DERIVATIVE-FREE BAYESIAN INVERSION USING MULTISCALE DYNAMICS

G. A. PAVLIOTIS, A. M. STUART, U. VAES

Abstract. Inverse problems are ubiquitous because they formalize the integration of data with mathematical models. In many scientific applications the forward model is expensive to evaluate, and adjoint computations are difficult to employ; in this setting derivative-free methods which involve a small number of forward model evaluations are an attractive proposition. Ensemble Kalman based interacting particle systems (and variants such as consensus based and unscented Kalman approaches) have proven empirically successful in this context, but suffer from the fact that they cannot be systematically refined to return the true solution, except in the setting of linear forward models [20]. In this paper, we propose a new derivative-free approach to Bayesian inversion, which may be employed for posterior sampling or for maximum a posteriori (MAP) estimation, and may be systematically refined. The method relies on a fast/slow system of stochastic differential equations (SDEs) for the local approximation of the gradient of the log-likelihood appearing in a Langevin diffusion. Furthermore the method may be preconditioned by use of information from ensemble Kalman based methods (and variants), providing a methodology which leverages the documented advantages of those methods, whilst also being provably refineable. We define the methodology, highlighting its flexibility and many variants, provide a theoretical analysis of the proposed approach, and demonstrate its efficacy by means of numerical experiments.

Key words. Inverse problems, Multiscale methods, Derivative-free methods.

AMS subject classifications. 62F15, 65C35, 65C30, 65N21.

1. Introduction.

1.1. Overview.

In this paper, we consider the inverse problem of finding an unknown parameter $\theta \in \mathbf{R}^d$ from data $y \in \mathbf{R}^K$ where

$$y = G(\theta) + \eta, \tag{1.1}$$

with $G : \mathbf{R}^d \to \mathbf{R}^K$ a forward operator and η the observational noise. In the Bayesian approach to inverse problems [35, 69, 14], the vectors θ, η and y are treated as random variables. If the unknown parameter and the noise are assumed to be independent and normally distributed, with distribution parameters $\theta \sim \mathcal{N}(m, \Sigma)$ and $\eta \sim \mathcal{N}(0, \Gamma)$, then the joint distribution of (θ, y) can be obtained from (1.1):

$$(\theta, y) \sim \frac{\mathrm{e}^{-\Phi_R(\theta;y)}}{\int_{\mathbf{R}^K}\int_{\mathbf{R}^d} \mathrm{e}^{-\Phi_R(\theta;y)}\,\mathrm{d}\theta\,\mathrm{d}y}.$$

Here Φ_R is a function that regularizes the *least-squares functional* Φ; these two functions are given by

$$\Phi(\theta; y) = \frac{1}{2}\,|y - G(\theta)|_\Gamma^2\,, \tag{1.2a}$$

$$\Phi_R(\theta; y) = \Phi(\theta; y) + \frac{1}{2}\,|\theta - m|_\Sigma^2\,, \tag{1.2b}$$

with the notation, for a positive definite matrix A,

$$\langle \bullet_1, \bullet_2 \rangle_A = \langle \bullet_1, A^{-1} \bullet_2 \rangle, \qquad |\bullet|_A^2 = \langle \bullet, \bullet \rangle_A\,,$$

where $\langle \bullet, \bullet \rangle$ is the Euclidean inner product. By Bayes' formula, the conditional probability density function of θ given y equals

$$\rho(\theta \,|\, y) = \frac{\exp\big(-\Phi_R(\theta; y)\big)}{\int_{\mathbf{R}^d} \exp\big(-\Phi_R(\theta; y)\big)\, \mathrm{d}\theta} =: \frac{1}{Z(y)} \exp\big(-\Phi_R(\theta; y)\big). \tag{1.3}$$

This probability distribution is the *Bayesian posterior*, and its pointwise maximizer is the *maximum a posteriori* (MAP) estimator.

There exist several approaches for solving inverse problems, which we review in the next section. In this paper, we present a new derivative-free approach for (1.1). Our method is based on a fast/slow system of SDEs, and it may be used for sampling from the Bayesian posterior (1.3) or for calculating the MAP estimator. Unlike the Ensemble Kalman Sampler (EKS), and variants such as the unscented Kalman sampler (UKS) [32] and consensus based sampler (CBS) [9], the method we present can be refined systematically in order to approach the true solution: in the refinement limit, it produces a stochastic process described by dynamics of the type

$$\mathrm{d}\theta_t = -K \nabla_\theta \Phi_R(\theta_t; y)\, \mathrm{d}t + \nu \sqrt{2K}\, \mathrm{d}w_t. \tag{1.4}$$

Here $\{w_t\}_{t \geq 0}$ is a standard d-dimensional Brownian motion, K is a symmetric positive definite matrix, and ν is a coefficient equal to 1 if the method is used for posterior sampling or 0 in optimization mode. When $\nu = 0$, equation (1.4) is a preconditioned gradient descent in the potential Φ_R, and, when $\nu = 1$, equation (1.4) is a preconditioned overdamped Langevin diffusion in the potential Φ_R. From now on, since the observation y is a fixed parameter of the inverse problem, we write $\Phi(\theta) = \Phi(\theta; y)$ and $\Phi_R(\theta) = \Phi_R(\theta; y)$ and $Z = Z(y)$ for simplicity.

1.2. Literature Review. There are two main approaches for solving inverse problems of the type (1.1): the classical approach and the Bayesian approach [35, 69]. Classical methods are generally based on an optimization problem of the form

$$\arg\min_{\theta \in \mathbf{R}^d} \frac{1}{2} |y - G(\theta)|_X^2 + R(\theta), \tag{1.5}$$

for some positive definite matrix X and where R is an optional regularization term. The aim of the regularization term is to ensure that the minimization problem is well-posed; without this term, there may be minimizing sequences that are not bounded in $\mathbf{R}^d$. See, for example, [2, 12, 13, 28, 29] for a discussion of the classical approach (1.5) and of regularization techniques. A widely used form for $R(\theta)$, known as a Tikhonov–Phillips regularization, is given by $R(\theta) = |\theta - z|_Y^2$, for some vector z and a positive definite matrix Y is . In the classical approach, the matrices X, Y and the vector z are generally parameters without a probabilistic interpretation.

The Bayesian approach to the inverse problem (1.1), on the other hand, relies on the statistical properties of the noise and on the specification of a prior probability distribution which encapsulates *a priori* knowledge on the unknown parameter, as shown in Subsection 1.1. In the Bayesian framework, the optimization problem (1.5) is relevant with $X = \Gamma$ and with the Tikhonov–Phillips regularization $R(\theta) = |\theta - m|_\Sigma^2$. In this case, the solution to (1.5) admits a clear interpretation: it is the pointwise maximizer of the Bayesian posterior (1.3), so it can be viewed as the most likely value of the parameter given the data. See [69] for more details on the connection between the classical and Bayesian approaches. Often, one is interested not in a

point estimator but in the statistical properties of the Bayesian posterior, which can be used, for example, for the derivation of confidence intervals. In most applications, the dimension of the parameter space is large, so it is necessary to generate samples from the Bayesian posterior in order to calculate its statistical properties.

Several methods can be employed for solving the inverse problem (1.1) via the optimization problem (1.5). In a number of important applications of inverse problems, such as parameter estimation in climate models [16], the derivatives of the forward operator G are unavailable or too computationally expensive to obtain, so, in this literature review, we only briefly review gradient-based methods and we focus mostly on derivative-free methods.

When the derivatives of the forward operator G are available, a natural approach is to employ the gradient descent algorithm or one of its variants: we mention, for example, the conjugate gradient descent [31], the stochastic gradient descent [63, 42], the Barzilai–Borwein method [4], and other gradient-based optimization techniques that rely on interacting particle systems [71, 5, 6]. One may also recur to the Newton or Gauss–Newton methods and their variants [24, 3], or to methodologies based on the Levenberg–Marquadt method [47, 51, 26].

When the derivatives of the forward operator G are unavailable, on the other hand, derivative-free methods are required for solving the optimization problem (1.5). A widely-used method of this type is Ensemble Kalman Inversion (EKI) [33], which is based on the ensemble Kalman filter (EnKF) [19] and relies on the least-squares structure of the optimization problem (1.5). The EKI has been shown, both theoretically in simple settings and empirically, to perform very well in the context of inverse problems [33, 67, 68], and it has also been applied successfully for training neural networks [40]. A recent variant on the EKI, unscented Kalman inversion (UKI), shows significant promise for problems in which the parameter dimension is low, but the forward model is expensive to evaluate and hard to differentiate [32]. Another alternative is the method developed in [25], which is based on similar ideas for gradient approximation but aims to drive a single distinguished particle to the optimizer. Another approach for solving (1.5) is to use a general-purpose derivative-free optimization method. We mention, for example, simulated annealing [38], particle-swarm optimization [37], and consensus-based optimization (CBO) [61, 8].

Likewise, there exist several methods for solving (1.1) via the Bayesian approach, i.e. for generating samples from the Bayesian posterior (1.3). If the derivatives of the log-posterior are available, the simplest option is to rely on a Langevin diffusion of the type (1.4), which enjoys the property of transforming any initial distribution into the Bayesian posterior (1.3) in the longtime limit $t \to \infty$. One may also employ higher-dimensional stochastic dynamics that admit the Bayesian posterior as a marginal of their ergodic measure, such as the underdamped Langevin dynamics [46, 56] or the generalized Langevin dynamics [54, 45, 57].

Another standard and related approach for sampling from a high-dimensional probability density is to use a Markov chain Monte Carlo method (MCMC), i.e. to construct a Markov chain whose unique invariant distribution is the target density. To this end, the most widely used method is the Metropolis–Hastings algorithm [53, 27] (MH). All that is required to define a MH algorithm is a proposal distribution, which may or may not be based on the derivatives of the target density (or of its logarithm). We mention, for example, the Metropolis-adjusted Langevin dynamics (MALA, which uses the derivative) and the random walk MH method (RWMH, which does not). There is also a substantial literature on the computation of, or exploiting, Gaussian approximations of the posterior; see [60] and the references therein. There is extensive literature on the convergence properties and optimal parametrization of these methods, and on their connections with overdamped Langevin diffusions [66, 64, 65] in the high-dimensional limit [66, 64, 65]. See also [34] for a study of the connection of high-dimensional RWMH with overdamped Langevin dynamics in the transient regime,

and [7] for a proof of convergence of MALA to an overdamped Langevin diffusion in the small timestep limit in fixed dimension.

In recent years, there has also been significant activity devoted to developing sampling methods based on interacting particle systems, which can leverage recent advances in parallel computing. These include, for example interacting particle MCMC methodologies [15] and [44], Stein variational gradient descent [49, 48], the ensemble Kalman sampler (EKS) [20], and affine-invariant Langevin dynamics (ALDI) [21]. The derivative-free formulations of the latter two methods were proposed specifically for Bayesian inverse problems – they rely on the least-squares structure (1.2a) of the log-posterior – and they were shown to produce good approximate samples of the posterior distribution at a relatively low computational cost. Both EKS and ALDI are strongly related to (1.4): in the linear setting, they are based on a system of preconditioned overdamped Langevin diffusions, with a time-dependent preconditioner given by the covariance of the ensemble. ALDI improves upon EKS by incorporating a correction term which guarantees that the ergodic measure of the finite-dimensional particle system is the product measure of J copies of the target distribution, where J denotes the number of particles. A variant on the EKS using the unscented transform (UKS) has recently been proposed [32], and a generalization of CBO to sampling (CBS) has recently been proposed [9]; both UKS and CBS are derivative-free.

The parallel MCMC method of [44], as well as the ensemble Kalman based methods for Bayesian inverse problems, i.e. EKI, EKS and ALDI, enjoy the property of being affine invariant in the sense of [22]; see also [43]. This property was studied carefully in [21] for EKS and ALDI, after it had been observed that, in the simple case of a linear forward model, the convergence rate of EKI and EKS was independent of the Hessian of the regularized least-squares functional Φ_R [20, 10]. As the terminology indicates, affine invariant methods are insensitive to affine transformations of the regularized least-squares functional Φ_R, which makes them particularly well-suited in cases where the Φ_R exhibits strong anisotropy at its minimizer.

As mentioned in Subsection 1.1, the method we present in this paper is based on a fast/slow system of stochastic differential equations, and it may be used both for MAP estimation and posterior sampling. Multiscale methods have been used before for optimization purposes. A multiscale dynamics is employed [11] for smoothing the loss function associated with deep neural networks, and the method is revisited later in [36]. A similar multiscale dynamics is also employed in [62] for calculating convolutions, with the aim of reducing metastability in the context of molecular dynamics. See also [70] for information on how smoothing the objective function by convolution with a Gaussian kernel can be helpful in optimization schemes. The method we propose in this paper is based on similar ideas, in that it employs a fast/slow system SDEs for approximating the gradient of the loss function, but it is gradient-free and relies on a different multiscale system. In addition, we demonstrate how preconditioning can be incorporated in the method in order to approach the solution to (1.4) with an appropriate symmetric, positive definite matrix K. We show, in particular, that an efficient numerical method can be obtained by constructing the preconditioner using information from EKS; similar ideas can be used based on information obtained through CBS or UKS.

1.3. Our Contributions. Our contributions in this paper are the following:

- We present a novel method based on a multiscale dynamics for MAP estimation and posterior sampling in Bayesian inverse problems. We discuss possible variations of the method and present a fully practical numerical discretization.
- In addition to motivating the method with formal arguments, we prove the pathwise convergence of the solution it produces to a gradient descent or to an overdamped Langevin diffusion, depending on whether the method is used for optimization or sampling, respectively. We also obtain a strong

convergence estimate for the numerical discretization of the multiscale dynamics.

- We present numerical experiments demonstrating the efficiency of the method, for the purposes of both sampling and optimization. We consider first a standard low-dimensional test problem and then a high-dimensional inverse problem where the forward model requires the solution of an elliptic PDE.
- We show how a significant improvement in performance can be obtained by preconditioning the method using information from EKS.

The rest of the paper is organized as follows. In Section 2, we introduce the multiscale method and present our main results. In Section 3, we present numerical experiments demonstrating the efficacy of the method, both for low-dimensional and high-dimensional parameter spaces, and we show how preconditioning can be incorporated in the method. In Section 4, we prove our main convergence results. Section 5 is reserved for conclusions and perspectives for future work.

2. Presentation of the Method and Main Results. This section is organized as follows: in Subsection 2.1, the multiscale method is presented as a continuous-time dynamics and motivated by formal arguments. In Subsection 2.2, a fully practical time discretization of the continuous dynamics is presented. Subsection 2.3 then presents the statements of our main results, the proofs of which are given in Section 4.

2.1. Continuous-time Dynamics. Our method is based on a multiscale system of stochastic differential equations (SDEs): A slow variable is employed for the purposes of finding the MAP estimator or sampling from the Bayesian posterior, and several fast variables provide information on the variation of the least-squares functional in the vicinity of the slow variable. At any time, the drift for the slow variable is calculated based on the values of the forward functional at the positions of these fast explorers, using a projected gradient approximation similar in structure to that used in the methods for inversion and sampling based on the ensemble Kalman filter (EnKF); see [67] (in the context of optimization) and [20] (in the context of sampling). The idea of employing an approximation based on ensemble Kalman methods for specifying the drift of a single distinguished particle is inspired by the paper [25], in which the authors proposed to use stochastic differences as a surrogate for gradients in an ensemble based optimization context.

In most applications, the space of the unknown parameter is $\mathbf{R}^d$ but, for simplicity of the analysis presented in Section 4, we also consider the case where this space is the d-dimensional torus $\mathbf{T}^d$. Therefore, we denote the parameter space by $\mathbf{K}^d$, with $\mathbf{K} = \mathbf{R}$ or $\mathbf{K} = \mathbf{T}$. At the continuous-time level, our method is based on the following system of interacting SDEs:

$$\dot\theta = -\frac{1}{J\sigma^2}\sum_{j=1}^{J}\langle G(\theta^{(j)}) - G(\theta), G(\theta) - y\rangle_\Gamma\,(\theta^{(j)} - \theta) - C(\Xi)\Sigma^{-1}(\theta - m) + \nu\sqrt{2C(\Xi)}\,\dot w, \tag{2.1a}$$

$$\theta^{(j)} = \theta + \sigma\,\xi^{(j)}, \qquad j = 1, \dots, J, \tag{2.1b}$$

$$\dot\xi^{(j)} = -\frac{1}{\delta^2}\,\xi^{(j)} + \sqrt{\frac{2}{\delta^2}}\,\dot w^{(j)}, \qquad \xi^{(j)}(0) \sim \mathcal{N}(0, I_d), \qquad j = 1, \dots, J, \tag{2.1c}$$

where $\theta \in \mathbf{K}^d$, $\Xi = (\xi^{(1)}, \dots, \xi^{(J)}) \in (\mathbf{R}^d)^J$, I_d is the $\mathbf{R}^{d\times d}$ is the identity matrix, the processes w and $\{w^{(j)}\}_{j=1}^J$ are independent standard Brownian motions and

$$C(\Xi) = \frac{1}{J}\sum_{j=1}^{J}(\xi^{(j)} \otimes \xi^{(j)}).$$

The processes $\{\xi^{(j)}\}_{j=1}^J$ are stationary Ornstein–Uhlenbeck (OU) processes with invariant measure $\mathcal{N}(0, I_d)$

and autocorrelation function $\mathrm{e}^{-|t|/\delta^2} I_d$. The parameter δ can therefore be viewed as the square root of the characteristic time scale of the fast processes. The coefficient $\nu \in \{0, 1\}$ controls whether noise should be included in the equation for θ: if $\nu = 0$, then (2.1) is a method for finding the minimizer of the regularized least-squares functional Φ_R, i.e. the MAP estimator; if $\nu = 1$, then (2.1) is a method for sampling from the posterior distribution $\frac{1}{Z}\mathrm{e}^{-\Phi_R}$, where Z is the normalization constant given in (1.3).

Note that (2.1) can also be employed without the prior regularization and with $\nu = 0$, as a method for finding the minimizer of the non-regularized least-squares functional Φ. This formally corresponds to taking the prior covariance to be infinite, i.e. $\Sigma = \infty I_d$, and can be useful when prior knowledge of the unknown parameter θ is not available, or is not needed because the problem is over-determined. In this case, any parameter $\theta \in G^{-1}(y)$ is a steady state of (2.1), which is not the case for the alternative derivative-free formulation (2.5) we present below.

