© 2022 Society for Industrial and Applied Mathematics

Derivative-Free Bayesian Inversion Using Multiscale Dynamics[*]

G. A. Pavliotis[†], A. M. Stuart[‡], and 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 [A. Garbuno-Inigo et al., *SIAM J. Appl. Dyn. Syst.*, 19 (2020), pp. 412–441]. 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 estimation, and may be systematically refined. The method relies on a fast/slow system of stochastic differential equations 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, while also being provably refinable. 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

DOI. 10.1137/21M1397416

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

$$(1.1) \qquad\qquad y = G(\theta) + \eta,$$

with $G : \mathbf{R}^d \to \mathbf{R}^K$ a forward operator and η the observational noise. In the Bayesian approach to inverse problems [48, 87, 23], the vectors θ, η, and y are treated as random variables. If the

[†]Department of Mathematics, Imperial College London, London, SW7 2AZ, UK (g.pavliotis@imperial.ac.uk).
[‡]Department of Computing and Mathematical Sciences, Caltech, Pasadena, CA 91125 USA (astuart@caltech.edu).
[§]MATHERIALS Team, Inria Paris, 75589, Paris, CEDEX 12, France (urbain.vaes@inria.fr).

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):

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

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

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

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

with the notation, for a symmetric positive definite matrix A,

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

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

$$(1.3) \qquad\qquad \pi(\theta) = \frac{\exp\left(-\Phi_R(\theta;y)\right)}{\int_{\mathbf{R}^d} \exp\left(-\Phi_R(\theta;y)\right) d\theta} =: \frac{1}{Z(y)} \exp\left(-\Phi_R(\theta;y)\right).$$

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 stochastic differential equations (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) [45] and consensus-based sampler (CBS) [15], 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

$$(1.4) \qquad\qquad d\theta_t = -K \nabla_\theta \Phi_R(\theta_t; y) \, dt + \nu \sqrt{2K} \, dw_t.$$

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 $\nu = 0$, (1.4) is a preconditioned gradient descent in the potential Φ_R, and, when $\nu = 1$, (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 [48, 87]. Classical methods are generally based on an optimization problem of the form

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

for some symmetric 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, 19, 22, 41, 42] 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 $R(\theta) = |\theta - z|^2_Y$ for some vector z and a symmetric positive definite matrix Y. 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 $R(\theta) = |\theta - m|^2_\Sigma$. 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 [87] 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 [26], 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 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 [44], the stochastic gradient descent [80, 54], the Barzilai–Borwein method [8], and other gradient-based optimization techniques that rely on interacting particle systems [90, 11, 10]. One may also recur to the Newton or Gauss–Newton methods and their variants [36, 3], or to methodologies based on the Levenberg–Marquardt method [61, 65, 39].

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 comprehensive presentation of standard derivative-free optimization methods is given in [20], and a review focusing on recent trends and developments in this field is given in [55]. A simple approach, which was popular in many of the early works on derivative-free optimization, is to employ a finite difference gradient approximation in place of the exact gradient in a derivative-based method; see, for example, [13, 33]. Many other general-purpose derivative-free optimization

methods can be employed for solving (1.5), and we mention, for example, simulated annealing [51], particle-swarm optimization [50], and consensus-based optimization (CBO) [76, 14]. One may also use methods based on the ensemble Kalman filter, which rely on the quadratic structure of the loss function in (1.5). Ensemble Kalman methods, introduced for dynamical state estimation in [30], were extended as derivative-free Bayesian inverse problem solvers in [18, 28]; they were modified to become derivative-free optimizers in [46], a method we refer to as ensemble Kalman inversion (EKI). EKI has been shown, both theoretically in simple settings and empirically, to perform very well in the context of inverse problems [46, 85, 86], and it has also been applied successfully for training neural networks [53]. A recent variant on 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 [45]. Another alternative is the method developed in [38], which is based on similar ideas for gradient approximation but aims to drive a single distinguished particle to the optimizer.

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 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 [60, 71] or the generalized Langevin dynamics [69, 59, 72].

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 (MH) algorithm [67, 40]. All that is required to define an 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 computing, or exploiting, Gaussian approximations of the posterior; see [75] and the references therein. There is an extensive literature on the convergence properties and optimal parametrization of these methods, and on their connections with overdamped Langevin diffusions in the high-dimensional limit [83, 81, 82]. See also [47] for a study of the connection of high-dimensional RWMH with overdamped Langevin dynamics in the transient regime, and [12] for a proof of convergence of MALA to an overdamped Langevin diffusion in the small time step 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, sequential Monte Carlo samplers [25], interacting particle MCMC methodologies [24, 58], Stein variational gradient descent [63, 62], EKS [31], and affine-invariant Langevin dynamics (ALDI) [32]. 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. Recently, a variant on EKS using the unscented transform (UKS) [45] was proposed, as was a generalization of CBO to sampling (CBS) [15]; both UKS and CBS are derivative-free.

