrm{KL}}(\nu^{\epsilon,\Delta t}\|\mu^\epsilon)$ in terms of ϵ and the step-size Δt.

We now spell out the construction of the measure $\nu^{\epsilon,\Delta t}$. Given an integer $K>0$, let $\Delta t = T/K$ and define, for $1 \le k \le K$,

$$\begin{aligned} m_{k+1}^{\Delta t} &:= m_k^{\Delta t} + \Delta t f(m_k^{\Delta t}), \quad m_0^{\Delta t} = m_0, \\ C_{k+1}^{\Delta t} &:= C_k^{\Delta t} + \Delta t \Big(Df(m_k^{\Delta t}) C_k^{\Delta t} + C_k^{\Delta t} Df(m_k^{\Delta t})^T + \Sigma \Big), \quad C_0^{\Delta t} = C_0. \end{aligned} \tag{3.1}$$

Let $t_k = k\Delta t, 0 \le k \le K$, and define piecewise linear functions $m^{\Delta t}$ and $C^{\Delta t}$ in $[0,T]$ by interpolation. That is, for $t \in (t_k, t_{k+1})$,

$$\begin{aligned} m^{\Delta t}(t) &:= m_k^{\Delta t} + (t-t_k) f(m_k^{\Delta t}), \\ C^{\Delta t}(t) &:= C_k^{\Delta t} + (t-t_k) \Big(Df(m_k^{\Delta t}) C_k^{\Delta t} + C_k^{\Delta t} Df(m_k^{\Delta t})^T + \Sigma \Big). \end{aligned} \tag{3.2}$$

Finally, we let $\nu^{\epsilon,\Delta t}$ be the law of $l^{\epsilon,\Delta t}$ defined via the piecewise linear SDE

$$dl^{\epsilon,\Delta t} = g_t^{\Delta t}(l^{\epsilon,\Delta t})\, dt + \sqrt{\epsilon\Sigma}\, dW, \quad l^{\epsilon,\Delta t}(0) \sim N\big(m_0, \epsilon C_0\big), \tag{3.3}$$

with

$$g_t^{\Delta t}(\cdot) := f(m_k^{\Delta t}) + Df(m_k^{\Delta t})\big(\cdot - m^{\Delta t}(t)\big), \quad t \in (t_k, t_{k+1}). \tag{3.4}$$

By construction $\nu^{\epsilon,\Delta t}$ has time marginals $\nu_t^{\epsilon,\Delta t} = N\big(m^{\Delta t}(t), C^{\epsilon,\Delta t}(t)\big)$ with $m^{\Delta t}$ and $C^{\epsilon,\Delta t} := \epsilon C^{\Delta t}$ defined by expression (3.2).

REMARK 3.1. For fixed $\Delta t > 0$, $m^{\Delta t} \neq m$ except in trivial cases. Thus, by Lemma 1.2 it is necessary to let Δt depend on ϵ in order to have accurate approximations $\nu^{\epsilon,\Delta t}$ of μ^ϵ in the limit $\epsilon \to 0$.

THEOREM 3.1. *Suppose that Assumption 1.1 holds. Let μ^ϵ be the law in $C([0,T],\mathbb{R}^D)$ of v^ϵ given by equation (1.1), and let $\nu^{\epsilon,\Delta t}$ be as above. For $K \in \mathbb{N}$ let $\Delta t = T/K$. Then there is $c > 0$, independent of all sufficiently small Δt and ϵ, such that*

$$D_{\mathrm{KL}}(\nu^{\epsilon,\Delta t}\|\mu^\epsilon) \le D_{\mathrm{KL}}(\nu_0^{\epsilon,\Delta t}\|\mu_0^\epsilon) + c\epsilon + c\frac{(\Delta t)^2}{\epsilon}.$$

Moreover, for $t \in (0,T]$,

$$D_{\mathrm{KL}}(\nu_t^{\epsilon,\Delta t} \| \mu_t^\epsilon) \le D_{\mathrm{KL}}(\nu_0^{\epsilon,\Delta t} \| \mu_0^\epsilon) + c\epsilon + c\frac{(\Delta t)^2}{\epsilon},$$

where $\nu_t^{\epsilon,\Delta t}$ *and* $\mu_t^{\epsilon,\Delta t}$ *denote the marginals of* $\nu^{\epsilon,\Delta t}$ *and* $\mu^{\epsilon,\Delta t}$ *at time* t.