REMARK 2.1 (Connection with the stochastic gradient descent). *The method (2.1) is most useful with σ small, in which case, neglecting quadratic or smaller terms in σ, the following approximation holds:*

$$G(\theta^{(k)}) - G(\theta) \approx (\theta^{(k)} - \theta) \cdot \nabla G(\theta). \tag{2.2}$$

Using this approximation, we can rewrite the equation for θ in (2.1) as

$$\begin{aligned}\dot\theta &\approx -\frac{1}{J}\sum_{k=1}^{J}\left(\xi^{(k)} \otimes \xi^{(k)}\right) \nabla\Phi(\theta) - C(\Xi)\Sigma^{-1}(\theta - m) + \nu\sqrt{2C(\Xi)}\,\dot w \\ &= -C(\Xi)\,\nabla\Phi_R(\theta) + \nu\sqrt{2C(\Xi)}\,\dot w.\end{aligned}$$

The term $C(\Xi)\,\nabla\Phi_R(\theta)$ can be viewed as a projection of $\nabla\Phi_R(\theta; y)$ on the subspace spanned by $\{\xi^{(1)}, \ldots, \xi^{(J)}\}$, which shows a link with the stochastic gradient descent algorithm. For large J, it holds formally that $C(\Xi) \approx I_d$ at all times, so the equation for θ reduces to a gradient descent when $\nu = 0$, or to the overdamped Langevin equation if $\nu = 1$, both both with respect to the potential Φ_R.

REMARK 2.2. *Note that (2.2) holds exactly when G is linear, for all $\sigma > 0$. In this case and in the presence of noise (i.e. when $\nu = 1$), equation (2.1) admits as invariant measure the distribution*

$$\rho_\infty(\theta, \Xi) = \frac{1}{Z}\,\exp\left(-\Phi_R(\theta)\right)\, g(\xi^{(1)}; 0, I_d)\ldots g(\xi^{(J)}; 0, I_d). \tag{2.3}$$

The associated marginal distribution for θ is given by $\frac{1}{Z}\mathrm{e}^{-\Phi_R}$, which is precisely the Bayesian posterior distribution. To show that (2.3) is indeed the unique invariant distribution when G is linear, we note that the Fokker–Planck operator associated with (2.1) in this case is given by [56, Chapter 4]

$$\begin{aligned}\mathcal{L}^\dagger\rho &= \nabla_\theta \cdot \Big(C(\Xi)\big(\nabla\Phi_R(\theta)\rho + \nabla_\theta\rho\big)\Big) + \frac{1}{\delta^2}\sum_{j=1}^{J}\nabla_{\xi^{(j)}}\cdot\left(\xi^{(j)}\rho + \nabla_{\xi^{(j)}}\rho\right) \\ &= \nabla_\theta\cdot\left(\rho_\infty C(\Xi)\nabla_\theta\left(\frac{\rho}{\rho_\infty}\right)\right) + \frac{1}{\delta^2}\sum_{j=1}^{J}\nabla_{\xi^{(j)}}\cdot\left(\rho_\infty\nabla_{\xi^{(j)}}\left(\frac{\rho}{\rho_\infty}\right)\right) = \nabla\cdot\left(\rho\mathcal{D}\nabla\log\left(\frac{\rho}{\rho_\infty}\right)\right),\end{aligned}$$

where $\mathcal{D}$ is the $\mathbf{R}^{d(J+1)\times d(J+1)}$ block diagonal matrix with diagonal blocks $C(\Xi), \frac{1}{\delta^2}I_d, \ldots, \frac{1}{\delta^2}I_d$. It is clear that ρ_∞ in (2.3) is in the kernel of this operator. To show formally that the invariant measure is unique, it suffices to multiply both sides of the equation $\mathcal{L}^\dagger\rho = 0$ by ρ/ρ_∞ and to integrate over the state space $\mathbf{K}^d \times (\mathbf{R}^d)^J$,

which gives

$$\int_{\mathbf{K}^d}\int_{\mathbf{R}^d}\cdots\int_{\mathbf{R}^d}\left(C(\Xi)\left|\nabla_\theta\left(\frac{\rho}{\rho_\infty}\right)\right|^2+\frac{1}{\delta^2}\sum_{j=1}^{J}\left|\nabla_{\xi^{(j)}}\left(\frac{\rho}{\rho_\infty}\right)\right|^2\right)\rho_\infty(\theta,\Xi)\,\mathrm{d}\xi^{(1)}\ldots\mathrm{d}\xi^{(J)}\,\mathrm{d}\theta=0,$$

and therefore $\rho=\rho_\infty$ necessarily.

For the purposes of analysis, we also consider a simplified version of (2.1) in which the coefficient of the noise is independent of the fast processes $\xi^{(1)},\ldots,\xi^{(J)}$:

$$\dot\theta=-\frac{1}{J\sigma^2}\sum_{j=1}^{J}\langle G(\theta^{(j)})-G(\theta),G(\theta)-y\rangle_\Gamma\,(\theta^{(j)}-\theta)-C(\Xi)\Sigma^{-1}(\theta-m)+\nu\sqrt{2}\dot w, \tag{2.4a}$$

$$\theta^{(j)}=\theta+\sigma\,\xi^{(j)},\qquad j=1,\ldots,J. \tag{2.4b}$$

$$\dot\xi^{(j)}=-\frac{1}{\delta^2}\,\xi^{(j)}+\sqrt{\frac{2}{\delta^2}}\,\dot w^{(j)}. \tag{2.4c}$$

Equation (2.4a) admits the same formal limit as $J\to\infty$ or $\delta\to 0$ as (2.1a), but it is simpler to analyze because the noise is additive. We note, however, that the invariant measure associated to (2.4a)–(2.4c) for finite J differs from (2.3), even in the case of a linear forward model.

Remark 2.3 (Second alternative derivative-free formulation). *Instead of (2.1) or (2.4), we could also use a system of equations of the form*

$$\dot\theta=-\frac{1}{J\sigma^2}\sum_{j=1}^{J}\Big(\Phi_R(\theta^{(j)})-\Phi_R(\theta)\Big)\Big(\theta^{(j)}-\theta\Big)+\nu\sqrt{2C(\Xi)}\,\dot w, \tag{2.5a}$$

$$\theta^{(j)}=\theta+\sigma\,\xi^{(j)},\qquad j=1,\ldots,J, \tag{2.5b}$$

$$\dot\xi^{(j)}=-\frac{1}{\delta^2}\,\xi^{(j)}+\sqrt{\frac{2}{\delta^2}}\,\dot w^{(j)}. \tag{2.5c}$$

This formulation has two advantages over (2.1) and (2.4): it does not require the prior distribution to be Gaussian, and it does not rely on the specific quadratic structure of Φ_R in (1.2a), making it more generally applicable. However, like (2.4), the system of equations (2.5) does not admit (2.3) as invariant measure when J is finite, not even in the case of a linear forward model.

2.2. Numerical Discretization. To integrate (2.1) numerically, we employ the Euler–Maruyama method for θ and a closed formula for the exact (in law) solution of the OU process for $\{\xi^{(j)}\}_{j=1}^J$. We use the notation Δ to denote the time step and the notation $(\hat\theta_n,\hat\Xi_n)$, with $\hat\Xi_n=(\hat\xi_n^{(1)},\ldots,\hat\xi_n^{(J)})$, to denote the numerical approximation of $(\theta_{n\Delta},\Xi_{n\Delta})$, i.e. the numerical approximation of the continuous-time solution at time $n\Delta$. We also denote by N the total number of iterations and by $T=N\Delta$ the final time of the simulation. The numerical scheme we propose reads

$$\begin{aligned}\hat\theta_{n+1}&=\hat\theta_n-\frac{1}{J\sigma}\sum_{j=1}^{J}\langle G(\hat\theta_n+\sigma\hat\xi_n^{(j)})-G(\hat\theta_n),G(\hat\theta_n)-y\rangle_\Gamma\,\hat\xi_n^{(j)}\,\Delta\\&\quad-C(\hat\Xi_n)\Sigma^{-1}(\hat\theta_n-m)\Delta+\nu\sqrt{2D(\hat\Xi_n)\Delta}\,x_n,\\\hat\xi_{n+1}^{(j)}&=\mathrm{e}^{-\frac{\Delta}{\delta^2}}\,\hat\xi_n^{(j)}+\sqrt{1-\mathrm{e}^{-\frac{2\Delta}{\delta^2}}}\,x_n^{(j)},\qquad \hat\xi_0^{(j)}\sim\mathcal{N}(0,I_d),\qquad j=1,\ldots,J,\end{aligned} \tag{2.6}$$

where x_n and $x_n^{(j)}$, for $n = 0, \dots, N-1$ and $j = 1, \dots, J$, are independent $\mathcal{N}(0, I_d)$ random variables. Here $D(\hat{\Xi}_n) = C(\hat{\Xi}_n)$ or $D(\hat{\Xi}_n) = I_d$ depending on whether a solution to (2.1) or to (2.4) is sought, respectively. At the numerical level, we can consider the limit $\delta \to 0^+$, in which case the numerical scheme (2.6) simplifies to

$$\begin{aligned} \hat{\theta}_{n+1} = \hat{\theta}_n - \frac{1}{J\sigma} \sum_{j=1}^{J} \langle G(\hat{\theta}_n + \sigma \hat{\xi}_n^{(j)}) - G(\hat{\theta}_n), G(\hat{\theta}_n) - y \rangle_\Gamma \, \hat{\xi}_n^{(j)} \, \Delta \\ - C(\hat{\Xi}_n)\Sigma^{-1}(\hat{\theta}_n - m)\Delta + \nu \sqrt{2D(\hat{\Xi}_n)\Delta} \, x_n, \end{aligned} \tag{2.7}$$

where $\hat{\xi}_n^{(j)}$, for $n = 0, \dots, N-1$ and $j = 1, \dots, J$, are drawn independently from $\mathcal{N}(0, I_d)$. This method is simpler to implement and analyze than (2.6), but is not always be the best option for the purposes of sampling and optimization. Indeed, it is possible that the choice $\delta > 0$, which introduces a correlation between the successive descent directions, promotes exploration, which could be useful e.g. when the energy landscape Φ_R is rugged. We leave this question for future work and, in all the numerical experiments presented in Section 3, we use the scheme (2.6) with $\delta > 0$. We do, however, analyze the convergence of (2.7) theoretically in Section 4, as a first step towards proving the convergence of (2.6).

2.3. Main Results. In this section, we present and comment on our main results, which are proved in Section 4. We assume throughout that the state space of the parameter is the d-dimensional torus $\mathbf{T}^d$, rather than $\mathbf{R}^d$, in order to simplify the analysis, but there is no conceptual obstruction to extending our findings to $\mathbf{R}^d$. In all our results, the minimum regularity requirement on the forward model is that $G \in C^2(\mathbf{T}^d, \mathbf{R}^K)$, but we also present refined estimates in the case of a more regular forward model, namely if $G \in C^3(\mathbf{T}^d, \mathbf{R}^K)$.

We first present a strong convergence result for the dynamics (2.4), which coincides with the dynamics (2.1) in the absence of noise. More precisely, Theorem 2.1 establishes, in the joint limit $\delta \to 0$ and $\sigma \to 0$, the pathwise convergence of the stochastic process $\{\theta_t\}_{t \in [0,T]}$ to the solution $\{\vartheta_t\}_{t \in [0,T]}$ of the averaged equation

$$\dot{\vartheta}_t = -\nabla \Phi_R(\vartheta_t) + \nu \sqrt{2} \, \dot{w}_t. \tag{2.8}$$

We use the notation μ_0 for the product measure $\tau_0 \times \mathcal{N}(0, I_d) \times \cdots \times \mathcal{N}(0, I_d)$, where τ_0 is a probability measure on $\mathbf{T}$ and J copies of $\mathcal{N}(0, I_d)$ are taken.

THEOREM 2.1. *Let $p > 1$ and assume that $G \in C^2(\mathbf{T}^d, \mathbf{R}^K)$. Then for any initial distribution τ_0, any $J > 0$ and any $T > 0$, there is $C = C(T, J)$ such that*

$$\forall (\delta, \sigma) \in \mathbf{R}_+ \times \mathbf{R}_+, \qquad \mathbf{E}\left(\sup_{0 \le t \le T} |\theta_t - \vartheta_t|^p \right) \le C(\delta^p + \sigma^{\beta p}). \tag{2.9}$$

Here $\{\vartheta_t\}_{t \in [0,T]}$ is the solution to (2.8) with initial condition $\vartheta_0 \sim \tau_0$, and $\{\theta_t, \xi^{(1)}, \dots, \xi^{(J)}\}_{t \in [0,T]}$ is the solution to (2.4) with initial condition μ_0. The exponent β is defined as follows:

$$\beta = \begin{cases} 1 & \text{if } G \in C^2(\mathbf{T}^d, \mathbf{R}^K), \\ 2 & \text{if } G \in C^3(\mathbf{T}^d, \mathbf{R}^K). \end{cases}$$

REMARK 2.4. *Since knowing the convergence rates with respect to δ and σ is helpful for the parametrization of the method in practice, we opted to explicitly consider both cases. The critical change from two to three*

derivatives occurs because three or more derivatives are required to exploit the mean zero property of third moments of Ξ. One might wonder whether an even higher regularity of G could lead to better convergence rates in the limit $\sigma \to 0$. An inspection of the proof of Theorem 2.1 reveals this is not the case.

In order to balance the two error terms on the right-hand side of (2.9), one may choose $\sigma \propto \delta$ when $G \in C^2(\mathbf{T}^d, \mathbf{R}^K)$, or just $\sigma \propto \sqrt{\delta}$ when $G \in C^3(\mathbf{T}^d, \mathbf{R}^K)$. Since a larger value of σ seems to favor exploration of the state space, as suggested by the numerical experiments in Subsection 3.1, choosing $\sigma \propto \sqrt{\delta}$ might indeed be advantageous for convergence when $G \in C^3(\mathbf{T}^d, \mathbf{R}^K)$.

Next, we present the counterpart of Theorem 2.1 for the numerical discretizations (2.6) and (2.7). We note that Theorem 2.2 employs a weaker metric than the one employed in (2.9).

THEOREM 2.2. *Assume that $G \in C^2(\mathbf{T}^d)$. Then, for any initial distribution τ_0 on $\mathbf{T}$, any $J > 0$ and any $T > 0$, there exists $C = C(T, J)$ such that*

$$\forall (\delta, \sigma, \Delta) \in \mathbf{R}_+^3, \qquad \sup_{0 \le n \le T/\Delta} \mathbf{E} \left| \hat{\theta}_n - \vartheta_{n\Delta} \right|^2 \le C \left(\Delta + \sigma^{2\beta} + \log(1 + \delta^{-1}) \, \delta^2 \right). \tag{2.10}$$

Here $\{\hat{\theta}_n, \hat{v}_n^{(1)}, \dots, \hat{v}_n^{(J)}\}_{n=0}^N$ is the solution to (2.6) with $D(\bullet) = I_d$ and initial condition μ_0, $\{\vartheta_t\}_{t \in [0,T]}$ is the solution to (2.8) with initial condition τ_0 and

$$\beta = \begin{cases} 1 & \text{if } G \in C^2(\mathbf{T}^d, \mathbf{R}^K), \\ 2 & \text{if } G \in C^3(\mathbf{T}^d, \mathbf{R}^K). \end{cases}$$

REMARK 2.5. *It is straightforward to obtain an error bound for (2.7) from this result. Indeed, denoting by $\tilde{\theta}_n$ the solution to (2.7) with initial condition τ_0 in order to differentiate it from the solution to (2.6), it holds by the triangle inequality and (2.10) that, for all $(\delta, \sigma, \Delta) \in \mathbf{R}_+^3$ and all $\varepsilon > 0$,*

$$\begin{aligned} \sup_{0 \le n \le N} \mathbf{E} \left| \tilde{\theta}_n - \vartheta_{n\Delta} \right|^2 \le \left(1 + \frac{1}{\varepsilon} \right) \sup_{0 \le n \le N} \mathbf{E} \left| \tilde{\theta}_n - \hat{\theta}_n \right|^2 \\ + (1 + \varepsilon) \, C \left(\Delta + \sigma^{2\beta} + \log(1 + \delta^{-1}) \, \delta^2 \right). \end{aligned}$$

Here we used that, by Young's inequality, it holds $(a+b)^2 \le (1+\varepsilon)a^2 + \left(1 + \frac{1}{\varepsilon}\right) b^2$ for any $\varepsilon > 0$ and any $a, b \in \mathbf{R}$. Therefore, taking the limit $\delta \to 0$, noting that the first term on the right-hand side vanishes in this limit (assuming that $\hat{\xi}_n^{(j)}$ in (2.7) coincides with $x_n^{(j)}$ in (2.6)), and then letting $\varepsilon \to 0$, we deduce

$$\forall (\sigma, \Delta) \in \mathbf{R}_+ \times \mathbf{R}_+, \qquad \sup_{0 \le n \le N} \mathbf{E} \left| \tilde{\theta}_n - \vartheta_{n\Delta} \right|^2 \le C \left(\Delta + \sigma^{2\beta} \right), \tag{2.11}$$

where $C = C(T, J)$ is the same constant as in (2.10). In Section 4, for clarity of exposition, we will in fact first prove the convergence estimate (2.11) before showing the more general Theorem 2.2.

REMARK 2.6. *In the limit $\Delta \to 0$, the error bound (2.10) becomes*

$$\sup_{0 \le n \le T/\Delta} \mathbf{E} \left| \hat{\theta}_n - \vartheta_{n\Delta} \right|^2 \le \left(\sigma^{2\beta} + \log(1 + \delta^{-1}) \, \delta^2 \right),$$

which is almost as sharp as the bound obtained in Theorem 2.1. The presence of the extra factor $\log(1+\delta^{-1})$ in front of δ^2 indicates that it may be possible to obtain a sharper bound.

3. Numerical Experiments. In this section, we present numerical experiments illustrating our method. Subsections 3.1 and 3.2 serve as proof of concept; in Subsection 3.1 a toy inverse problem with low-dimensional parameter and data is considered, and in Subsection 3.2 preconditioning is discussed and exemplified. A realistic and challenging example is then considered in Subsection 3.3.