The parallel MCMC method of [58], 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 [34]; see also [37] and [56]. 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 Φ_R exhibits strong anisotropy at its minimizer. The affine invariance of EKS, ALDI, and the ensemble Kalman–Bucy filter was demonstrated carefully in [32], where the authors also show that the Bayesian posterior is invariant and ergodic under ALDI. Around the same time, it was observed that the rate of convergence to equilibrium for the nonlocal PDEs associated with EKI and EKS was independent of the parameters of the regularized least-squares functional Φ_R, in the simple case of a linear forward model [31, 16]; this independence is in fact a consequence of affine invariance, although this fact is not identified in these references.

As mentioned in subsection 1.1, the method we present in this paper is based on a fast/slow system of SDEs, 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 in [17] for smoothing the loss function associated with deep neural networks, and the method is revisited later in [49]. A similar multiscale dynamics is also employed in [77] for calculating convolutions, with the aim of reducing metastability in the context of molecular dynamics. See also [89] 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 of 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 also show how a good preconditioner K can be constructed using information from ALDI; similar approaches can be used based on information obtained through EKS, UKS, or CBS.

Our work is aimed at sampling posterior distributions which are not Gaussian. The paper [29] demonstrates clearly that standard ensemble Kalman–based methods for inverse problems do not reproduce the correct posterior distribution in the large particle size limit in the non-Gaussian setting. Although there is interesting empirical work which addresses this shortcoming through iteration of ensemble Kalman filters [84, 9], it is not clear that this methodology may be applied systematically to arbitrary inverse problems. Our proposed methodology, on the other hand, addresses shortcomings of standard ensemble methods in this setting and is founded on refinable approximations which, in certain limits, will reproduce the true posterior distribution.

Although our method applies, in principle, to multimodal distributions, it will not be efficient in this scenario; this is because our method is based on approximation of an over-

damped Langevin equation, and hence may suffer from metastability issues when the posterior is multimodal. For this reason our focus is on non-Gaussian unimodal distributions; our primary purpose in this paper is not to address multimodality. Indeed it is likely that a generic solution to the problem of multimodality will be very hard to find [88]—problem-specific multimodal approaches are more likely to yield fruitful research. There is, however, interesting work by [57] which uses a localized sample covariance matrix in a parallel MCMC method to address this problem in a generic fashion; we briefly touch on this in subsection 2.4. The idea of using localized covariances is incorporated into ALDI in [79] in order to extend the range of applicability of the method beyond the unimodal setting.

1.3. Our contributions. This paper, then, is focused on the construction of provably refinable derivative-free methods for Bayesian inverse problems characterized by unimodal but non-Gaussian posterior distributions. The primary 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 ALDI.

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 into the method. Section 4 is reserved for conclusions and perspectives for future work. In Appendix A, we prove our main convergence results, and in Appendices B and C, we obtain auxiliary technical results.

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 Appendix A.

2.1. Continuous-time dynamics. Our method is based on a multiscale system of 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 [85] (in the context of optimization) and [31] (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 [38], in which the authors proposed using 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 Appendix A, 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:

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

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

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

where $\theta \in \mathbf{K}^d$, $\Xi = (\xi^{(1)}, \ldots, \xi^{(J)}) \in (\mathbf{R}^d)^J$, 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 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 π 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 nonregularized 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 overdetermined. 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:

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

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

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

$$= -C(\Xi)\, \nabla \Phi_R(\theta) + \nu\sqrt{2C(\Xi)}\, \dot{w}.$$

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 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 $\nu = 1$), (2.1) admits as invariant measure the distribution

$$(2.3) \qquad \rho_\infty(\theta, \Xi) = \pi(\theta)\, g(\xi^{(1)}) \ldots g(\xi^{(J)}),$$

where g denotes the density of the standard normal distribution. The associated marginal distribution for θ is given by the Bayesian posterior distribution π. 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 [71, Chapter 4]