Proof. Throughout the proof c is a positive constant that may change from line to line, and is independent of all sufficiently small ϵ and Δt. We recall that $\nu^{\epsilon,\Delta t} = N(m^{\Delta t}, C^{\epsilon,\Delta t})$ is the law of the SDE (3.3). This, combined with Lemma 2.1 applied with $g_t = g_t^{\Delta t}$ as defined by (3.4), gives

$$D_{\mathrm{KL}}(\nu^{\epsilon,\Delta t} \| \mu^\epsilon) = D_{\mathrm{KL}}(\nu_0^{\epsilon,\Delta t} \| \mu_0^\epsilon) + \frac{1}{2\epsilon} \int_0^T \mathbb{E}^{\nu_t^{\epsilon,\Delta t}} |f(v) - g_t^{\Delta t}(v)|_\Sigma^2 \, dt. \tag{3.5}$$

We split the integral as follows:

$$|f(v) - g_t^{\Delta t}(v)|_\Sigma^2 \le c|g_t(v) - g_t^{\Delta t}(v)|_\Sigma^2 + c|f(v) - g_t(v)|_\Sigma^2, \tag{3.6}$$

and bound each of the two terms. For the first one note that, for $t \in (t_k, t_{k+1})$ and all sufficiently small Δt,

$$\begin{aligned} |g_t(v) - g_t^{\Delta t}(v)| &\le |f\big(m(t)\big) - f(m_k)| + |Df(m_k)||m(t) - m_k| + \cdots \\ &\quad |Df\big(m(t)\big) - Df(m_k)||v(t) - m^{\Delta t}(t)| + |Df\big(m(t)\big) - Df(m_k)||m^{\Delta t} - m_k| \\ &\le c\Delta t + c\Delta t|v(t) - m^{\Delta t}|. \end{aligned}$$

Thus,

$$|g_t(v) - g_t^{\Delta t}(v)|_\Sigma^2 \le c(\Delta t)^2 + c(\Delta t)^2 |v(t) - m^{\Delta t}|^2. \tag{3.7}$$

For the second one, as in the proof of Theorem 2.1, we have for small enough Δt that

$$\begin{aligned} |f(v) - g_t(v)|^2 &\le c \sum_{r=0}^s |v - m|^{r+2} \\ &\le c \sum_{r=0}^s |v - m^{\Delta t}|^{r+2} + c \sum_{r=0}^s |m^{\Delta t} - m|^{r+2} \\ &\le c \sum_{r=0}^s |v - m^{\Delta t}|^{r+2} + c(\Delta t)^2. \end{aligned}$$

Putting everything together, for all Δt and ϵ sufficiently small,

$$\begin{aligned} D_{\mathrm{KL}}(\nu^{\epsilon,\Delta t} \| \mu^\epsilon) &= D_{\mathrm{KL}}(\nu_0^{\epsilon,\Delta t} \| \mu_0^\epsilon) + \frac{1}{2\epsilon} \int_0^T \mathbb{E}^{\nu_t^{\epsilon,\Delta t}} |f(v) - g_t^{\Delta t}(v)|_\Sigma^2 \, dt \\ &\le D_{\mathrm{KL}}(\nu_0^{\epsilon,\Delta t} \| \mu_0^\epsilon) + \frac{c}{\epsilon}(\Delta t)^2 + \frac{c}{\epsilon} \int_0^T \mathbb{E}^{N(m^{\Delta t}(t), C^{\epsilon,\Delta t})} \sum_{r=0}^s |v - m^{\Delta t}|^{r+2} \\ &\le D_{\mathrm{KL}}(\nu_0^{\epsilon,\Delta t} \| \mu_0^\epsilon) + \frac{c}{\epsilon}(\Delta t)^2 + c\epsilon. \end{aligned}$$

∎

4. Gaussian approximation of discretized SDEs via discretized ODEs Consider the Euler–Maruyama discretization of equation (1.1)

$$v_{k+1}^{\epsilon} = v_k^{\epsilon} + f(v_k^{\epsilon})\Delta t + \sqrt{\epsilon\Sigma\Delta t}\,\xi_k, \quad v_0^{\epsilon} \sim \mu_0^{\epsilon}, \tag{4.1}$$

where the ξ_k are independently drawn from a standard Gaussian distribution. In this section we consider $\Delta t > 0$ small and fixed, and analyze small ϵ limits. A structure of the form (4.1) arises from discretization of SDEs, but also in many stochastic algorithms [7]. In the case of interest where the function f is nonlinear, the distribution μ_k^{ϵ} of v_k^{ϵ} is not Gaussian. We study the approximation of these measures by Gaussians ν_k^{ϵ}, whose means and covariances are built using discretizations of the ODEs (1.2) and (1.3).