3.1. Low-dimensional Parameter Space. We first consider the inverse problem with low-dimensional parameter space that was first presented in [18] and later employed as a test problem in [30, 20]. In this problem, the forward model maps the unknown $(u_1, u_2) \in \mathbf{R}^2$ to the observation $(p(x_1), p(x_2)) \in \mathbf{R}^2$, where $x_1 = 0.25$ and $x_2 = 0.75$ and where $p(x)$ denotes the solution to the boundary value problem

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\mathrm{e}^{u_1}\frac{\mathrm{d}p}{\mathrm{d}x}\right) = 1, \qquad x \in [0,1], \tag{3.1}$$

with boundary conditions $p(0) = 0$ and $p(1) = u_2$. This problem admits the following explicit solution [30]:

$$p(x) = u_2 x + \mathrm{e}^{-u_1}\left(-\frac{x^2}{2} + \frac{x}{2}\right).$$

We employ the same parameters as in [20]: the prior distribution is $\mathcal{N}(0, \sigma^2 I_2)$ with $\sigma = 10$, and the noise distribution is $\mathcal{N}(0, \gamma^2 I_2)$ with $\gamma = 0.1$. The observed data is taken to be $y = (27.5, 79.7)$. Since (3.1) admits an explicit solution, the forward model can be evaluated very quickly. As a result, obtaining a good approximation of the MAP with our multiscale method takes less than a minute on a personal computer.

We first investigate the performance of the algorithm (2.6) when $\nu = 0$, i.e. when an approximation of the MAP estimator is sought. All the numerical results related to this problem were obtained with $J = 8$ auxiliary processes, with a fixed time step $\Delta = 10^{-3}$, and with $\hat{\theta}_0 = (1, 103)$ as the initial condition.

The effect of the parameter σ, which encodes the radius of exploration around θ, is illustrated in Figure 1. In the left panel, we present the trajectories of the solution $\hat{\theta}_n$ obtained with (2.6) for fixed small $\delta = 10^{-7}$ and several values of σ. In the right panel, we present the evolution of the error, in the Euclidean norm, along the trajectories. In contrast with deterministic algorithms, the iterates produced by our method do not converge to a limit, which is reflected in the fact that the error oscillates indefinitely at a small value as the simulation progresses. The reason for this is that there does not exist a value of θ for which the right-hand side of (2.1a) is zero for all $\Xi \in (\mathbf{R}^d)^J$. We also notice that a larger value of σ seems to increase the convergence speed in the initial stage of the simulation but it leads to a larger error in the later stages, as the iterates get close to the MAP estimator.

The effect of the parameter δ, which influences the correlation between the directions of successive steps, is illustrated in Figure 2 for fixed $\sigma = 0.1$. We observe that the direction of successive steps seems to oscillate more rapidly when δ is small, which is consistent with our understanding of the effect of this parameter. In this particular example, choosing a large δ does not appear to improve convergence.

Let us now investigate the efficiency of (2.6) for sampling from the posterior distribution. For this simulation, we used the parameters $\delta = 10^{-4}$ and $\sigma = 0.01$. We ran the simulation for 20,000 iterations and, discarding the first 1,000 iterates, we computed an approximation of the posterior by kernel density estimation with the function `gaussian_kde` from the `scipy.stats` module. The iterates 1,000 to 20,000, the approximation of the Bayesian posterior based on these iterates, and the true posterior are depicted in the left, middle and right panels of Figure 3, respectively. It appears from the figure that the approximate posterior is close to the true posterior. Indeed, the mean and covariance of the true and approximate posterior

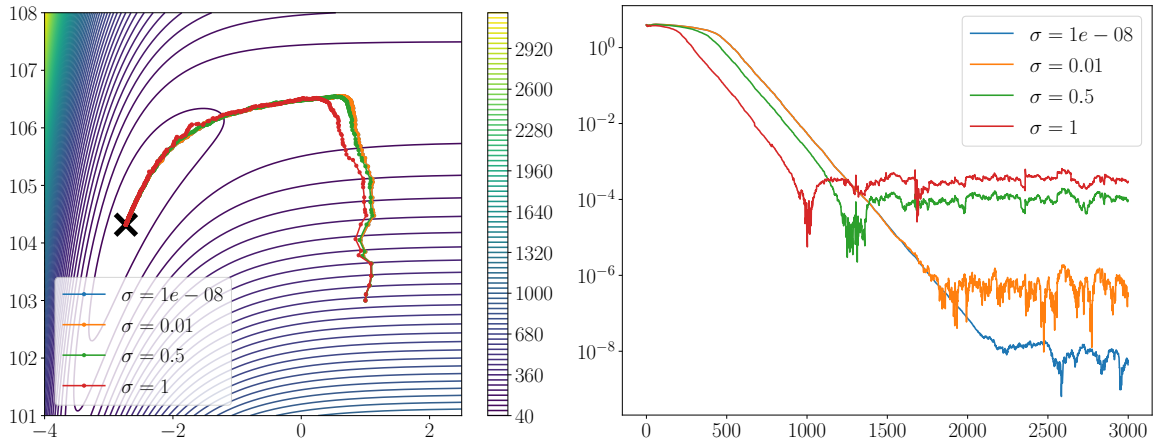

FIG. 1. **Left:** *Trajectories of the numerical solution* $\hat{\theta}_n$ *for fixed* $\delta = 10^{-8}$ *and different values of* σ. **Right:** *Error* $|\hat{\theta}_n - \theta_{\mathrm{MAP}}|_2$ *along the trajectories, where* θ_{MAP} *is the MAP estimator.*

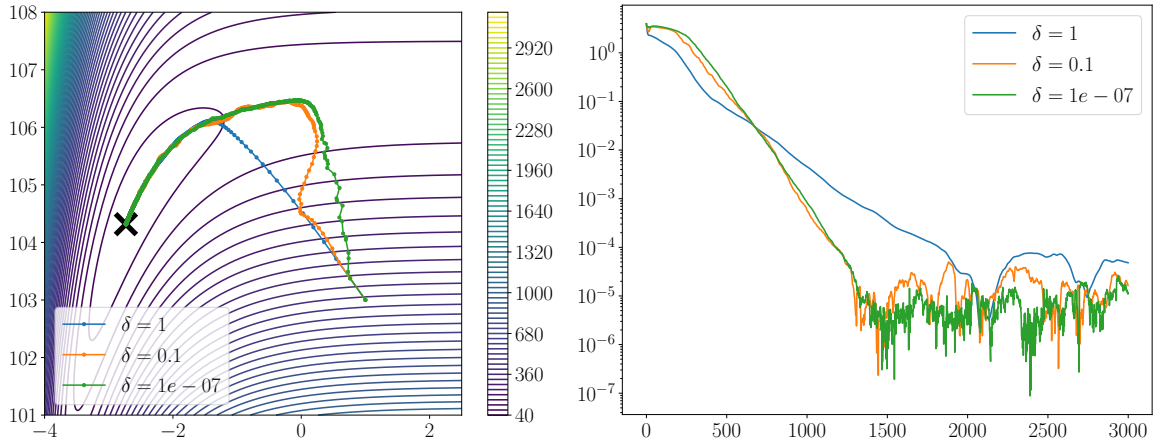

FIG. 2. **Left:** *Trajectories of the numerical solution* $\hat{\theta}_n$ *for fixed* $\sigma = 0.1$ *and different values of* δ. **Right:** *Error* $|\hat{\theta}_n - \theta_{\mathrm{MAP}}|_2$ *along the trajectories, where* θ_{MAP} *is the MAP estimator.*

distributions, given, respectively, by

$$m_p = \begin{pmatrix} -2.714... \\ 104.346... \end{pmatrix} \quad C_p = \begin{pmatrix} 0.0129... & 0.0288... \\ 0.0288... & 0.0808... \end{pmatrix},$$

and

$$\tilde{m}_p = \begin{pmatrix} -2.686... \\ 104.411... \end{pmatrix} \quad \tilde{C}_p = \begin{pmatrix} 0.0147... & 0.0308... \\ 0.0308... & 0.0866... \end{pmatrix}$$

are fairly close.

3.2. Accelerating Convergence with Preconditioning. In many applications, the condition number of the Hessian of Φ_R at and around the MAP estimator is very large. In this situation, the fastest time scale of (2.8), i.e. of gradient descent ($\nu = 0$) or overdamped Langevin ($\nu = 1$) dynamics, is much smaller than its slowest time scale. This is evident when Φ_R is quadratic, in which case the slowest and fastest time scales correspond to the reciprocals of the smallest and largest eigenvalues of $\mathrm{D}^2\Phi_R$, respectively. As a result of this wide scale separation, a very small time step, relatively to the slowest time scale which is limiting for

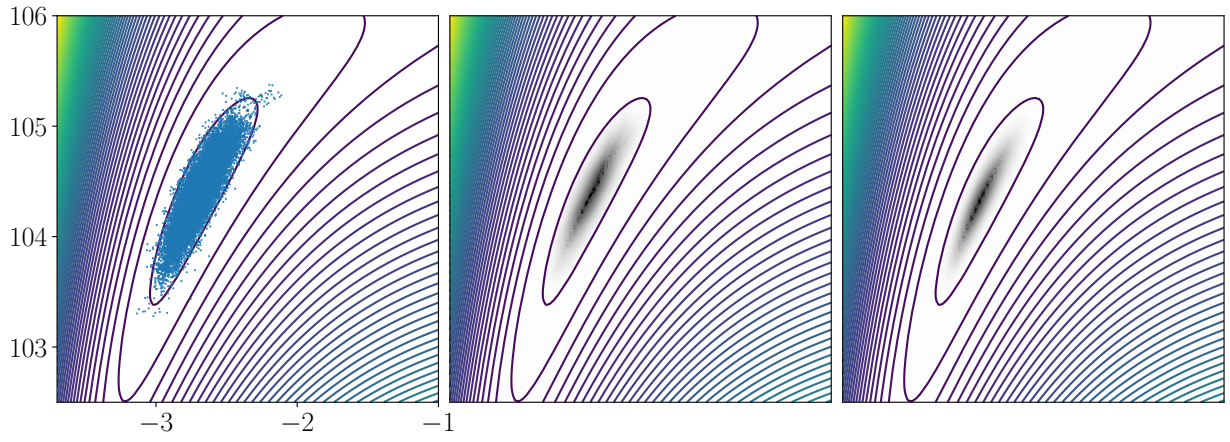

FIG. 3. **Left:** *Iterates 1,000 to 20,000.* **Middle:** *Approximation of the Bayesian posterior based on these iterates, using kernel density estimation.* **Right:** *True Bayesian posterior.*

convergence, is required in order to resolve the dynamics precisely using a numerical method. For explicit numerical methods, a wide separation of time scales also leads to a stringent constraints (with respect to the time scale of convergence) on the time step in order to guarantee stability, leading to often prohibitive computational costs.

Empirically, we observed that our multiscale method suffers from a similar issue, which is not surprising given that (2.4) converges to (2.8) as $(\delta, \sigma) \to (0, 0)$. This is in contrast with the sampling and inversion methods for inverse problems that are based on the ensemble Kalman filter, essentially because these methods are affine-invariant [21]: they behave similarly across the class of problems that differ only by an affine transformation. Ensemble Kalman methods can be viewed, at least in the case of a linear forward model, as coupled gradient descents dynamics or overdamped Langevin diffusions preconditioned by the covariance of the ensemble, providing good stability and convergence properties [20, 10].

To remedy this issue of overly restrictive constraints on the time step (relatively to the slowest time scales of the problem), preconditioning can be incorporated in our multiscale method. More precisely, given a symmetric positive definite matrix K, the dynamics (2.1) (resp. (2.4)) can be modified as follows,

$$\dot\theta = -\frac{1}{J\sigma^2}\sum_{j=1}^{J}\langle G(\theta^{(j)}) - G(\theta), G(\theta) - y\rangle_\Gamma\,(\theta^{(j)} - \theta) - C_K(\Xi)\Sigma^{-1}(\theta - m) + \nu\sqrt{2D_K(\Xi)}\,\dot w, \tag{3.2a}$$

$$\theta^{(j)} = \theta + \sigma\sqrt{K}\xi^{(j)}, \qquad j = 1, \dots, J, \tag{3.2b}$$

$$\dot\xi^{(j)} = -\frac{1}{\delta^2}\,\xi^{(j)} + \sqrt{\frac{2}{\delta^2}}\,\dot w^{(j)}, \qquad \xi^{(j)}(0) \sim \mathcal{N}(0, I_d), \qquad j = 1, \dots, J, \tag{3.2c}$$

where $C_K(\Xi) := \sqrt{K}\,C(\Xi)\sqrt{K}$ and $D_K(\Xi) = C_K(\Xi)$ (resp. $D_K(\Xi) = K$). Under the linear approximation

$$G(\theta^{(j)}) - G(\theta) \approx (\theta^{(j)} - \theta)\cdot\nabla G(\theta),$$

which is accurate for small σ, we can rewrite (3.2a) as

$$\dot\theta \approx -C_K(\Xi)\nabla\Phi_R(\theta) + \nu\sqrt{2D_K(\Xi)}\dot w,$$

which suggests that $\{\theta_t\}_{t\geq 0}$ should converge, in the limit as $\delta \to 0$ and $\sigma \to 0$, to the preconditioned overdamped Langevin dynamics (1.4). In order to make this more precise, notice that if the stochastic process $\{\theta_t, \Xi_t\}_{t\geq 0}$ solves (3.2), then $\{u_t, \Xi_t\}_{t\geq 0} := \{\sqrt{K^{-1}}\theta_t, \Xi_t\}_{t\geq 0}$ is equal in law to the solution of (2.1) (resp. (2.4)) with the modified forward model

$$\tilde{G}(u) = G(\sqrt{K}\, u), \qquad u \in \mathbf{R}^d,$$

with the modified initial condition $u_0 = \sqrt{K^{-1}}\theta_0$, and with the modified prior parameters $\tilde{m} = \sqrt{K^{-1}}m$ and $\tilde{\Sigma} = \sqrt{K^{-1}}\Sigma\sqrt{K^{-1}}$, i.e. with the prior distribution $\mathcal{N}(\sqrt{K^{-1}}m, \sqrt{K^{-1}}\Sigma\sqrt{K^{-1}})$. Here we employed the fact that $\sqrt{C_K(\Xi)}w_t = \sqrt{K}\sqrt{C(\Xi)}w_t$ in law. In view of this connection, Theorems 2.1 and 2.2 apply *mutatis mutandis* to the dynamics (3.2) with $D_K(\Xi) = K$.

When the forward model is linear, a good preconditioning matrix K is given by the covariance of the Bayesian posterior, as motivated in the example below. In practice, we observed that preconditioning with the posterior covariance works well also for nonlinear forward models. We emphasize that this approach to preconditioning can be applied both with ($\nu = 1$) and without ($\nu = 0$) noise. In order to approximate the posterior covariance at a reasonable computational cost, we employ the Ensemble Kalman Sampler (EKS) proposed in [20], a method based on interacting Langevin diffusions that enables generating approximate samples from the Bayesian posterior. The EKS is based on the dynamics

$$\dot{\theta}^{(\ell)} = -\frac{1}{L}\sum_{\ell=1}^{L} d_{\ell k}(\Theta)(\theta^{(k)} - \bar{\theta}) - C(\Theta)\Sigma^{-1}(\theta^{(\ell)} - m) + \sqrt{2C(\Theta)}\,\dot{w}^{(\ell)}, \qquad \ell = 1, \dots, L, \tag{3.3a}$$

$$d_{\ell k}(\Theta) = \langle G(\theta^{(k)}) - \bar{G}, G(\theta^{(\ell)}) - y\rangle_\Gamma, \tag{3.3b}$$

where $\Theta = (\theta^{(1)}, \dots, \theta^{(L)})$ and

$$\bar{\theta} = \frac{1}{L}\sum_{\ell=1}^{L}\theta^{(\ell)}, \qquad \bar{G} = \frac{1}{L}\sum_{\ell=1}^{L} G(\theta^{(\ell)}).$$

The processes $w^{(\ell)}$ are independent Brownian motions and $C(\Theta)$ is the covariance of the empirical distribution associated with the particles:

$$C(\Theta) = \frac{1}{L}\sum_{\ell=1}^{L}(\theta^{(\ell)} - \bar{\theta}) \otimes (\theta^{(\ell)} - \bar{\theta}).$$

The first argument in the inner product on the right-hand side of (3.3b) is related to consensus, whereas the second argument measures the mismatch with the observed data. In the limit $t \to \infty$, the empirical distribution $\frac{1}{L}\sum_{\ell=1}^{L}\delta_{\theta^{(\ell)}}$ associated with the particles provides an approximation of the Bayesian posterior. The matrix K could also be learned from the UKS or CBS approaches to approximate sampling, rather than from the EKS.

We now illustrate this preconditioning methodology for a simple inverse problem where the forward model is given by the linear function $G : \mathbf{R}^3 \ni \theta \mapsto (\theta_1, k\theta_2, k^2\theta_3)$ with $k = 5$. We choose the other parameters of the inverse problem as follows: $y = (1, k, k^2)$, $\Sigma = +\infty I_3$ and $\Gamma = I_3$, so that the MAP estimator is given

explicitly by $\theta_{\mathrm{MAP}} = (1, 1, 1)$. For these parameters, the least-squares functional is given by

$$\Phi(\theta; y) = \frac{1}{2}\left(|\theta_1 - y_1|^2 + k^2|\theta_2 - y_2|^2 + k^4|\theta_3 - y_3|^2\right). \tag{3.4}$$

The largest eigenvalue of the Hessian of $\Phi(\bullet; y)$ at θ_{MAP} is equal to k^4 so, if we were to use a gradient descent for (3.4) with the explicit Euler method in order to find the minimizer of $\phi(\bullet; y)$, then the constraint on the time step Δ in order to ensure stability would be that $k^4\Delta < 2$. Since our method converges to a gradient descent in the limit as $\delta \to 0$ and $\sigma \to \infty$, it is reasonable to expect that a similar constraint should hold to ensure stability of (2.6), and this is indeed what we observed numerically; in particular, we verified in the case where $\sigma = \delta = 10^{-5}$ that (2.6) is stable when $\Delta = 1/k^4$ but unstable when $\Delta = 3/k^4$.