$$\mathcal{L}^\dagger \rho = \nabla_\theta \cdot \left(C(\Xi)\left(\nabla \Phi_R(\theta)\rho + \nabla_\theta \rho \right) \right) + \frac{1}{\delta^2} \sum_{j=1}^{J} \nabla_{\xi^{(j)}} \cdot \left(\xi^{(j)}\rho + \nabla_{\xi^{(j)}}\rho \right)$$

$$= \nabla_\theta \cdot \left(\rho_\infty C(\Xi)\nabla_\theta \left(\frac{\rho}{\rho_\infty} \right) \right) + \frac{1}{\delta^2} \sum_{j=1}^{J} \nabla_{\xi^{(j)}} \cdot \left(\rho_\infty \nabla_{\xi^{(j)}} \left(\frac{\rho}{\rho_\infty} \right) \right) = \nabla \cdot \left(\rho_\infty \mathcal{D} \nabla \log \left(\frac{\rho}{\rho_\infty} \right) \right),$$

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 $\xi^{(1)}, \ldots, \xi^{(J)}$:

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

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

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

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

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

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

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

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 Ornstein–Uhlenbeck 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

$$\text{(2.6a)} \qquad \hat{\theta}_{n+1} = \hat{\theta}_n - \frac{1}{J\sigma} \sum_{j=1}^{J} \langle G(\hat{\theta}_n + \sigma\hat{\xi}_n^{(j)}) - G(\hat{\theta}_n), G(\hat{\theta}_n) - y \rangle_{\Gamma}\, \hat{\xi}_n^{(j)}\,\Delta$$

$$- C(\hat{\Xi}_n)\Sigma^{-1}(\hat{\theta}_n - m)\Delta + \nu\sqrt{2D(\hat{\Xi}_n)\Delta}\,x_n,$$

$$\text{(2.6b)} \qquad \hat{\xi}_{n+1}^{(j)} = e^{-\frac{\Delta}{\delta^2}}\,\hat{\xi}_n^{(j)} + \sqrt{1 - e^{-\frac{2\Delta}{\delta^2}}}\,x_n^{(j)}, \qquad \hat{\xi}_0^{(j)} \sim \mathcal{N}(0, I_d), \qquad j = 1, \ldots, J,$$

where $x_n = \Delta^{-1/2}(w_{(n+1)\Delta} - w_{n\Delta}) \sim \mathcal{N}(0, I_d)$ and $x_n^{(j)}$, for $n = 0, \ldots, N-1$ and $j = 1, \ldots, J$, 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

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

where $\hat{\xi}_n^{(j)}$, for $n = 0, \ldots, N-1$ and $j = 1, \ldots, J$, are drawn independently from $\mathcal{N}(0, I_d)$. In the noise-free case $\nu = 0$, the algorithm in the form (2.7) is precisely what was proposed and implemented in the paper [38], in the context of optimizing parameters of neural networks, and inspired the work presented here. This algorithm is simpler to implement and analyze than (2.6), but may not always be the best option for the purposes of sampling and optimization. Although we have found small values of δ to be preferable in our numerical experiments, it is indeed conceivable that problems with rugged energy landscapes may benefit from stronger correlation between successive descent directions. We leave this question for future work and, in all the numerical experiments presented in section 3, we use the scheme (2.6) with $\delta > 0$ (but in most examples very small). We do, however, analyze the convergence of (2.7) theoretically in Appendix A, 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 Appendix A. In order to simplify the analysis, we assume throughout that the state space of the parameter is the d-dimensional torus $\mathbf{T}^d$, rather than $\mathbf{R}^d$, and that the prior distribution is the uniform density over $\mathbf{T}^d$, in which case $\Phi_R = \Phi$. 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 $G \in C^3(\mathbf{T}^d, \mathbf{R}^K)$. In particular, the case of a linear forward model is excluded.

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

$$
(2.8) \qquad \dot{\vartheta}_t = -\nabla \Phi_R(\vartheta_t) + \nu \sqrt{2} \, \dot{w}_t, \qquad \vartheta_0 = \theta_0.
$$

Note that the Brownian motion and initial condition in this equation are the same as in (2.4), which allows us to establish a strong convergence estimate.