We now detail the construction of the measures ν_k^{ϵ}. Let

$$m_{k+1} = m_k + f(m_k)\Delta t, \tag{4.2}$$

$$C_{k+1} = \big(I + Df(m_k)\Delta t\big)C_k\big(I + Df(m_k)\Delta t\big)^T + \Sigma\Delta t. \tag{4.3}$$

These agree with the discretization (3.1) used in Section 3, except for an extra $(\Delta t)^2$ term in the covariance. We set $C_k^{\epsilon} := \epsilon C_k$, and finally $\nu_k^{\epsilon} = N(m_k, C_k^{\epsilon})$.

The subsequent analysis is parallel to that of the previous sections. We again use two observations. First, that the ν_k^{ϵ} are the laws of l_k^{ϵ} given by

$$l_{k+1}^{\epsilon} = l_k^{\epsilon} + \big(f(m_k) + Df(m_k)(l_k^{\epsilon} - m_k)\big)\Delta t + \sqrt{\epsilon\Sigma\Delta t}\,\xi_k. \tag{4.4}$$

Second, the following lemma, analogous to Lemma 2.1. A derivation of this result can be found for instance in the appendix of [1]. We include a short proof that highlights how the chain rule Lemma 1.1 makes the Kullback–Leibler divergence well suited for the analysis of conditionally Gaussian dynamics, as those defined by equation (4.1).

LEMMA 4.1. *Let $g_k : \mathbb{R}^D \to \mathbb{R}^D$. Let μ_k^{ϵ} and ν_k^{ϵ} be the laws of v_k^{ϵ} and l_k^{ϵ} in $\mathbb{R}^D$ given by*

$$l_{k+1}^{\epsilon} = l_k^{\epsilon} + g_k(l_k^{\epsilon})\Delta t + \sqrt{\epsilon\Sigma\Delta t}\,\xi_k, \quad l_0 \sim \nu_0^{\epsilon},$$
$$v_{k+1}^{\epsilon} = v_k^{\epsilon} + f(v_k^{\epsilon})\Delta t + \sqrt{\epsilon\Sigma\Delta t}\,\xi_k, \quad v_0^{\epsilon} \sim \mu_0^{\epsilon}.$$

Denote by $\mu_{0:k}^{\epsilon}$ and $\nu_{0:k}^{\epsilon}$ the law of $(v_j^{\epsilon})_{j=0}^k$ and $(l_j^{\epsilon})_{j=0}^k$ in $\mathbb{R}^{D(k+1)}$. Then

$$D_{\mathrm{KL}}(\nu_{0:k}^{\epsilon} \| \mu_{0:k}^{\epsilon}) = D_{\mathrm{KL}}(\nu_0^{\epsilon} \| \mu_0^{\epsilon}) + \frac{\Delta t}{2\epsilon}\sum_{j=0}^{k-1} \mathbb{E}^{\nu_j^{\epsilon}} |f - g_j|_{\Sigma}^2. \tag{4.5}$$

Proof. We show that

$$D_{\mathrm{KL}}(\nu_{0:k+1}^{\epsilon} \| \mu_{0:k+1}^{\epsilon}) = D_{\mathrm{KL}}(\nu_{0:k}^{\epsilon} \| \mu_{0:k}^{\epsilon}) + \frac{\Delta t}{2\epsilon}\mathbb{E}^{\nu_k^{\epsilon}} |f - g_k|_{\Sigma}^2. \tag{4.6}$$