For this problem, the covariance of the Bayesian posterior is given by

$$K = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{k^2} & 0 \\ 0 & 0 & \frac{1}{k^4} \end{pmatrix}.$$

This is clearly the optimal preconditioner for calculating the MAP estimator. Indeed, in this case, the limiting equation associated with (3.2) in the limit as $\max(\delta, \sigma) \to 0$ is

$$\mathrm{d}\theta_t = -(\theta_t - \theta_{\mathrm{MAP}})\,\mathrm{d}t + \nu\sqrt{2K}\,\mathrm{d}w_t. \tag{3.5}$$

When employed for integrating this equation, a gradient descent implemented with explicit Euler when $\nu = 0$ would be stable for time steps $\Delta < 2$. In order to calculate the preconditioning parameter K, we integrate the dynamics (3.3a) and (3.3b) using the Euler–Maruyama method over a time that is sufficiently long to provide a rough approximation of the Bayesian posterior. After 1,000 iterations of the EKS, with 256 particles initialized independently as $\mathcal{N}(0, I_3)$ and with an adaptive timestep computed as in [40], the covariance $C(\Theta)$ of the ensembles from EKS is close to the optimal preconditioning parameter:

$$C(\Theta) = \begin{pmatrix} 0.989... & 0.00102... & -0.00212... \\ 0.00102... & 0.198... & 0.00131... \\ -0.00212... & 0.00131... & 0.0395... \end{pmatrix}.$$

The effect of using this preconditioner is illustrated in Figure 4 in the case where (2.6) is used in optimization mode, i.e. with $\nu = 0$. The left and right panels present the evolution of the error with and without preconditioning. The parameters employed are $\sigma = \delta = 10^{-5}$ and $J = 8$, and the time step was set to $\Delta = 1$ with preconditioning and $\Delta = 1/k^4$ without preconditioning. The initial condition taken in both cases was $\theta_0 = (0, 0, 0)^T$. It appears clearly from the figure that preconditioning significantly accelerates the convergence.

In practice, the mean of the ensembles obtained after application of the EKS is also useful; it can be employed as initial condition for (3.2). This is the approach taken in the next section.

3.3. High-dimensional Parameter Space. We now present an example from [20] in which the calculation of the forward map requires the solution to a partial differential equation (PDE) in two dimensions, and is therefore computationally expensive. More precisely, we consider the inverse problem of finding the permeability from noisy pressure measurements in a Darcy flow. This problem falls into the framework developed in [14], and it is natural to model it as an inverse problem with infinite-dimensional parameter

space. In order to be amenable to the numerical methods developed in this paper, however, the problem needs to be discretized: this requires defining a finite-dimensional approximation space for the unknown parameter and specifying a numerical approximation for the calculation of the forward map. We begin by presenting the inverse problem in its natural infinite-dimensional setting, and then we give the associated discrete approximation, which we solve numerically using (2.6) together with the preconditioning approach proposed in Subsection 3.2.

At the infinite-dimensional level, the abstract inverse problem we consider is that of estimating the logarithm of the permeability profile, denoted by $a(x)$, based on noisy measurements of the solution $p(x)$ to the PDE

$$-\nabla \cdot \left(\mathrm{e}^{a(x)} \nabla p(x) \right) = f(x), \qquad x \in D, \tag{3.6a}$$

$$p(x) = 0, \qquad x \in \partial D. \tag{3.6b}$$

Here $D = [0,1]^2$ is the domain and $f(x) = 50$ represents a source of fluid. For the prior distribution, we employ a Gaussian measure on $L^2(D)$ with mean zero and precision (inverse covariance) operator given by

$$\mathcal{C}^{-1} = (-\Delta + \tau^2 \mathcal{I})^{\alpha},$$

equipped with Neumann boundary conditions on the space of mean-zero functions. The eigenfunctions and eigenvalues associated with the covariance operator are given by

$$\psi_\ell(x) = \cos\big(\pi(\ell_1 x_1 + \ell_2 x_2)\big), \qquad \lambda_\ell = \left(\pi^2 |\ell|^2 + \tau^2\right)^{-\alpha}, \qquad \ell \in \mathbf{N}^2.$$

The parameters τ and α control the characteristic length scale and the smoothness of samples drawn from the prior, respectively. For the numerical experiments presented in this section, we take the same values for these parameters as in [20]: $\tau = 3$ and $\alpha = 2$. In this setting, it can be shown by reasoning as in [14, Example 2.19] that the log-permeability $a(x)$ is almost surely continuous on the closed set D, so there exists a unique solution $p \in H^1(D)$ to (3.6) almost surely.

If $a(x) \sim \mathcal{N}(0, \mathcal{C})$, then $\mathbf{E}\big((a, \psi_\ell)\,(a, \psi_m)\big) = \lambda_\ell \, \delta_{\ell,m}$ by definition of the covariance operator and by orthonormality of the eigenfunctions $\{\psi_\ell\}_{\ell \in \mathbf{N}^2}$, where $(\bullet, \bullet)$ denotes the inner product in $L^2(D)$. Since $a(x)$

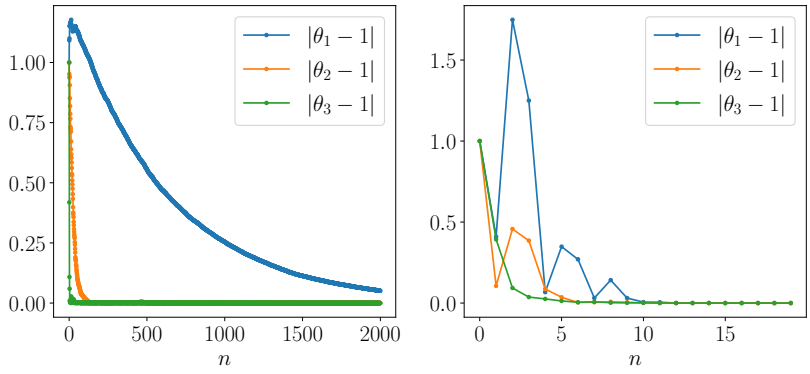

FIG. 4. *Error between the iterates and the MAP estimator, without (left) and with (right) preconditioning. Here the label n of the x axis denotes the iteration index. We observe that a better approximation of the MAP estimator is obtained after 20 iterations with preconditioning, rather than after 2,000 iterations without preconditioning.*

is almost surely in $L^2(D)$, we deduce that, almost surely,

$$a = \sum_{\ell \in \mathbf{N}^2} (a, \psi_\ell)\, \psi_\ell =: \sum_{\ell \in \mathbf{N}^2} \sqrt{\lambda_\ell}\, \theta_\ell\, \psi_\ell, \tag{3.7}$$

where the factors $\{\theta_\ell\}_{\ell \in \mathbf{N}^2}$ are independent $\mathcal{N}(0,1)$ random variables. This is the Karhunen–Loève (KL) expansion, which can be used as a starting point for the definition of probability distributions in infinite dimensions; see e.g. [14] for more details.

In theoretical works on Bayesian inverse problems of the type considered in this section, the data are usually modeled as the values, perturbed by Gaussian noise, taken at the solution p to (3.6) by a finite number of continuous linear functionals $\ell_1, \ldots, \ell_K$ over $H_0^1(D)$. That is, the forward model maps the unknown permeability $a(\bullet)$ to

$$\Big(\ell_1(p), \ldots, \ell_K(p)\Big) \in \mathbf{R}^K.$$

In practice, however, we consider that the data consist of pointwise measurements of the solution to (3.6), up to noise. We assume that these are taken at a finite number equidistant points given by

$$x_{ij} = \left(\frac{i}{M}, \frac{j}{M}\right), \qquad 1 \le i, j \le M-1. \tag{3.8}$$

Since pointwise evaluation is not a continuous functional on $H_0^1(D)$, our example deviates here from the framework in [14]. As mentioned in [20], pointwise evaluation could in principle be approximated by integration against a narrow mollifier, which would ensure continuity of the functionals, but we do not discuss this here. We take the distance between measurement points equal to $1/10$, i.e. $M = 10$, and we work with the noise distribution $\mathcal{N}(0, \gamma^2 I_K)$, with $\gamma = 0.01$ and $K = (M-1)^2$.

In order to approximate the solution to the inverse problem numerically, we truncate the KL series (3.7) after a finite number of terms and take this truncated series as the object of inference. More precisely, we take as unknown the vector of parameters $\theta = \{\theta_\ell : |\ell|_\infty \le N\} \in \mathbf{R}^{(N+1)^2}$, and as prior distribution the Gaussian $\mathcal{N}(0, I_d)$, with $d = (N+1)^2$. For each $\theta \in \mathbf{R}^{(N+1)^2}$, a log-permeability field is constructed by summation as $a(\bullet; \theta) := \sum_{|\ell|_\infty \le N} \sqrt{\lambda_\ell}\, \theta_\ell\, \psi_\ell$, and the corresponding solution to (3.6) is approximated with a finite element method (FEM). This defines an inverse problem with finite-dimensional parameter space, which is amenable to (2.6) or (2.7). In practice, we use $N = 2$ and a FEM with quadratic elements over a regular mesh with 80 elements per direction. Below, we refer to this inverse problem as the *finite-dimensional inverse problem*, in order to distinguish it from the inverse problem with infinite-dimensional parameter space it aims to discretize.

In order to generate data for the finite-dimensional inverse problem, we employ the same FEM as is employed in the numerical algorithm for the evaluation of the forward model, with a true permeability also given by a truncated KL expansion. We consider below two possibilities:

Simulation without model misspecification. Here, we generate the true permeability using as many KL terms as in the numerical inference, i.e.

$$a^\dagger(x) = \sum_{|\ell|_\infty \le N} \sqrt{\lambda_\ell} \theta_\ell^\dagger \psi_\ell, \qquad \theta_\ell^\dagger \sim \mathcal{N}(0,1). \tag{3.9}$$

Clearly, this does not provide a sample from $\mathcal{N}(0, \mathcal{C})$, but it does provide a sample consistent with the prior distribution assumed in the finite-dimensional inverse problem. The logarithm of the true permeability field,

as well as its MAP approximation by (2.6) with $\nu = 0$, is illustrated in Figure 5. The preconditioning matrix K was calculated using the methodology outlined in Subsection 3.2, by running 100 iterations of the EKS with an ensemble size equal to 256 and with the adaptive time-stepping scheme of [40]. The parameters used in (2.6) are $\delta = \sigma = 10^{-5}$ and $J = 8$, and we run 100 iterations of the algorithm with a fixed time step equal to 0.1. It appears from the figure that the approximation of the MAP estimator provided by the multiscale method leads to a permeability field close to the truth; we calculate that, denoting by θ^{MAP} the MAP estimator,

$$\frac{\sqrt{\sum_{|\ell|_\infty \le N} \left|\theta_\ell^{\mathrm{MAP}} - \theta_\ell^\dagger\right|^2}}{\sqrt{\sum_{|\ell|_\infty \le N} \left|\theta_\ell^\dagger\right|^2}} = 0.184... \tag{3.10a}$$

and

$$\frac{\left\|a^\dagger - a^{\mathrm{MAP}}\right\|_{L^2(D)}}{\left\|a^\dagger\right\|_{L^2(D)}} = \frac{\sqrt{\sum_{|\ell|_\infty \le N} \lambda_\ell \left|\theta_\ell^{\mathrm{MAP}} - \theta_\ell^\dagger\right|^2}}{\sqrt{\sum_{|\ell|_\infty \le N} \lambda_\ell \left|\theta_\ell^\dagger\right|^2}} = 0.085... \tag{3.10b}$$

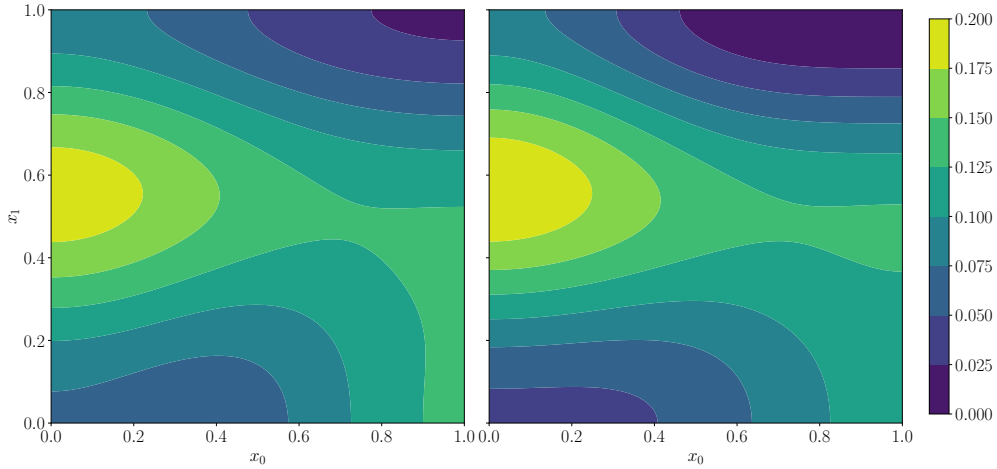

FIG. 5. *Logarithms of true (left) and approximate permeability profiles (right), without model misspecification. The approximate permeability profile was constructed from the approximation of the MAP estimator provided by (2.6) with $\nu = 0$.*

We now turn our attention to the problem of sampling from the Bayesian posterior. The ensembles obtained with 1,000 iterations of the multiscale method (2.6), with $\nu = 1$ and all other parameters unchanged, are illustrated in Figure 8. The agreement between the true parameter and the posterior samples is good: overall, the marginals of the approximate posterior, which were obtained by kernel estimation, are close to the true values of the KL coefficients.

Simulation with model misspecification. Next, we generate the true permeability with twice as many KL terms per dimension for the truth as in the numerical algorithm, i.e. we take the true permeability to be

$$a^\dagger(x) = \sum_{|\ell|_\infty \le 2N+1} \sqrt{\lambda_\ell}\theta_\ell^\dagger \psi_\ell, \qquad \theta_\ell^\dagger \sim \mathcal{N}(0,1). \tag{3.11}$$

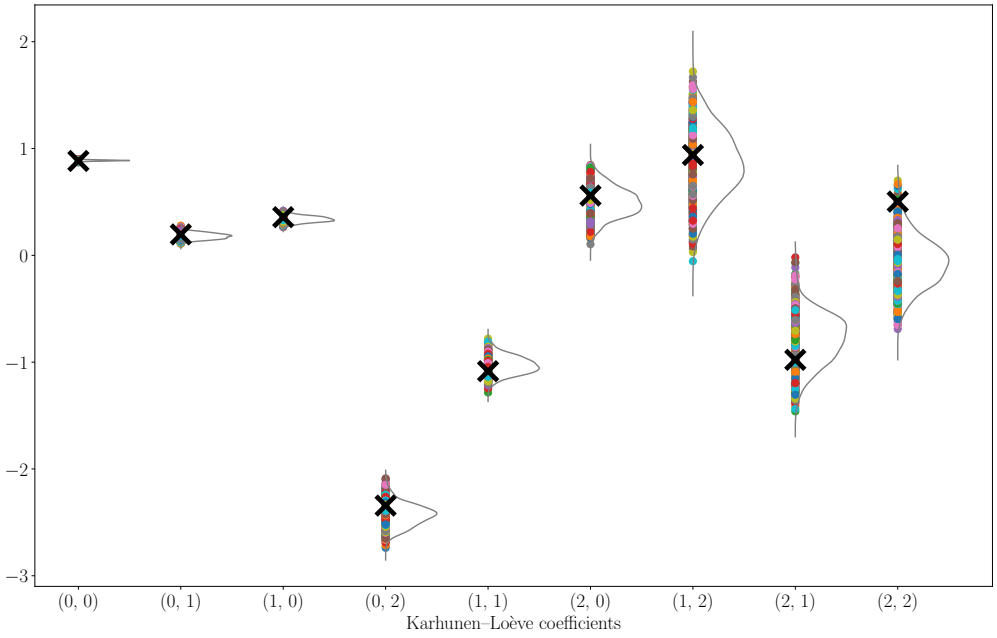

FIG. 6. *Approximate posterior samples produced by (2.6) with $\nu = 1$, without model misspecification. From these posterior samples, the marginal distributions of the KL coefficients were approximated by kernel density estimation using Gaussian kernels; they are depicted (non-normalized) in solid lines. The crosses are the true values of the coefficients, i.e. the values employed to generate the data.*

This does not provide a sample from $\mathcal{N}(0, \mathcal{C})$ either, but keeping more terms in the truth than in the prior distribution of the finite-dimensional inverse problem enables to understand the effect of model misspecification. In particular, even in the absence of noise on the observations, we expect that the MAP estimator associated with the finite-dimensional inverse problem for data generated using (3.11) would differ from the true value of the corresponding $(N+1)^2$ coefficients in $\theta^\dagger$.

In this case, the logarithm of the true permeability field and its MAP approximation, obtained by (2.6) with the same parameters as those employed for Figure 5, are illustrated in Figure 7. The error between the MAP estimator provided by the multiscale method and the true permeability is higher that in (3.10a) and (3.10b): counting only the first $(N+1)^2$ KL modes, we have

$$\frac{\sqrt{\sum_{|\ell|_\infty \le N} \left|\theta_\ell^{\mathrm{MAP}} - \theta_\ell^\dagger\right|^2}}{\sqrt{\sum_{|\ell|_\infty \le N} \left|\theta_\ell^\dagger\right|^2}} = 0.204... \quad \text{and} \quad = \frac{\sqrt{\sum_{|\ell|_\infty \le N} \lambda_\ell \left|\theta_\ell^{\mathrm{MAP}} - \theta_\ell^\dagger\right|^2}}{\sqrt{\sum_{|\ell|_\infty \le N} \lambda_\ell \left|\theta_\ell^\dagger\right|^2}} = 0.121... \tag{3.12}$$

We now present the counterpart of Figure 6 when the data is generated from the permeability (3.11) instead of (3.9); see Figure 8. Overall, the marginals of the approximate posterior are quite close to the true values of the KL coefficients. On close inspection, however, we notice that the true value of the coefficient associated with $\ell = (0, 0)$ is more than four standard deviations away from the mean of the associated approximate marginal. This discrepancy, which is surprising at first, originates in large part from the fact that the method is based on a prior with fewer KL coefficients than were used to generate the data. We note, in particular, that the agreement is not as good as in Figure 6.