Iterating equation (4.6) gives equation (4.5). To prove equation (4.6) we apply Lemma 1.1 with $\nu = (\nu_{0:k}, \nu_{k+1})$, $\mu = (\mu_{0:k}, \mu_{k+1})$, and the kernels

$$\mu(\cdot|x) = N\Big(x + f(x)\Delta t, \epsilon\Sigma\Delta t\Big), \quad \nu(\cdot|x) = N\Big(x + g_k(x)\Delta t, \epsilon\Sigma\Delta t\Big).$$

To conclude, recall that the Kullback–Leibler between two Gaussians with the same covariance is

$$D_{\mathrm{KL}}\big(N(m, C) \| N(\bar{m}, C)\big) = \frac{1}{2}|\bar{m} - m|_C^2.$$

□

The following result shows that the Gaussians ν_k^ϵ accurately approximate the distribution μ_k^ϵ of v_k^ϵ in the small noise limit. More so, the result holds in path space.

THEOREM 4.1. *Suppose that Assumption 1.1 holds. Let $\mu_{0:k}^\epsilon$ be the law of $(v_j^\epsilon)_{j=0}^k$ given by (4.1), and let $\nu_{0:k}^\epsilon$ be the law of $(l_j^\epsilon)_{j=0}^k$ given by equation (4.4). Then, for all sufficiently small Δt and ϵ, and for all k with $k\Delta t < T$,*

$$D_{\mathrm{KL}}(\nu_{0:k}^\epsilon \| \mu_{0:k}^\epsilon) \le D_{\mathrm{KL}}(\nu_0^\epsilon \| \mu_0^\epsilon) + c\epsilon, \quad D_{\mathrm{KL}}(\nu_k^\epsilon \| \mu_k^\epsilon) \le D_{\mathrm{KL}}(\nu_0^\epsilon \| \mu_0^\epsilon) + c\epsilon,$$

where $c > 0$ is independent of the noise strength ϵ.

Proof. Throughout the proof c denotes a positive constant that is independent of ϵ and may change from line to line. We use the previous lemma with the choice

$$g_k(u) := f(m_k) + Df(m_k)(u - m_k)$$

together with the observation made in equation (4.4) to deduce that, for k with $k\Delta t \le T$,

$$\begin{aligned} D_{\mathrm{KL}}(\nu_{0:k}^\epsilon \| \mu_{0:k}^\epsilon) &= D_{\mathrm{KL}}(\nu_0^\epsilon \| \mu_0^\epsilon) + \frac{\Delta t}{2\epsilon} \sum_{j=0}^{k-1} \mathbb{E}^{\nu_j^\epsilon} |f - g_j|_\Sigma^2 \\ &\le D_{\mathrm{KL}}(\nu_0^\epsilon \| \mu_0^\epsilon) + \frac{c}{\epsilon} \max_{k:k\Delta t \le T} \mathbb{E}^{N(m_k, C_k^\epsilon)} |f(u) - g_k(u)|^2. \end{aligned} \tag{4.7}$$

Now, using Remark 1.2 it can be shown as in Theorem 2.1 that

$$|f(u) - g_k(u)|^2 \le c \sum_{r=0}^{s} |u - m_k|^{2r+4}. \tag{4.8}$$

To conclude we combine inequalities (4.7) and (4.8), and recall that $C_k^\epsilon = \epsilon C_k$ and Remark 1.2 to deduce that, for k with $k\Delta t \le T$ and all sufficiently small ϵ,

$$\begin{aligned} D_{\mathrm{KL}}(\nu_{0:k}^\epsilon \| \mu_{0:k}^\epsilon) &\le D_{\mathrm{KL}}(\nu_0^\epsilon \| \mu_0^\epsilon) + \frac{c}{\epsilon} \max_{k:k\Delta t \le T} \mathbb{E}^{N(m_k, C_k^\epsilon)} \sum_{r=0}^{s} |u - m_k|^{2r+4} \\ &\le D_{\mathrm{KL}}(\nu_0^\epsilon \| \mu_0^\epsilon) + c\epsilon. \end{aligned}$$

□

EXAMPLE 4.1. *Again if f is linear then, for $k \ge 1$,*

$$D_{\mathrm{KL}}(\nu_{0:k}^\epsilon \| \mu_{0:k}^\epsilon) = D_{\mathrm{KL}}(\nu_0^\epsilon \| \mu_0^\epsilon), \quad D_{\mathrm{KL}}(\nu_k^\epsilon \| \mu_k^\epsilon) \le D_{\mathrm{KL}}(\nu_0^\epsilon \| \mu_0^\epsilon).$$