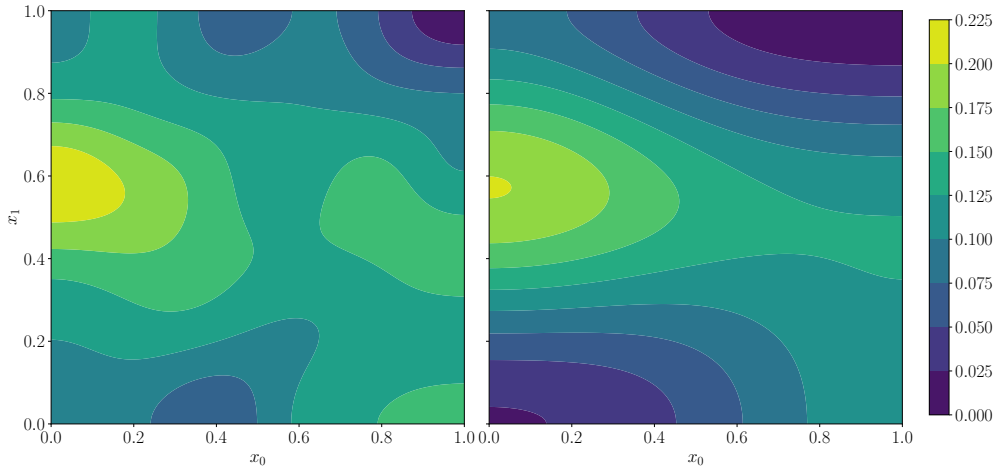

FIG. 7. *Logarithms of true (left) and approximate permeability profiles (right), with model misspecification.*

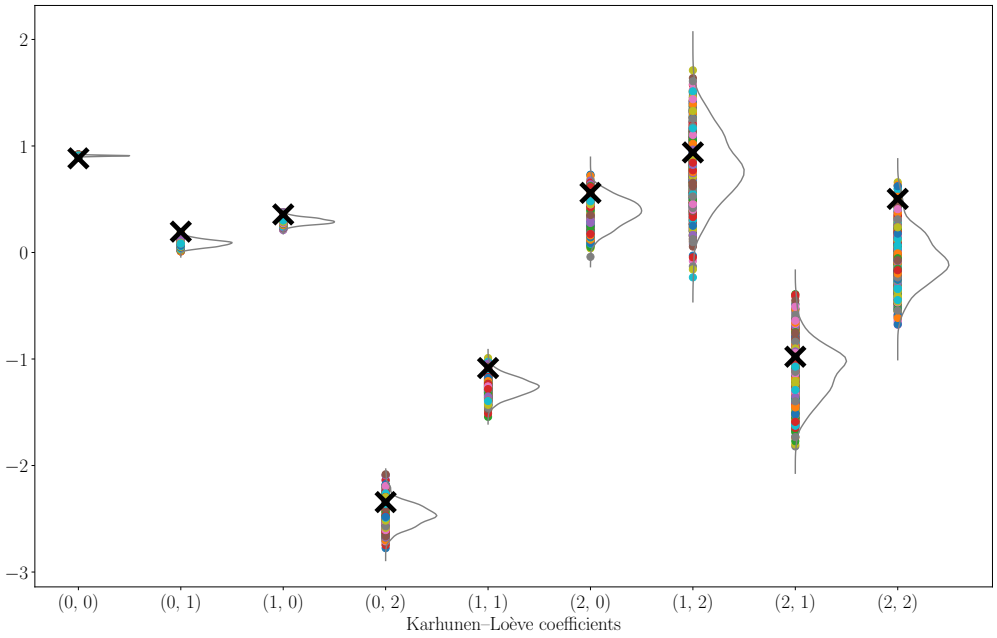

FIG. 8. *Approximate posterior samples produced by* (2.6) *with* $\nu = 1$*, with model misspecification.*

4. Proof of the Main Results. Throughout this section, we consider that the number of particles J is a fixed parameter. We often denote the drift in (2.1a) by

$$F_\sigma(\theta, \Xi) = -\frac{1}{J\sigma} \sum_{j=1}^{J} \langle G(\theta + \sigma\xi^{(j)}) - G(\theta), G(\theta) - y \rangle_\Gamma \, \xi^{(j)}.$$

We recall that we are working on the multi-dimensional torus $\mathbf{T}^d$ with a uniform prior, so $\Phi_R = \Phi$. If the forward model satisfies $G \in C^3(\mathbf{T}^d)$ then, by Taylor's formula, it holds for all $(\theta, \xi) \in \mathbf{T}^d \times \mathbf{R}^d$ that

$$
\begin{aligned}
-\frac{1}{\sigma}\langle G(\theta+\sigma\xi) - G(\theta), G(\theta) - y\rangle_\Gamma \, \xi = &- (\xi \otimes \xi)\nabla\Phi_R(\theta) - \frac{\sigma}{2}\langle (\xi\otimes\xi) : \mathrm{D}^2 G(\theta), G(\theta) - y\rangle_\Gamma \xi \\
&- \frac{\sigma^2}{6}\langle (\xi\otimes\xi\otimes\xi) \vdots \mathrm{D}^3 G(\theta + \alpha\sigma\xi), G(\theta) - y\rangle_\Gamma \xi,
\end{aligned} \tag{4.1}
$$

for some $\alpha \in [0,1]$ depending on θ, on ξ and on σ. Here $\xi \otimes \xi \in \mathbf{R}^{d\times d}$ and $\xi\otimes\xi\otimes\xi \in \mathbf{R}^{d\times d\times d}$ denote the matrix and third-order tensor with components $(\xi\otimes\xi)_{ij} = \xi_i\xi_j$ and $(\xi\otimes\xi\otimes\xi)_{ijk} = \xi_i\xi_j\xi_k$, respectively. We used the notation $\mathrm{D}^2 G(\theta) \in \mathbf{R}^{d\times d}$ for the Hessian of G at θ, and the notation $\mathrm{D}^3 G(\theta) \in \mathbf{R}^{d\times d\times d}$ for the tensor of third derivatives of G at θ, i.e. the third-order tensor with components $(\mathrm{D}^3 G)_{ijk}(\theta) = \partial_i\partial_j\partial_k G(\theta)$. The symbol : denotes the Frobenius inner product on $\mathbf{R}^{d\times d}$, and the symbol $\vdots$ denotes the $\mathbf{R}^{d\times d\times d}$ inner product defined by $S \vdots T = \sum_{i=1}^d \sum_{j=1}^d \sum_{k=1}^d S_{ijk}T_{ijk}$, for $S, T \in \mathbf{R}^{d\times d\times d}$.

Equation (4.1) motivates the following notation:

$$
\begin{aligned}
F_0(\theta,\Xi) &= -C(\Xi)\nabla\Phi_R(\theta), \\
F_1(\theta,\Xi) &= -\frac{1}{J}\sum_{j=1}^J \frac{1}{2}\langle (\xi\otimes\xi) : \mathrm{D}^2 G(\theta), G(\theta) - y\rangle_\Gamma \xi, \\
F_2(\theta,\Xi) &= \frac{1}{\sigma^2}\big(F_\sigma(\theta,\Xi) - F_0(\theta,\Xi) - \sigma F_1(\theta,\Xi)\big).
\end{aligned}
$$

With this notation, the multiscale system (2.4) can be rewritten as

$$
\begin{aligned}
\dot\theta &= F_0(\theta,\Xi) + \sigma F_1(\theta,\Xi) + \sigma^2 F_2(\theta,\Xi) + \nu\sqrt{2}\dot w, \\
\dot\xi^{(j)} &= -\frac{1}{\delta^2}\,\xi^{(j)} + \sqrt{\frac{2}{\delta^2}}\,\dot w^{(j)}, \qquad j = 1,\dots,J.
\end{aligned}
$$

If $G \in C^2(\mathbf{T}^d)$, then by Taylor's theorem

$$
\forall (\theta,\Xi) \in \mathbf{T}^d \times (\mathbf{R}^d)^J, \qquad |F_\sigma(\theta,\Xi) - F_0(\theta,\Xi)| \le C\sigma \sum_{j=1}^J |\xi^{(j)}|^3, \tag{4.2}
$$

for a constant C independent of σ. Likewise, if $G \in C^3(\mathbf{T}^d)$, then by (4.1) it holds

$$
\forall (\theta,\Xi) \in \mathbf{T}^d \times (\mathbf{R}^d)^J, \qquad |F_\sigma(\theta,\Xi) - F_0(\theta,\Xi) - \sigma F_1(\theta,\Xi)| \le C\sigma^2 \sum_{j=1}^J |\xi^{(j)}|^4, \tag{4.3}
$$

for a possible different constant C, also independent of σ. We divide the proof of Theorem 2.1 in two parts. In the first part, we assume that the forward map satisfies only $G \in C^2(\mathbf{T}^d, \mathbf{R}^K)$, and in the second part we obtain a refined estimate for when $G \in C^3(\mathbf{T}^d, \mathbf{R}^K)$.

Proof of Theorem 2.1 when $G \in C^2(\mathbf{T}^d, \mathbf{R}^k)$. Our approach for this proof is based on [59, Chapter 17]. Throughout the proof, C denotes a constant independent of δ and σ that is allowed to change from occurrence to occurrence. The generator of the dynamics associated to (2.4) is given by

$$
\mathcal{L} = \frac{1}{\delta^2}\mathcal{L}_0 + \mathcal{L}_1,
$$

where

$$\mathcal{L}_0 = \sum_{j=1}^{J} -\xi^{(j)} \cdot \nabla_{\xi^{(j)}} + \Delta_{\xi^{(j)}},$$
$$\mathcal{L}_1 = -C(\Xi)\,\nabla\Phi_R(\theta)\cdot\nabla_\theta + (F_\sigma(\theta,\Xi) - F_0(\theta,\Xi))\cdot\nabla_\theta + \nu\Delta_\theta.$$

The Poisson equation

$$-\mathcal{L}_0 A(\Xi) = C(\Xi) - I_d = \frac{1}{J}\sum_{j=1}^{J}(\xi^{(j)}\otimes\xi^{(j)} - I_d)$$

admits as unique mean-zero solution

$$A(\Xi) = \frac{1}{2J}\sum_{j=1}^{J}(\xi^{(j)}\otimes\xi^{(j)} - I_d).$$

Applying Itô's formula to the function $\psi(\theta,\Xi) = A(\Xi)\,\nabla\Phi_R(\theta)$, we obtain

$$\begin{aligned}
&\psi(\theta_t,\Xi_t) - \psi(\theta_0,\Xi_0) = \\
&\quad \frac{1}{\delta^2}\int_0^t \big(I_d - C(\Xi_s)\big)\nabla\Phi_R(\theta_s)\,\mathrm{d}s + \int_0^t \big(F_\sigma(\theta_s,\Xi_s)\cdot\nabla_\theta\big)\psi(\theta_s,\Xi_s)\mathrm{d}s \\
&\quad + \sqrt{\frac{2}{\delta^2}}\sum_{j=1}^{J}\int_0^t (\mathrm{d}w_s^{(j)}\cdot\nabla_{\xi^{(j)}})\psi(\theta_s,\Xi_s) + \nu\sqrt{2}\sum_{j=1}^{J}\int_0^t (\mathrm{d}w_s\cdot\nabla_\theta)\psi(\theta_s,\Xi_s).
\end{aligned}$$

Since the rigorous interpretation of the Itô SDE (2.4a) is in integral form, we have

$$\theta_t - \theta_0 = -\int_0^t C(\Xi_s)\nabla\Phi_R(\theta_s)\,\mathrm{d}s + \int_0^t \big(F_\sigma(\theta,\Xi_s) - F_0(\theta,\Xi_s)\big)\,\mathrm{d}s + \sqrt{2}w_t,$$

and so we deduce

$$\begin{aligned}
\theta_t - \theta_0 = &-\int_0^t \nabla\Phi_R(\theta_s)\,\mathrm{d}s + \int_0^t \big(F_\sigma(\theta_s,\Xi_s) - F_0(\theta_s,\Xi_s)\big)\,\mathrm{d}s + \sqrt{2}w_t \\
&+ \delta^2\big(\psi(\theta_t;\Xi_t) - \psi(\theta_0;\Xi_0)\big) - \delta^2\int_0^t \big(F_\sigma(\theta_s,\Xi_s)\cdot\nabla_\theta\big)\psi(\theta_s;\Xi_s)\,\mathrm{d}s - \delta M_t - \delta^2 N_t
\end{aligned}$$

where M_t and N_t are the martingale terms

$$M_t = \sqrt{2}\sum_{j=1}^{J}\int_0^t \left(\mathrm{d}w_s^{(j)}\cdot\nabla_{\xi^{(j)}}\right)\psi(\theta_s,\Xi_s) =: \sum_{j=1}^{J} M_t^{(j)},$$
$$N_t = \nu\sqrt{2}\int_0^t (\mathrm{d}w_s\cdot\nabla_\theta)\,\psi(\theta_s,\Xi_s).$$

Using the fact that ϑ solves the averaged equation (2.8), we deduce

$$\begin{aligned}
\theta_t - \vartheta_t = &-\int_0^t \big(\nabla\Phi_R(\theta_s) - \nabla\Phi_R(\vartheta_s)\big)\,\mathrm{d}s + \int_0^t \big(F_\sigma(\theta_s,\Xi_s) - F_0(\theta_s,\Xi_s)\big)\,\mathrm{d}s \\
&+ \delta^2\big(\psi(\theta_t;\Xi_t) - \psi(\theta_0;\Xi_0)\big) - \delta^2\int_0^t \big(F_\sigma(\theta_s,\Xi_s)\cdot\nabla_\theta\big)\psi(\theta_s;\Xi_s)\,\mathrm{d}s - \delta M_t - \delta^2 N_t.
\end{aligned}$$

Let $e_t = \theta_t - \vartheta_t$. Using the Lipschitz continuity of $\nabla\Phi_R$, the bound (4.2), and the fact that $|\psi(\theta_s, \Xi_s)|$ and its gradient $|\nabla_\theta \psi(\theta_s, \Xi_s)|$ are bounded from above by $C\left(1 + \sum_{j=1}^J |\xi_s^{(j)}|^2\right)$ by definition of ψ,

$$|e_t| \le L \int_0^t |e_s| \, \mathrm{d}s + C\sigma \sum_{j=1}^J \int_0^t \left|\xi_s^{(j)}\right|^3 \mathrm{d}s + C\delta^2 \left(\sum_{j=1}^J \left|\xi_0^{(j)}\right|^2 + \sum_{j=1}^J \left|\xi_t^{(j)}\right|^2\right) + C\delta^2 \int_0^t \left(1 + \sum_{j=1}^J \left|\xi_s^{(j)}\right|^5\right) \mathrm{d}s + \delta \, |M_t| + \delta^2 \, |N_t| \, .$$

Here L is the Lipschitz constant of $\nabla\Phi_R$. Raising to the power p, taking the supremum and taking the expectation, we obtain

$$\begin{aligned}\mathbf{E}\left(\sup_{0\le s\le t} |e_s|^p\right) \le\ & CT^{p-1} \int_0^t \mathbf{E}\, |e_s|^p \, \mathrm{d}s + C\, \sigma^p T^{p-1} \sum_{j=1}^J \int_0^t \mathbf{E}\left|\xi_s^{(j)}\right|^{3p} \mathrm{d}s \\ &+ C\delta^{2p}\left(\sum_{j=1}^J \mathbf{E}|\xi_0^{(j)}|^{2p} + \sum_{j=1}^J \mathbf{E}\left(\sup_{0\le s\le t} |\xi_s^{(j)}|^{2p}\right)\right) + C\delta^{2p}T^{p-1}\int_0^t \left(1 + \sum_{j=1}^J \mathbf{E}\left|\xi_s^{(j)}\right|^{5p}\right) \mathrm{d}s \\ &+ C\delta^p \mathbf{E}\left(\sup_{0\le s\le t} |M_t|^p\right) + C\delta^{2p}\mathbf{E}\left(\sup_{0\le s\le t} |N_t|^p\right) \qquad \forall t \in [0,T].\end{aligned}$$

Since $\xi^{(1)}, \dots, \xi^{(J)}$ are identically distributed stationary stochastic processes, their moments are constant in time and they coincide, so we deduce

$$\begin{aligned}\mathbf{E}\left(\sup_{0\le s\le t} |e_s|^p\right) \le & C\int_0^t \mathbf{E}\left(\sup_{0\le u\le s} |e_u|^p\right) \mathrm{d}s + C\sigma^p + C\delta^{2p} + C\delta^{2p}\mathbf{E}\left(\sup_{0\le s\le t} |\xi_s^{(1)}|^{2p}\right) \\ &+ C\delta^p \mathbf{E}\left(\sup_{0\le s\le t} |M_t|^p\right) + C\delta^{2p}\mathbf{E}\left(\sup_{0\le s\le t} |N_t|^p\right) \qquad \forall t \in [0,T].\end{aligned}$$

By Lemma A.1, there exists a constant C depending only on p such that

$$\forall \delta > 0, \qquad \mathbf{E}\left(\sup_{0\le t\le T} \left|\xi_t^{(1)}\right|^p\right) \le C\left(1 + \log\left(1 + \frac{T}{\delta^2}\right)\right)^{p/2}, \tag{4.4}$$

Therefore, since

$$\delta \log\left(1 + \left(1 + \frac{T}{\delta^2}\right)\right) \xrightarrow[\delta\to 0]{} 0,$$

it holds

$$\forall \delta \in (0,1], \qquad \delta^p \mathbf{E}\left(\sup_{0\le t\le T} |\xi_1|^{2p}\right) \le C. \tag{4.5}$$

Let us now bound the martingale terms. Using the fact that the summands in the definition of M_t are identically distributed and using the moment inequality [50, Theorem 7.2], which is valid for $p \ge 2$, we

obtain

$$\mathbf{E}\left(\sup_{0\le t\le T}|M_t|^p\right) \le \mathbf{E}\left(J^{p-1}\sum_{j=1}^{J}\sup_{0\le t\le T}\left|M_t^{(j)}\right|^p\right) = J^p\mathbf{E}\left(\sup_{0\le t\le T}\left|M_t^{(1)}\right|^p\right)$$
$$\le J^p\left(\frac{p^3}{p-1}\right)^{\frac{p}{2}} T^{\frac{p-2}{p}}\int_0^T \mathbf{E}\left|\nabla_{\xi^{(1)}}\left(A(\Xi_t)\nabla\Phi_R(\theta_t)\right)\right|^p \,\mathrm{d}t$$

Since $\nabla\Phi_R$ is bounded on $\mathbf{T}^d$ and the moments of $\xi_t^{(1)},\dots,\xi_t^{(J)}$ are constant in time, we obtain the bound

$$\forall p\ge 2,\quad \forall T\ge 0,\qquad \mathbf{E}\left(\sup_{0\le t\le T}|M_t|^p\right)\le CT^{1+\frac{p-2}{p}},$$

and for $1<p\le 2$ we have

$$\mathbf{E}\left(\sup_{0\le t\le T}|M_t|^p\right)\le \mathbf{E}\left(\sup_{0\le t\le T}\left(1+|M_t|^2\right)\right)=1+\mathbf{E}\left(\sup_{0\le t\le T}|M_t|^2\right)\le 1+CT.$$

Similarly, for fixed T it holds

$$\forall p>1,\qquad \mathbf{E}\left(\sup_{0\le t\le T}|N_t|^p\right)\le C.$$

Using these bounds together with (4.5), we obtain

$$\mathbf{E}\left(\sup_{0\le s\le t}|e_s|^p\right)\le C\int_0^t\mathbf{E}\left(\sup_{0\le u\le s}|e_u|^p\right)\mathrm{d}s + C\sigma^p + C\delta^p.$$

We then obtain the required bound by Grönwall's inequality, which concludes the proof. ∎

Proof of Theorem 2.1 when $G\in C^3(\mathbf{T}^d,\mathbf{R}^K)$. The idea of the proof is the same. The only difference is that now we consider additionally the Poisson equation

$$-\mathcal{L}_0\phi(\Xi;\theta) = -F_1(\theta,\Xi).$$

Since the right-hand side is a cubic polynomial in $\xi^{(1)},\dots,\xi^{(J)}$ for fixed θ, its average with respect to the invariant measure of Ξ is zero and the solution to the equation is itself cubic (with only cubic and linear terms) in the variables $\xi^{(1)},\dots,\xi^{(J)}$. Indeed, the eigenfunctions of $\mathcal{L}_0$ are given by tensor products of Hermite polynomials; see, for example, [56, Section 4.4] and [1]. Therefore, after applying Itô's formula to $\varphi(\theta;\Xi) := A(\Xi)\,\nabla\Phi_R(\theta)+\phi(\Xi;\theta)$, we obtain

$$\begin{aligned}\theta_t-\theta_0 = &-\int_0^t\nabla\Phi_R(\theta_s)\,\mathrm{d}s+\sqrt{2}w_t+\int_0^t\big(F_\sigma(\theta_s,\Xi_s)-F_0(\theta_s,\Xi_s)-\sigma F_1(\theta_s,\Xi_s)\big)\,\mathrm{d}s\\
&+\delta^2\big(\varphi(\theta_t;\Xi_t)-\varphi(\theta_0;\Xi_0)\big)-\delta^2\int_0^t\left(F_\sigma(\theta_s,\Xi_s)\cdot\nabla_\theta\right)\varphi(\theta;\Xi)\,\mathrm{d}s\\
&+\delta\sqrt{2}\sum_{j=1}^{J}\int_0^t\left(\mathrm{d}w_s^{(j)}\cdot\nabla_{\xi^{(j)}}\right)\varphi(\theta_s;\Xi_s)+\nu\delta^2\sqrt{2}\int_0^t\left(\mathrm{d}w_s\cdot\nabla_\theta\right)\varphi(\theta_s;\Xi_s).\end{aligned}$$

By (4.3), the last term on the first line leads to a bound scaling as σ^2. The other terms are bounded as in the proof of Theorem 2.1 in the case where $G\in C^2(\mathbf{T}^d,\mathbf{R}^K)$, which is possible because, like ψ in that proof, the function φ and its derivatives are polynomial functions in the variables $\xi^{(1)},\dots,\xi^{(J)}$. ∎

4.1. Analysis of the Discrete-time Numerical Method. Before showing Theorem 2.2, we show a preparatory result.

PROPOSITION 4.1. *Assume that $G \in C^2(\mathbf{T}^d)$ and let $\hat{\theta}_n$ be the solution obtained by (2.7). Then there exists a constant $C = C(T, J)$ such that*

$$\forall (\sigma, \Delta) \in \mathbf{R}^+ \times \mathbf{R}^+, \qquad \sup_{0 \le n \le N} \mathbf{E} \left| \hat{\theta}_n - \vartheta_{n\Delta} \right|^2 \le C \left(\Delta + \sigma^{2\beta} \right),$$

where

$$\beta = \begin{cases} 1 & \text{if } G \in C^2(\mathbf{T}^d), \\ 2 & \text{if } G \in C^3(\mathbf{T}^d). \end{cases}$$

Proof. Our strategy of proof is loosely based on that of [17, Theorem 2.4]. Let us denote by $\{\hat{\vartheta}_n\}_{n=0}^N$ the Euler–Maruyama approximation of the solution $\{\vartheta_t\}_{t \in [0,T]}$ to the averaged equation (2.8), i.e. the discrete-time solution obtained from the iteration

$$\hat{\vartheta}_{n+1} = \hat{\vartheta}_n + \nabla \Phi_R(\hat{\vartheta}_n)\, \Delta + \nu \sqrt{2\Delta}\, x_n, \qquad \hat{\vartheta}_0 = \hat{\theta}_0. \tag{4.6}$$

By the standard theory of numerical methods for SDEs [23, 39], the difference between $\vartheta_{n\Delta}$ and its approximation $\hat{\vartheta}_n$ satisfies the bound

$$\mathbf{E} \left(\sup_{0 \le n \le T/\Delta} \left| \hat{\vartheta}_n - \vartheta_{n\Delta} \right|^2 \right) \le C\Delta. \tag{4.7}$$

Here and below, C denotes a constant independent of σ and Δ that is allowed to change from occurrence to occurrence. Subtracting (4.6) from the equation for $\hat{\theta}_n$ (2.7), we obtain that the error $e_n := \hat{\theta}_n - \hat{\vartheta}_n$ at step n satisfies

$$\begin{aligned} e_{n+1} &= e_n + \left(F_\sigma(\hat{\theta}_n) + C(\hat{\Xi}_n) \nabla \Phi_R(\hat{\theta}_n) \right) \Delta + \left(\nabla \Phi_R(\hat{\vartheta}_n) - \nabla \Phi_R(\hat{\theta}_n) \right) \Delta \\ &\qquad + \left(\nabla \Phi_R(\hat{\theta}_n) - C(\hat{\Xi}_n) \nabla \Phi_R(\hat{\theta}_n) \right) \Delta \\ &=: e_n + X_n \Delta + Y_n \Delta + Z_n \Delta. \end{aligned}$$

Therefore

$$\mathbf{E} \left| e_{n+1} \right|^2 = \mathbf{E} \left| e_n + \Delta X_n + \Delta\, Y_n \right|^2 + \Delta^2 \mathbf{E} \left| Z_n \right|^2 + 2\Delta \mathbf{E} \left((e_n + \Delta X_n + \Delta Y_n) \cdot Z_n \right).$$

Using the tower property of conditional expectation, it holds

$$\left| \mathbf{E} \left(e_n \cdot Z_n \right) \right| = \left| \mathbf{E} \left(\mathbf{E} \left(e_n \cdot Z_n \,|\, \hat{\theta}_n, \hat{\vartheta}_n \right) \right) \right| = \left| \mathbf{E} \left(e_n \cdot \mathbf{E} \left(Z_n \,|\, \hat{\theta}_n, \hat{\vartheta}_n \right) \right) \right| = 0,$$

because, by the definition of Z_n and the fact that $\xi_n^{(1)}, \dots, \xi_n^{(J)}$ are independent of $\hat{\theta}_n$ and $\hat{\vartheta}_n$, it holds $\mathbf{E} \left(Z_n \,|\, \hat{\theta}_n, \hat{\vartheta}_n \right) = 0$. Thus, since by Young's inequality $(a+b)^2 \le (1+\varepsilon) a^2 + \left(1 + \frac{1}{\varepsilon}\right) b^2$ for any $a, b \in \mathbf{R}$,

$$\begin{aligned} \mathbf{E} \left| e_{n+1} \right|^2 &= \mathbf{E} \left| e_n + \Delta X_n + \Delta\, Y_n \right|^2 + \Delta^2 \mathbf{E} \left| Z_n \right|^2 + 2\Delta^2 \mathbf{E} (X_n \cdot Z_n + Y_n \cdot Z_n) \\ &\le (1 + \Delta) \mathbf{E} \left| e_n \right|^2 + \left(2\Delta + 3\Delta^2 \right) \left(\mathbf{E} \left| X_n \right|^2 + \mathbf{E} \left| Y_n \right|^2 \right) + 3\Delta^2 \mathbf{E} \left| Z_n \right|^2. \end{aligned} \tag{4.8}$$

We now bound the terms one by one.

- By (4.2), it holds

$$\forall(\theta,\Xi)\in\mathbf{T}^d\times(\mathbf{R}^d)^J,\qquad |F_\sigma(\theta,\Xi)+C(\Xi)\nabla\Phi_R(\theta)|\le C\sigma\sum_{j=1}^J\left|\xi^{(j)}\right|^3.$$

  Taking the expectation and using the fact that the moments of $\hat\xi_n^{(1)},\dots,\hat\xi_n^{(J)}$ are constant in n, it holds $\mathbf{E}\left|X_n\right|^2\le C\sigma^2$. If G is three times differentiable, we can carry out the Taylor expansion to the next order as in (4.3), leading to the refined bound $\mathbf{E}\left|X_n\right|^2\le C\sigma^4$.
- The expectation of $\left|Y_n\right|^2$ can be bounded by using the Lipschitz continuity of Φ_R:

$$\mathbf{E}\left|Y_n\right|^2=\mathbf{E}\left|\nabla\Phi_R(\hat\vartheta_n)-\nabla\Phi_R(\hat\theta_n)\right|^2\le C\mathbf{E}\left|e_n\right|^2.$$

- To bound $\mathbf{E}\left|Z_n\right|^2$, we use the fact that Φ_R is uniformly bounded and that the moments of $\xi_n^{(1)},\dots,\xi_n^{(1)}$ are constant in n:

$$\mathbf{E}\left|Z_n\right|^2\le 2\mathbf{E}\left|\nabla\Phi_R(\hat\theta_n)\right|^2+2\mathbf{E}\left|C(\hat\Xi_n)\nabla\Phi_R(\hat\theta_n)\right|^2\le C. \tag{4.9}$$

Going back to (4.8) and combining the bounds, we deduce

$$\begin{aligned}\forall\delta\le 1,\qquad \mathbf{E}\left|e_{n+1}\right|^2&\le(1+\Delta)\mathbf{E}\left|e_n\right|^2+C\Delta(\sigma^{2\beta}+\mathbf{E}\left|e_n\right|^2)+C\Delta^2\\&\le(1+C\Delta)\mathbf{E}\left|e_n\right|^2+C\Delta\left(\sigma^{2\beta}+\Delta\right).\end{aligned}$$

Let $\varepsilon=\sigma^\beta+\Delta$. Applying the previous bound recursively,

$$\begin{aligned}\mathbf{E}\left|e_{n+1}\right|^2&\le(1+C\Delta)((1+C\Delta)\mathbf{E}\left|e_{n-1}\right|^2+C\Delta\varepsilon)+C\Delta\varepsilon\\&\le\cdots\le(1+C\Delta)^{n+1}\,\mathbf{E}\left|e_0\right|^2+C\Delta\varepsilon\sum_{i=1}^n(1+C\Delta)^{n-i}\\&\le\mathbf{E}\left|e_0\right|^2\,\mathrm{e}^{CT}+(n\Delta)\,\mathrm{e}^{CT}\,C\varepsilon\le C(\mathbf{E}\left|e_0\right|^2+\varepsilon).\end{aligned}$$

Since $\mathbf{E}\left|e_0\right|^2=0$, and in view of (4.7), this concludes the proof. ∎

We now prove the more general Theorem 2.2, applicable to (2.6). To this end, it is useful to introduce, for $j=1,\dots,J$ and $n=0,\dots,N$, an approximation $\hat v_n^{(j)}$ of $\hat\xi_n^{(j)}$ in (2.6) such that the processes $\hat v_\bullet^{(j)}$ have a compactly-supported autocorrelation function. Notice first that, with the same notation as in (2.6), it holds

$$\hat\xi_n^{(j)}=\hat\xi_0^{(j)}\,\mathrm{e}^{-n\frac{\Delta}{\delta^2}}+\sqrt{1-\mathrm{e}^{-2\frac{\Delta}{\delta^2}}}\sum_{m=1}^n x_{m-1}^{(j)}\,\mathrm{e}^{-(n-m)\frac{\Delta}{\delta^2}}.$$

In order to work with iterates that are uncorrelated when far apart in time, it is natural to define the approximation

$$\hat v_n^{(j)}=\begin{cases}\hat\xi_n^{(j)} & \text{if } n\le M\\ \sqrt{1-\mathrm{e}^{-2\frac{\Delta}{\delta^2}}}\sum_{m=n-M}^n x_{m-1}^{(j)}\,\mathrm{e}^{-(n-m)\frac{\Delta}{\delta^2}} & \text{if } n>M.\end{cases} \tag{4.10}$$

We denote the collection $(\hat{v}_n^{(1)}, \dots, \hat{v}_n^{(J)})$ by $\hat{\Upsilon}_n$. The following lemma, proved in Appendix B, is useful in the proof of Theorem 2.2 below.

LEMMA 4.2. *Let $\{\hat{v}_n^{(j)}\}_{n=0}^N$ and $\{\hat{\xi}_n^{(j)}\}_{n=0}^N$, for $j = 1, \dots, J$, be the discrete-time processes obtained by (2.6) and (4.10). Then the following bound holds for a constant C independent of δ:*

$$\forall j \in \{1, \dots, J\}, \quad \forall n \in \{0, \dots, N\}, \qquad \mathbf{E}\left|\hat{\xi}_n^{(j)} - \hat{v}_n^{(j)}\right|^4 \leq C\, \mathrm{e}^{-4(M+1)\frac{\Delta}{\delta^2}}. \tag{4.11}$$

Consequently, it holds

$$\forall n \in \{0, \dots, N\}, \qquad \mathbf{E}\left|C(\hat{\Upsilon}_n) - C(\hat{\Xi}_n)\right|^2 \leq C\, \mathrm{e}^{-2(M+1)\frac{\Delta}{\delta^2}}, \tag{4.12}$$

where we used the notation $C(\hat{\Upsilon}_n) = \frac{1}{J}\sum_{j=1}^J \hat{v}_n^{(j)} \otimes \hat{v}_n^{(j)}$.

Proof of Theorem 2.2. Throughout the proof, C denotes a constant independent of δ, σ and Δ, allowed to change from occurrence to occurrence. Let $e_n = \hat{\theta}_n - \hat{\vartheta}_n$, where $\hat{\vartheta}_n$ is as defined in (4.6). In view of (4.7), it is sufficient to obtain a bound on e_n with the same right-hand side as in (2.10). Using the same definition for X_n and Y_n as in the proof of Proposition 4.1, we have

$$e_{n+1} = e_n + X_n\Delta + Y_n\Delta + Z_n\Delta + W_n\Delta,$$

with now

$$Z_n = \left(C(\hat{\Upsilon}_n)\nabla\Phi_R(\hat{\theta}_n) - C(\hat{\Xi}_n)\nabla\Phi_R(\hat{\theta}_n)\right), \qquad C(\hat{\Upsilon}_n) = \frac{1}{J}\sum_{j=1}^J \hat{v}_n^{(j)} \otimes \hat{v}_n^{(j)},$$
$$W_n = \left(\nabla\Phi_R(\hat{\theta}_n) - C(\hat{\Upsilon}_n)\nabla\Phi_R(\hat{\theta}_n)\right).$$

Using the fact that $e_0 = 0$, we deduce

$$\begin{aligned}|e_n|^2 &\leq 4\Delta^2\left|\sum_{m=0}^{n-1} X_m\right|^2 + 4\Delta^2\left|\sum_{m=0}^{n-1} Y_m\right|^2 + 4\Delta^2\left|\sum_{m=0}^{n-1} Z_m\right|^2 + 4\Delta^2\left|\sum_{m=0}^{n-1} W_m\right|^2 \\ &\leq 4n\Delta^2\sum_{m=0}^{n-1}|X_m|^2 + 4n\Delta^2\sum_{m=0}^{n-1}|Y_m|^2 + 4n\Delta^2\sum_{m=0}^{n-1}|Z_m|^2 + 4\Delta^2\left|\sum_{m=0}^{n-1} W_m\right|^2.\end{aligned}$$

Consequently, for $n \leq T/\Delta$,

$$\sup_{0\leq m\leq n}\mathbf{E}\,|e_m|^2 \leq 4T\Delta\sum_{m=0}^{n-1}(\mathbf{E}\,|X_m|^2 + \mathbf{E}\,|Y_m|^2 + \mathbf{E}\,|Z_m|^2) + 4\Delta^2\sup_{0\leq m\leq n}\left(\mathbf{E}\left|\sum_{\ell=0}^{m-1} W_\ell\right|^2\right). \tag{4.13}$$

The first term is bounded as in the proof of Proposition 4.1. For the second term, we use the Lipschitz continuity of Φ_R to deduce

$$\mathbf{E}\,|Y_m|^2 = \mathbf{E}\left|\nabla\Phi_R(\hat{\theta}_m) - \nabla\Phi_R(\hat{\vartheta}_m)\right|^2 \leq C\mathbf{E}\,|e_m|^2 \leq C\sup_{0\leq\ell\leq m}\mathbf{E}\,|e_\ell|^2.$$

For the third term, we use (4.12), which gives

$$\mathbf{E}\,|Z_n|^2 \le C\mathbf{E}\left|C(\hat{\Upsilon}_n) - C(\hat{\Xi}_n)\right|^2 \le C\,\mathrm{e}^{-2(M+1)\frac{\Delta}{\delta^2}}\,.$$