As an example, suppose that $D = 1$, $f(u) = au + b$, and μ_0^ϵ is Gaussian in $\mathbb{R}$. Then a direct proof of the last inequality can be easily obtained. Indeed, it boils down to showing that, for any $m, \tilde{m}, \sigma, \tilde{\sigma}, a, b \in \mathbb{R}$ and $\epsilon > 0$,

$$D_{\mathrm{KL}}\big(N(am+b, a^2\sigma^2 + \epsilon) \| N(a\tilde{m} + b, a^2\tilde{\sigma}^2 + \epsilon)\big) \le D_{\mathrm{KL}}\big(N(m, \sigma^2) \| N(\tilde{m}, \tilde{\sigma}^2)\big).$$

In other words,

$$\frac{a^2\sigma^2 + \epsilon}{a^2\tilde{\sigma}^2 + \epsilon} - \log\Big(\frac{a^2\sigma^2 + \epsilon}{a^2\tilde{\sigma}^2 + \epsilon}\Big) - 1 + \frac{a^2(m - \tilde{m})^2}{a^2\tilde{\sigma}^2 + \epsilon} \le \frac{\sigma^2}{\tilde{\sigma}^2} - \log\Big(\frac{\sigma^2}{\tilde{\sigma}^2}\Big) - 1 + \frac{(m - \tilde{m})^2}{\tilde{\sigma}^2},$$

which is immediate since the sum of the first two terms in the LHS is smaller than the sum of the first two terms in the RHS and the last term in the LHS is smaller than the last term in the RHS.

REFERENCES

[1] C. Archambeau, D. Cornford, M. Opper, and J. Shawe-Taylor, *Gaussian process approximations of stochastic differential equations*, Journal of Machine Learning Research, 11:1–16, 2007.

[2] J.M.C. Clark, D.L. Ocone, and C. Coumarbatch, *Relative entropy and error bounds for filtering of Markov processes*, Mathematics of Control, Signals and Systems, Springer, 12(4):346–360, 1999.

[3] P. Dupuis and R.S. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, Wiley, 1997.

[4] K.D. Elworthy, *Stochastic Differential Equations on Manifolds*, Cambridge University Press, 1982.

[5] M.I. Freidlin and A.D. Wentzell, *Random Perturbation of Dynamical Systems*, Springer Science & Business Media, 260, 2012.

[6] M. Hairer, A.M. Stuart, and J. Voss, *Analysis of SPDEs arising in path sampling part II: the nonlinear case*, Annals of Applied Probability, 17:1657–1706, 2007.

[7] H. Kushner and G.G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, Springer Science & Business Media, 35, 2003.

[8] K.J.H. Law, A.M. Stuart, and K.C. Zygalakis, *Data Assimilation: A Mathematical Introduction*, Springer, 62, 2015.

[9] K.J.H. Law, H. Tembine, and R. Tempone, *Deterministic mean-field ensemble Kalman filtering*, SIAM J. Sci. Comput., 38(3):A1251–A1279, 2016.

[10] F. LeGland and B. Wang, *Asymptotic normality and partially observed diffusions with small noise: application to FDI*, Stochastic Theory and Control, Springer, 267–282, 2002.

[11] Y. Lu, A.M. Stuart, and H. Weber, *Gaussian approximations for transition paths in Brownian dynamics*, SIAM J. Math. Anal., 49(4):3005–3047, 2017.

[12] M.S. Pinsker, *Information and information stability of random variables and processes*, Holdtein-Day, American Mathematical Monthly, 73(8), 1964.

[13] F.J. Pinski, G. Simpson, A.M. Stuart, and H. Weber, *Algorithms for Kullback–Leibler approximation for probability measures in infinite dimensions*, SIAM J. Sci. Comput., 37(6):2733–2757, 2015.

[14] F.J. Pinski, G. Simpson, A.M. Stuart, and H. Weber, *Kullback–Leibler approximation for probability measures on infinite dimensional spaces*, SIAM J. Math. Anal., 47(11):4091–4122, 2015.

[15] C.E. Rasmussen and C.K.I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 62, 2006.