In order to bound the last term of (4.13), we calculate

$$\begin{aligned}
\mathbf{E}\left|\sum_{\ell=0}^{m-1} W_\ell\right|^2 &= \sum_{\ell=0}^{m-1}\sum_{k=0}^{m-1}\mathbf{E}(W_\ell\cdot W_k)\\
&= \sum_{\ell=0}^{m-1}\mathbf{E}\,|W_\ell|^2 + 2\sum_{\ell=0}^{m-1}\sum_{k=\ell+1}^{(\ell+M)\wedge(m-1)}\mathbf{E}(W_\ell\cdot W_k) + 2\sum_{\ell=0}^{m-1}\sum_{k=\ell+M+1}^{m-1}\mathbf{E}(W_\ell\cdot W_k)\\
&\le \sum_{\ell=0}^{m-1}\mathbf{E}\,|W_\ell|^2 + 2M\sum_{\ell=0}^{m-1}\mathbf{E}\,|W_\ell|^2 + 2\sum_{\ell=0}^{m-1}\sum_{k=\ell+M+1}^{m-1}\mathbf{E}(W_\ell\cdot W_k).
\end{aligned}$$

Since $\nabla\Phi_R$ is bounded and the moments of $\hat{v}_n^{(j)}$ are bounded uniformly in n and j, we can bound the first and second sums, which leads to

$$\mathbf{E}\left|\sum_{\ell=0}^{m-1} W_\ell\right|^2 \le Cm(1+2M) + 2\sum_{\ell=0}^{m-1}\sum_{k=\ell+M+1}^{m-1}\mathbf{E}(W_\ell\cdot W_k) \tag{4.14}$$

We now bound the second term uniformly for ℓ and k satisfying $k \ge \ell+M+1$. Using the tower property of conditional expectation, we notice that

$$\begin{aligned}
|\mathbf{E}(W_\ell\cdot W_k)| &= \left|\mathbf{E}(\mathbf{E}(W_\ell\cdot W_k\,|\,\hat{\theta}_\ell,\hat{\Upsilon}_\ell))\right| = \left|\mathbf{E}(W_\ell\cdot\mathbf{E}(W_k\,|\,\hat{\theta}_\ell,\hat{\Upsilon}_\ell))\right|\\
&= \left|\mathbf{E}(W_\ell\cdot\mathbf{E}(W_k\,|\,\hat{\theta}_\ell,\hat{\Upsilon}_\ell))\right| \le \sqrt{\mathbf{E}\,|W_\ell|^2\,\mathbf{E}\left|\mathbf{E}(W_k\,|\,\hat{\theta}_\ell,\hat{\Upsilon}_\ell)\right|^2} \le C\sqrt{\mathbf{E}\left|\mathbf{E}(W_k\,|\,\hat{\theta}_\ell,\hat{\Upsilon}_\ell)\right|^2}.
\end{aligned}$$

Using the notations $\mathbf{E}_\ell = \mathbf{E}(\bullet\,|\,\hat{\theta}_\ell,\hat{\Upsilon}_\ell)$ and $\mathbf{E}_\ell(\bullet_1\,|\,\bullet_2) = \mathbf{E}(\bullet_1\,|\,\hat{\theta}_\ell,\hat{\Upsilon}_\ell,\bullet_2)$ for conciseness, we obtain

$$\begin{aligned}
\mathbf{E}_\ell(W_k) &= \frac{1}{J}\sum_{j=1}^{J}\mathbf{E}_\ell\Big(\big(I_d-\hat{v}_k^{(j)}\otimes\hat{v}_k^{(j)}\big)\nabla\Phi_R(\hat{\theta}_k)\Big) = \frac{1}{J}\sum_{j=1}^{J}\mathbf{E}_\ell\Big(\mathbf{E}_\ell\Big(\big(I_d-\hat{v}_k^{(j)}\otimes\hat{v}_k^{(j)}\big)\nabla\Phi_R(\hat{\theta}_k)\,|\,\hat{v}_k^{(j)}\Big)\Big)\\
&= \frac{1}{J}\sum_{j=1}^{J}\mathbf{E}_\ell\Big(\big(I_d-\hat{v}_k^{(j)}\otimes\hat{v}_k^{(j)}\big)\mathbf{E}_\ell\left(\nabla\Phi_R(\hat{\theta}_k)\,|\,\hat{v}_k^{(j)}\right)\Big)\\
&= \frac{1}{J}\sum_{j=1}^{J}\mathbf{E}_\ell\left(I_d-\hat{v}_k^{(j)}\otimes\hat{v}_k^{(j)}\right)\mathbf{E}_\ell\left(\nabla\Phi_R(\hat{\theta}_k)\right) \qquad (4.15)\\
&\quad+\frac{1}{J}\sum_{j=1}^{J}\mathbf{E}_\ell\Big(\big(I_d-\hat{v}_k^{(j)}\otimes\hat{v}_k^{(j)}\big)\big(\mathbf{E}_\ell\big(\nabla\Phi_R(\hat{\theta}_k)\,|\,\hat{v}_k^{(j)}\big)-\mathbf{E}_\ell\big(\nabla\Phi_R(\hat{\theta}_k)\big)\big)\Big).
\end{aligned}$$

Now note that, for any $k > \ell+M$ and any $j\in\{1,\dots,J\}$, the random variable $\hat{v}_k^{(j)}$ is independent of $\hat{\theta}_\ell$ and $\hat{\Upsilon}_\ell$, and it has distribution

$$\mathcal{N}\Big(0,\big(1-\mathrm{e}^{-2(M+1)\frac{\Delta}{\delta^2}}\big)\,I_d\Big).$$

Consequently, we can calculate the first expectation in (4.15) exactly. Using Hölder's inequality for the other

term, we obtain

$$|\mathbf{E}_\ell(W_k)| \le C\,\mathrm{e}^{-2(M+1)\frac{\Delta}{\delta^2}} + C\sum_{j=1}^{J}\mathbf{E}_\ell\left|\mathbf{E}_\ell\left(\nabla\Phi_R(\hat\theta_k)\,|\,\upsilon_k^{(j)}\right) - \mathbf{E}_\ell\left(\nabla\Phi_R(\hat\theta_k)\right)\right|^2.$$

Therefore, employing Jensen's inequality, we deduce

$$\begin{aligned}\mathbf{E}\,|\mathbf{E}_\ell(W_k)|^2 &\le C\,\mathrm{e}^{-4(M+1)\frac{\Delta}{\delta^2}} + C\sum_{j=1}^{J}\mathbf{E}\left|\mathbf{E}_\ell\left(\nabla\Phi_R(\hat\theta_k)\,|\,\upsilon_k^{(j)}\right) - \mathbf{E}_\ell\left(\nabla\Phi_R(\hat\theta_k)\right)\right|^4 \\ &=: C\,\mathrm{e}^{-4(M+1)\frac{\Delta}{\delta^2}} + C\sum_{j=1}^{J}\mathbf{E}(B_j).\end{aligned} \tag{4.16}$$

Since $\mathbf{E}_\ell\big(\nabla\Phi_R(\hat\theta_{k-M})\,|\,\hat\upsilon_k^{(j)}\big) = \mathbf{E}_\ell\big(\nabla\Phi_R(\hat\theta_{k-M})\big)$, it holds

$$\begin{aligned}B_j &\le 2^3\left|\mathbf{E}_\ell\left(\nabla\Phi_R(\hat\theta_k)\,|\,\hat\upsilon_k^{(j)}\right) - \mathbf{E}_\ell\left(\nabla\Phi_R(\hat\theta_{k-M})\,|\,\hat\upsilon_k^{(j)}\right)\right|^4 + 2^3\left|\mathbf{E}_\ell\left(\nabla\Phi_R(\hat\theta_{k-M})\right) - \mathbf{E}_\ell\left(\nabla\Phi_R(\hat\theta_k)\right)\right|^4 \\ &= 2^3\left|\mathbf{E}_\ell\left(\nabla\Phi_R(\hat\theta_k) - \nabla\Phi_R(\hat\theta_{k-M})\,|\,\hat\upsilon_k^{(j)}\right)\right|^4 + 2^3\left|\mathbf{E}_\ell\left(\nabla\Phi_R(\hat\theta_{k-M}) - \nabla\Phi_R(\hat\theta_k)\right)\right|^4.\end{aligned}$$

Using Jensen's inequality and the Lipschitz continuity of $\nabla\Phi_R$, we deduce

$$B_j \le 2^3\mathbf{E}_\ell\left(\left|\hat\theta_k - \hat\theta_{k-M}\right|^4\,|\,\hat\upsilon_k^{(j)}\right) + 2^3\mathbf{E}_\ell\left|\hat\theta_{k-M} - \hat\theta_k\right|^4,$$

so $\mathbf{E}(B_j) \le 2^4\mathbf{E}\left|\hat\theta_{k-M} - \hat\theta_k\right|^4$. Using the bound

$$\mathbf{E}\left|\hat\theta_k - \hat\theta_{k-M}\right|^4 = \mathbf{E}\left|\sum_{i=k-M}^{k-1} F_\sigma(\hat\theta_i, \hat\Upsilon_i)\Delta + \sum_{i=k-M}^{k-1}\nu\sqrt{2\Delta}x_i\right|^4 \le C\,(M^4\Delta^4 + M^2\Delta^2),$$

which is justified because $F_\sigma(\theta,\Xi) \le C\left(1 + \sum_{j=1}^{J}\left|\xi^{(j)}\right|^3\right)$ by (4.2), we deduce by going back to (4.16) that

$$\mathbf{E}\,|\mathbf{E}_\ell(W_k)|^2 \le C\,\mathrm{e}^{-4(M+1)\frac{\Delta}{\delta^2}} + CM^4\Delta^4 + CM^2\Delta^2.$$

Taking the square root and returning to the usual notation, we obtain

$$\sqrt{\mathbf{E}\left|\mathbf{E}(W_k\,|\,\hat\theta_\ell, \hat\Upsilon_\ell)\right|^2} \le C\,\mathrm{e}^{-2(M+1)\frac{\Delta}{\delta^2}} + CM^2\Delta^2 + CM\Delta.$$

Employing this bound in (4.14), we deduce

$$\begin{aligned}\Delta^2\mathbf{E}\left|\sum_{\ell=0}^{m-1} W_\ell\right|^2 &\le C\Delta(1+2M) + C\left(\mathrm{e}^{-2(M+1)\frac{\Delta}{\delta^2}} + M^2\Delta^2 + M\Delta\right) \\ &\le C\left(\Delta + M\Delta + M^2\Delta^2 + \mathrm{e}^{-2(M+1)\frac{\Delta}{\delta^2}}\right).\end{aligned}$$

Letting $M = \lfloor \log(1+\delta^{-1})\,\frac{\delta^2}{\Delta} \rfloor$, we obtain

$$\Delta^2 \mathbf{E}\left|\sum_{\ell=0}^{m-1} W_\ell\right|^2 \leq C\left(\Delta + \log(1+\delta^{-1})\delta^2 + \left(\log(1+\delta^{-1})\right)^2 \delta^4 + \frac{1}{|1+\delta^{-1}|^2}\right)$$

$$\leq C\big(\Delta + \log(1+\delta^{-1})\delta^2\big) \qquad \forall \delta \in (0,1]. \tag{4.17}$$

Here we used that, by concavity of the logarithm,

$$\log(1+\delta^{-1}) = \log(1+\delta^{-1}) - \log(1) \geq \frac{\delta^{-1}}{1+\delta^{-1}} \geq \frac{\delta^{-2}}{|1+\delta^{-1}|^2},$$

so

$$\log(1+\delta^{-1})\delta^2 \geq \frac{1}{|1+\delta^{-1}|^2}.$$

Note that the second term in (4.17) vanishes in the limit $\delta \to 0$, i.e. when $\xi_n^{(j)}$ are drawn independently from $\mathcal{N}(0, I_d)$ at each iteration. In this case, we recover the statement of Proposition 4.1.

Combining everything in (4.13), we obtain

$$\sup_{0\leq m\leq n} \mathbf{E}\,|e_m|^2 \leq C\left(\Delta + \sigma^{2\beta} + \log(1+\delta^{-1})\delta^2 + \Delta \sum_{m=0}^{n-1}\left(\sup_{0\leq \ell\leq m} \mathbf{E}\,|e_\ell|^2\right)\right).$$

By the discrete Grönwall's lemma, we deduce

$$\sup_{0\leq n\leq N} \mathbf{E}\,|e_n|^2 \leq C\,\mathrm{e}^T\left(\Delta + \sigma^{2\beta} + \log(1+\delta^{-1})\delta^2\right),$$

which concludes the proof. ∎

5. Conclusions and Perspectives for Future Work. In this paper, we introduce a new derivative-free and adjoint-free method for solving Bayesian inverse problems, specifically for the tasks of sampling from the Bayesian posterior or finding the maximum *a posteriori* estimate. The method relies on a gradient approximation with a structure similar to that in the ensemble Kalman methods for sampling (EKS) and inversion (EKI). In contrast with these algorithms and other approximate sampling methods such as UKS and CBS, however, the method we propose is provably refineable: using tools from multiscale analysis, we prove its convergence to the gradient descent or overdamped Langevin dynamics, depending on whether it is employed for inversion or sampling, respectively.

Since our method is a variation on standard gradient descent, it suffers from slow convergence when the Bayesian posterior exhibits strong anisotropy or, relatedly, when the Hessian of the regularized least-squares functional has a large condition number in the part of the domain close to the MAP estimator. In order to remedy this possible issue, we propose a preconditioning methodology based on information from EKS (or UKS, CBS), and we demonstrate its efficacy for both inversion and sampling through careful numerical experiments.

Several exciting research avenues remain open for future work. On the theoretical front, it would be interesting to obtain a uniform-in-time weak error estimate, both for the continuous-time dynamics and its discrete-time approximation. This might prove challenging even in the case of a compact parameter space, because the state space of the auxiliary processes employed for the gradient approximation is unbounded. Relatedly, it would be useful to obtain a bound, in terms of the parameters σ and δ and in an appropriate

metric, on the distance between the true Bayesian posterior and that approximated by the method, i.e. the θ-marginal of the invariant measure of (2.1). It would also be interesting to study other time discretizations of (2.1a) than the one employed in this paper; for example, we could consider discretizations where Bernoulli random variables are employed instead of exact (in law) Brownian increments, which should not change the weak convergence properties of the method [39, 52], or semi-implicit discretizations (which preserve linearity of the updates) based on the formulation of ensemble Kalman methods in [41, Section 4.3.3]. Finally, our results could be extended to the case of an unbounded parameter space, which, in principle, should be possible by using results from [55].

On the practical side, it will be important to determine how the method can be coupled to computationally less expensive preconditioners than those computed from the EKS, or how EKS, and related methods such as UKS and CBS, can be accelerated. One may also explore questions related to the parametrization of the multiscale method. Since a larger value of the parameter σ in (2.1) seems to be associated with faster convergence initially but a larger error later on, as noted in the description of Figure 1, it would be interesting to investigate whether a computational gain can be obtained by adapting σ during a simulation. One might, for example, start the dynamics with a relatively large value of σ in order to favor exploration initially, and then progressively decrease this parameter in order to increase accuracy once the distinguished particle has reached regions of high posterior probability density. Finally, it would be interesting to study more precisely and more generally the influence of the parameter δ, in order to determine, for example, whether a large value of this parameter can be advantageous for promoting exploration in rugged landscapes.

Acknowledgements. G.A.P. was partially supported by the EPSRC through the grant number EP/P031587/1 and by JPMorgan Chase & Co under a J.P. Morgan A.I. Research Award 2019. (Any views or opinions expressed herein are solely those of the authors listed, and may differ from the views and opinions expressed by JPMorgan Chase & Co. or its affiliates. This material is not a product of the Research Department of J.P. Morgan Securities LLC. This material does not constitute a solicitation or offer in any jurisdiction.) The work of A.M.S. is supported by NSF (award DMS-1818977) and by the Office of Naval Research (award N00014-17-1-2079). The work of U.V. was partially funded by the Fondation Sciences Mathématique de Paris (FSMP), through a postdoctoral fellowship in the "mathematical interactions" program.

Appendix A. Auxiliary Results.

LEMMA A.1. *Let $\delta > 0$ be a fixed parameter and let X_t denote the solution to the scalar Ornstein–Uhlenbeck equation with stationary initial condition,*

$$\mathrm{d}X_t = -\frac{1}{\delta^2} X_t \,\mathrm{d}t + \sqrt{\frac{2}{\delta^2}}\,\mathrm{d}W_t, \qquad X_0 \sim \mathcal{N}(0,1).$$

It holds

$$\forall p \in (1,\infty), \qquad \mathbf{E}\left(\sup_{0\le t\le T} |X_t|^p\right) \le C\,\left(1+\log\left(1+\frac{T}{\delta^2}\right)\right)^{p/2},$$

for a constant C independent of T.

Proof. The strategy of the proof parallels that in [58, Theorem A.1], so here we give only a sketch. We use the notations $W(t)$ and W_t interchangeably. The process $\{X_t\}$ is equal in law to $\{Y_{t/\delta^2}\}$, where Y_t is the solution to

$$\mathrm{d}Y_t = -Y_t\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t, \qquad Y_0 \sim \mathcal{N}(0,1).$$

so we can assume without loss of generality that $\delta = 1$. The process $\{Y_t\}_{t\ge 0}$ is equivalent in law to $Z_t =$

$\mathrm{e}^{-t}\, W(\mathrm{e}^{2t})$; see, for example, [56, Chapter 1]. Therefore, it holds

$$\begin{aligned}\mathbf{E}\left(\sup_{0\le t\le T}|Y_t|^p\right) &= \mathbf{E}\left(\sup_{0\le s\le \mathrm{e}^{2T}}\left|\frac{W_s}{\sqrt{s}}\right|^p\right)\\ &\le \left(\log\log(2+\mathrm{e}^{2T})\right)^{p/2}\,\mathbf{E}\left(\sup_{0\le s\le \mathrm{e}^{2T}}\left|\frac{W_s}{\sqrt{s\log\log(2+s)}}\right|^p\right)\\ &= \left(\log\log(2+\mathrm{e}^{2T})\right)^{p/2}\,\mathbf{E}\left|N\left(\mathrm{e}^{2T}\right)\right|^p,\end{aligned}$$

where, for $S\ge 0$,

$$N(S)=\sup_{0\le s\le S}\left|\frac{W_s}{\sqrt{s\log\log(2+s)}}\right|.$$

Now clearly $\log(\mathrm{e}^{2T}+2)\le \log(\mathrm{e}^{2T})+2 = 2T+2$, so $\log\log(2+\mathrm{e}^{2T})\le 1+\log(1+T)$. It is shown in the proof of [58, Theorem A.1], and in the references therein, that there are constants C and σ such that

$$\forall S>0,\quad \forall\lambda>0,\qquad \mathbf{P}(N(S)>\lambda)\le C\,\mathrm{e}^{-\frac{\lambda^2}{4\sigma^2}}.$$

Therefore

$$\begin{aligned}\mathbf{E}\left|N(\mathrm{e}^{2T})\right|^p &= \int_0^\infty \mathbf{P}(\left|N(\mathrm{e}^{2T})\right|^p>\lambda)\,\mathrm{d}\lambda\\ &= \int_0^\infty \mathbf{P}(N(\mathrm{e}^{2T})>\lambda^{1/p})\,\mathrm{d}\lambda \le \int_0^\infty C\,\mathrm{e}^{-\frac{\lambda^{2/p}}{4\sigma^2}}\,\mathrm{d}\lambda\le K,\end{aligned}$$

for K independent of S, which concludes the proof. ∎

Appendix B. Proof of Lemma 4.2.

If $n\le M$, then $\hat{v}_n^{(j)}$ and $\hat{\xi}_n^{(j)}$ coincide and the statement is true. If $n>M$, then by definition it holds

$$\begin{aligned}\hat{\xi}_n^{(j)}-\hat{v}_n^{(j)} &= \hat{\xi}_0^{(j)}\,\mathrm{e}^{-n\frac{\Delta}{\delta^2}}+\sqrt{1-\mathrm{e}^{-2\frac{\Delta}{\delta^2}}}\sum_{m=1}^{n-M-1}x_{m-1}^{(j)}\,\mathrm{e}^{-(n-m)\frac{\Delta}{\delta^2}}\\ &= \mathrm{e}^{-(M+1)\frac{\Delta}{\delta^2}}\left(\hat{\xi}_0^{(j)}\,\mathrm{e}^{-(n-M-1)\frac{\Delta}{\delta^2}}+\sqrt{1-\mathrm{e}^{-2\frac{\Delta}{\delta^2}}}\sum_{m=1}^{n-M-1}x_{m-1}^{(j)}\,\mathrm{e}^{-(n-M-1-m)\frac{\Delta}{\delta^2}}\right).\end{aligned}$$

Using the inequality $|a+b|^4\le 8\,|a|^4+8\,|b|^4$ for all $a,b\in\mathbf{R}^d$, together with the working assumption that $n>M$, we deduce

$$\mathbf{E}\left|\hat{\xi}_n^{(j)}-\hat{v}_n^{(j)}\right|^4\le \mathrm{e}^{-4(M+1)\frac{\Delta}{\delta^2}}\left(8\mathbf{E}\left|\hat{\xi}_0^{(j)}\right|^4+8\left(1-\mathrm{e}^{-2\frac{\Delta}{\delta^2}}\right)^2\mathbf{E}\left|\sum_{m=1}^{n-M-1}x_{m-1}^{(j)}\,\mathrm{e}^{-(n-M-1-m)\frac{\Delta}{\delta^2}}\right|^4\right). \tag{B.1}$$

It remains to prove that the second term in the round brackets is bounded from above independently of δ. Since it holds $|a-b|^4\le C\sum_{i=1}^d|a_i-b_i|^4$, for all $a,b\in\mathbf{R}^d$, we assume without loss of generality that $d=1$

in order to establish (4.11). For simplicity of notation, let $v_m = x^{(j)}_{n-M-2-m}$, so that

$$\mathbf{E}\left|\sum_{m=1}^{n-M-1} x^{(j)}_{m-1}\,\mathrm{e}^{-(n-M-1-m)\frac{\Delta}{\delta^2}}\right|^4 = \mathbf{E}\left|\sum_{m=0}^{n-M-2} v_m\,\mathrm{e}^{-m\frac{\Delta}{\delta^2}}\right|^4.$$

Expanding the sum and using the independence of $v_0,\dots,v_{n-M-2}$, we calculate

$$\begin{aligned}\mathbf{E}\left|\sum_{m=0}^{n-M-2} v_m\,\mathrm{e}^{-m\frac{\Delta}{\delta^2}}\right|^4 &= \sum_{m=0}^{n-M-2}\mathbf{E}\,|v_m|^4\,\mathrm{e}^{-4m\frac{\Delta}{\delta^2}} + 6\sum_{m=0}^{n-M-2}\sum_{\ell=m+1}^{n-M-2}\mathbf{E}\,|v_m|^2\,\mathbf{E}\,|v_\ell|^2\,\mathrm{e}^{-2(m+\ell)\frac{\Delta}{\delta^2}}\\ &\le 3\sum_{m=0}^{\infty}\mathrm{e}^{-4m\frac{\Delta}{\delta^2}} + 6\sum_{m=0}^{\infty}\sum_{\ell=m+1}^{\infty}\mathrm{e}^{-2(m+\ell)\frac{\Delta}{\delta^2}}\\ &\le \frac{3}{1-\mathrm{e}^{-4\frac{\Delta}{\delta^2}}} + \frac{6}{\left(1-\mathrm{e}^{-2\frac{\Delta}{\delta^2}}\right)^2} \le \frac{9}{\left(1-\mathrm{e}^{-2\frac{\Delta}{\delta^2}}\right)^2}.\end{aligned}$$

Therefore we deduce (4.11), because the denominator cancels out with the factor of the second term in the round brackets in (B.1). In order to derive (4.12), we use Hölder's inequality and (4.11):

$$\begin{aligned}\mathbf{E}\left|C(\hat{\Upsilon}_n) - C(\hat{\Xi}_n)\right|^2 &\le C\sum_{j=1}^{J}\mathbf{E}\left|\hat{v}_n^{(j)}\otimes\hat{v}_n^{(j)} - \hat{\xi}_n^{(j)}\otimes\hat{\xi}_n^{(j)}\right|^2\\ &\le C\sum_{j=1}^{J}\mathbf{E}\left|\hat{v}_n^{(j)}\otimes(\hat{v}_n^{(j)}-\hat{\xi}_n^{(j)}) - (\hat{\xi}_n^{(j)}-\hat{v}_n^{(j)})\otimes\hat{\xi}_n^{(j)}\right|^2\\ &\le C\sum_{j=1}^{J}\mathbf{E}\left(\left(\left|\hat{v}_n^{(j)}\right|^2 + \left|\hat{\xi}_n^{(j)}\right|^2\right)\left|\hat{v}_n^{(j)}-\hat{\xi}_n^{(j)}\right|^2\right)\\ &\le C\sum_{j=1}^{J}\mathbf{E}\left(\left|\hat{v}_n^{(j)}-\hat{\xi}_n^{(j)}\right|^4\right)^{1/2} \le C\,\mathrm{e}^{-2(M+1)\frac{\Delta}{\delta^2}},\end{aligned}$$

which is the required bound.

REFERENCES

[1] A. Abdulle, G. A. Pavliotis, and U. Vaes. Spectral methods for multiscale stochastic differential equations. *SIAM/ASA J. Uncertain. Quantif.*, 5(1):720–761, 2017.

[2] S. Agapiou, M. Burger, M. Dashti, and T. Helin. Sparsity-promoting and edge-preserving maximum *a posteriori* estimators in non-parametric Bayesian inverse problems. *Inverse Problems*, 34(4):045002, 37, 2018.

[3] V. Akcelik, G. Biros, and O. Ghattas. Parallel multiscale Gauss–Newton–Krylov methods for inverse wave propagation. In *SC'02: Proceedings of the 2002 ACM/IEEE Conference on Supercomputing*, pages 41–41. IEEE, 2002.

[4] J. Barzilai and J. M. Borwein. Two-point step size gradient methods. *IMA J. Numer. Anal.*, 8(1):141–148, 1988.

[5] A. Borovykh, N. Kantas, P. Parpas, and G. A. Pavliotis. On stochastic mirror descent with interacting particles: convergence properties and variance reduction. *arXiv e-prints*, July 2020.

[6] A. Borovykh, N. Kantas, P. Parpas, and G. A. Pavliotis. To interact or not? the convergence properties of interacting stochastic mirror descent. In *International Conference on Machine Learning (ICML) Workshop on 'Beyond first order methods in machine learning systems'*, 2020.

[7] N. Bou-Rabee and E. Vanden-Eijnden. Pathwise accuracy and ergodicity of metropolized integrators for SDEs. *Comm. Pure Appl. Math.*, 63(5):655–696, 2010.

[8] J. A. Carrillo, Y.-P. Choi, C. Totzeck, and O. Tse. An analytical framework for consensus-based global optimization method. *Math. Models Methods Appl. Sci.*, 28(6):1037–1066, 2018.

[9] J. A. Carrillo, F. Hoffmann, A. M. Stuart, and U. Vaes. Consensus-based sampling, 2021 (in preparation).

[10] J. A. Carrillo and U. Vaes. Wasserstein stability estimates for covariance-preconditioned Fokker-Planck equations. *arXiv e-prints*, October 2019.

[11] P. Chaudhari, A. Oberman, S. Osher, S. Soatto, and G. Carlier. Deep relaxation: partial differential equations for optimizing deep neural networks. *Res. Math. Sci.*, 5(3):Paper No. 30, 30, 2018.

[12] C. Clason, T. Helin, R. Kretschmann, and P. Piiroinen. Generalized modes in Bayesian inverse problems. *SIAM/ASA J. Uncertain. Quantif.*, 7(2):652–684, 2019.

[13] M. Dashti, K. J. H. Law, A. M. Stuart, and J. Voss. MAP estimators and their consistency in Bayesian nonparametric inverse problems. *Inverse Problems*, 29(9):095017, 27, 2013.

[14] M. Dashti and A. M. Stuart. The Bayesian approach to inverse problems. In *Handbook of uncertainty quantification. Vol. 1, 2, 3*, pages 311–428. Springer, Cham, 2017.

[15] P. Del Moral. *Mean field simulation for Monte Carlo integration*, volume 126 of *Monographs on Statistics and Applied Probability*. CRC Press, Boca Raton, FL, 2013.

[16] O. R. A. Dunbar, A. Garbuno-Inigo, T. Schneider, and A. M. Stuart. Calibration and Uncertainty Quantification of Convective Parameters in an Idealized GCM. *arXiv e-prints*, December 2020.

[17] W. E, D. Liu, and E. Vanden-Eijnden. Analysis of multiscale methods for stochastic differential equations. *Comm. Pure Appl. Math.*, 58(11):1544–1585, 2005.

[18] O. G. Ernst, B. Sprungk, and H. Starkloff. Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):823–851, 2015.

[19] G. Evensen. *Data assimilation*. Springer-Verlag, Berlin, second edition, 2009. The ensemble Kalman filter.

[20] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart. Interacting Langevin diffusions: gradient structure and ensemble Kalman sampler. *SIAM J. Appl. Dyn. Syst.*, 19(1):412–441, 2020.

[21] A. Garbuno-Inigo, N. Nüsken, and S. Reich. Affine invariant interacting Langevin dynamics for Bayesian inference. *SIAM Journal on Applied Dynamical Systems*, 19(3):1633–1658, 2020.

[22] J. Goodman and J. Weare. Ensemble samplers with affine invariance. *Commun. Appl. Math. Comput. Sci.*, 5(1):65–80, 2010.

[23] C. Graham and D. Talay. *Stochastic simulation and Monte Carlo methods*, volume 68 of *Stochastic Modelling and Applied Probability*. Springer, Heidelberg, 2013. Mathematical foundations of stochastic simulation.

[24] S. Gratton, A. S. Lawless, and N. K. Nichols. Approximate Gauss-Newton methods for nonlinear least squares problems. *SIAM J. Optim.*, 18(1):106–132, 2007.

[25] E. Haber, F. Lucka, and L. Ruthotto. Never look back - A modified EnKF method and its application to the training of neural networks without back propagation. *arXiv e-prints*, May 2018.

[26] M. Hanke. A regularizing Levenberg-Marquardt scheme, with applications to inverse groundwater filtration problems. *Inverse Problems*, 13(1):79–95, 1997.

[27] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[28] T. Helin and M. Burger. Maximum *a posteriori* probability estimates in infinite-dimensional Bayesian inverse problems. *Inverse Problems*, 31(8):085009, 22, 2015.

[29] T. Helin and M. Lassas. Hierarchical models in statistical inverse problems and the Mumford-Shah functional. *Inverse Problems*, 27(1):015008, 32, 2011.

[30] M. Herty and G. Visconti. Kinetic Methods for Inverse Problems. *Kinetic & Related Models*, 12(1937-5093-2019-5-1109):1109, 2019.

[31] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Research Nat. Bur. Standards*, 49:409–436 (1953), 1952.

[32] D. Z. Huang, T. Schneider, and A. M. Stuart. Unscented Kalman Inversion, 2021 (in preparation).

[33] M. A. Iglesias, K. J. H. Law, and A. M. Stuart. Ensemble Kalman methods for inverse problems. *Inverse Problems*, 29(4):045001, 20, 2013.

[34] B. Jourdain, T. Lelièvre, and B. Miasojedow. Optimal scaling for the transient phase of the random walk Metropolis algorithm: the mean-field limit. *Ann. Appl. Probab.*, 25(4):2263–2300, 2015.

[35] J. Kaipio and E. Somersalo. *Statistical and computational inverse problems*, volume 160 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2005.

[36] N. Kantas, P. Parpas, and G. A. Pavliotis. The sharp, the flat and the shallow: Can weakly interacting agents learn to escape bad minima? *arXiv e-prints*, May 2019.

[37] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of ICNN'95-International Conference on Neural Networks*, volume 4, pages 1942–1948. IEEE, 1995.

[38] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[39] Peter E. Kloeden and Eckhard Platen. *Numerical solution of stochastic differential equations*, volume 23 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1992.

[40] N. B. Kovachki and A. M. Stuart. Ensemble Kalman inversion: a derivative-free technique for machine learning tasks. *Inverse Problems*, 35(9):095005, 35, 2019.

[41] N. B. Kovachki and A. M. Stuart. Ensemble kalman inversion: a derivative-free technique for machine learning tasks. *Inverse Problems*, 35(9):095005, aug 2019.

[42] H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 2003. Stochastic Modelling and Applied Probability.

[43] B. Leimkuhler and C. Matthews. *Molecular dynamics*, volume 39 of *Interdisciplinary Applied Mathematics*. Springer, Cham, 2015. With deterministic and stochastic numerical methods.

[44] B. Leimkuhler, C. Matthews, and J. Weare. Ensemble preconditioning for Markov chain Monte Carlo simulation. *Stat. Comput.*, 28(2):277–290, 2018.

[45] B. Leimkuhler and M. Sachs. Ergodic properties of quasi-Markovian generalized Langevin equations with configuration dependent noise and non-conservative force. In G. Giacomin, S. Olla, E. Saada, H. Spohn, and G. Stoltz, editors, *Stochastic dynamics out of equilibrium*, volume 282 of *Springer Proc. Math. Stat.*, pages 282–330. Springer, Cham, 2019.

[46] T. Lelièvre and G. Stoltz. Partial differential equations and stochastic methods in molecular dynamics. *Acta Numer.*, 25:681–880, 2016.

[47] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quart. Appl. Math.*, 2:164–168, 1944.

[48] Q. Liu and D. Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pages 2378–2386, 2016.

[49] J. Lu, Y. Lu, and J. Nolen. Scaling limit of the Stein variational gradient descent part i: the mean field regime. *arXiv preprint arXiv:1805.04035*, 2018.

[50] X. Mao. *Stochastic differential equations and applications*. Horwood Publishing Limited, Chichester, second edition, 2008.

[51] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Indust. Appl. Math.*, 11:431–441, 1963.

[52] J. C. Mattingly, A. M. Stuart, and M. V. Tretyakov. Convergence of numerical time-averaging and stationary measures via Poisson equations. *SIAM J. Numer. Anal.*, 48(2):552–577, 2010.

[53] N. Metropolis, A. W Rosenbluth, M. N Rosenbluth, A. H Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953.

[54] M. Ottobre and G. A. Pavliotis. Asymptotic analysis for the generalized Langevin equation. *Nonlinearity*, 24(5):1629–1653, 2011.

[55] E. Pardoux and A. Yu. Veretennikov. On the Poisson equation and diffusion approximation. I. *Ann. Probab.*, 29(3):1061–1085, 2001.

[56] G. A. Pavliotis. *Stochastic processes and applications*, volume 60 of *Texts in Applied Mathematics*. Springer, New York, 2014. Diffusion processes, the Fokker-Planck and Langevin equations.

[57] G. A. Pavliotis, G. Stoltz, and U. Vaes. Scaling Limits for the Generalized Langevin Equation. *J. Nonlinear Sci.*, 31(1):Paper No. 8, 2021.

[58] G. A. Pavliotis and A. M. Stuart. White noise limits for inertial particles in a random field. *Multiscale Model. Simul.*, 1(4):527–533, 2003.

[59] G. A. Pavliotis and A. M. Stuart. *Multiscale methods*, volume 53 of *Texts in Applied Mathematics*. Springer, New York, 2008. Averaging and homogenization.

[60] N. Petra, J. Martin, G. Stadler, and O. Ghattas. A computational framework for infinite-dimensional Bayesian inverse problems, Part II: Stochastic Newton MCMC with application to ice sheet flow inverse problems. *SIAM J. Sci. Comput.*, 36(4):A1525–A1555, 2014.

[61] R. Pinnau, C. Totzeck, O. Tse, and S. Martin. A consensus-based model for global optimization and its mean-field limit. *Math. Models Methods Appl. Sci.*, 27(1):183–204, 2017.

[62] J. Quer. *Importance sampling for metastable dynamical systems in molecular dynamics*. PhD thesis, Feie Universität Berlin, 2018.

[63] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statistics*, 22:400–407, 1951.

[64] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms.

Ann. Appl. Probab., 7(1):110–120, 1997.

[65] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.*, 16(4):351–367, 2001.

[66] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[67] C. Schillings and A. M. Stuart. Analysis of the ensemble Kalman filter for inverse problems. *SIAM J. Numer. Anal.*, 55(3):1264–1290, 2017.

[68] C. Schillings and A. M. Stuart. Convergence analysis of ensemble Kalman inversion: the linear, noisy case. *Appl. Anal.*, 97(1):107–123, 2018.

[69] A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numer.*, 19:451–559, 2010.

[70] Z. Wu. The effective energy transformation scheme as a special continuation approach to global optimization with application to molecular conformation. *SIAM J. Optim.*, 6(3):748–768, 1996.

[71] S. Zhang, A. E. Choromanska, and Y. LeCun. Deep learning with elastic averaging sgd. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 685–693. Curran Associates, Inc., 2015.