Theorem 2.1. *Assume that* $\{\theta_t, \xi_t^{(1)}, \ldots, \xi_t^{(J)}\}_{t \in [0,T]}$ *is a solution to* (2.4) *supplemented with any initial condition* $(\theta_0, \xi_0^{(1)}, \ldots, \xi_0^{(J)})$ *such that*

$$
(2.9) \qquad (\xi_0^{(1)}, \ldots, \xi_0^{(J)}) \sim \mathcal{N}(0, I_d) \times \cdots \times \mathcal{N}(0, I_d).
$$

If $G \in C^2(\mathbf{T}^d, \mathbf{R}^K)$, *then for any* $p > 1$, *any* $J > 0$, *and any* $T > 0$ *there is* $C = C(p, T, J)$ *such that*

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

The exponent β is defined as follows:

$$(2.11) \qquad \beta = \begin{cases} 1 & if\ G \in C^2(\mathbf{T}^d, \mathbf{R}^K), \\ 2 & if\ G \in C^3(\mathbf{T}^d, \mathbf{R}^K). \end{cases}$$

Remark 2.4. It should in principle be possible, but may prove technically challenging, to extend our findings to $\mathbf{R}^d$, Gaussian priors, and unbounded vector fields by using results from [70]. Furthermore, although the convergence result is over a finite time interval, it is to be expected that convergence of invariant measures might also be established, using (for example) the ideas in [66] or [6, 7]. In this regard we notice that we have focused on pathwise convergence on finite time intervals; typically only weak convergence results would be required to obtain convergence of the invariant measure. Indeed weak convergence results will be needed to study the formulation from Remark 2.3 because the noise is multiplicative.

Remark 2.5. 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 (2.10), 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). We note that a weaker metric is employed in this result than in (2.10).

Theorem 2.2. *Assume that $\{\hat{\theta}_n, \hat{\xi}_n^{(1)}, \ldots, \hat{\xi}_n^{(J)}\}_{n=0}^N$ is a solution to (2.6) with $D(\bullet) = I_d$ and an initial condition $(\theta_0, \xi_0^{(1)}, \ldots, \xi_0^{(J)})$ satisfying (2.9). If $G \in C^2(\mathbf{T}^d, \mathbf{R}^K)$, then for any $J > 0$ and any $T > 0$, there exists a constant $C = C(T, J)$ such that*

$$(2.12) \qquad \forall (\delta, \sigma, \Delta) \in \mathbf{R}_+^3, \qquad \sup_{0 \le n \le T/\Delta} \mathbf{E} \left| \hat{\theta}_n - \vartheta_{n\Delta} \right|^2 \le C \left(\Delta + \sigma^{2\beta} + \log(1 + \delta^{-1}) \delta^2 \right).$$

The exponent β is defined as in (2.11).

Remark 2.6. 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) in order to differentiate it from the solution to (2.6), it holds by the triangle inequality and (2.12) that, for all $(\delta, \sigma, \Delta) \in \mathbf{R}_+^3$ and all $\varepsilon > 0$,

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

Here we used that, by Young's inequality, it holds that $(a + b)^2 \leq (1 + \varepsilon)a^2 + \left(1 + \frac{1}{\varepsilon}\right)b^2$ for any $\varepsilon > 0$ and any $a, b \in \mathbf{R}$. It is clear that if $\hat{\xi}_0^{(j)}, \hat{\xi}_1^{(j)}, \hat{\xi}_2^{(j)}, \ldots$ in (2.7) coincide with $\hat{\xi}_0^{(j)}, x_0^{(j)}, x_1^{(j)}, \ldots$ in (2.6) for all $1 \leq j \leq J$, then the first term on the right-hand side vanishes in the limit as $\delta \to 0$. Therefore, letting $\delta \to 0$ and then $\varepsilon \to 0$, we deduce

$$(2.13) \qquad \forall (\sigma, \Delta) \in \mathbf{R}_+ \times \mathbf{R}_+, \qquad \sup_{0 \leq n \leq N} \mathbf{E} \left| \tilde{\theta}_n - \vartheta_{n\Delta} \right|^2 \leq C \left(\Delta + \sigma^{2\beta} \right),$$

where $C = C(T, J)$ is the same constant as in (2.12). In Appendix A, for clarity of exposition, we will in fact first prove the convergence estimate (2.13) before showing the more general Theorem 2.2.

Remark 2.7. In the limit $\Delta \to 0$, the error bound (2.12) becomes

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

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.

2.4. 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, compared to the time scale of 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 stringent constraints on the time step in order to guarantee stability, leading to often prohibitive computational costs.

Empirically, we observe—see subsection 3.3—that our multiscale method suffers from a similar issue, which is not surprising given that (2.4) converges to (2.8) as $(\delta, \sigma) \to (0, 0)$ by Theorem 2.1. 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 [32]: 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 descent dynamics or overdamped Langevin diffusions preconditioned by the covariance of the ensemble, which provides good stability and convergence properties [31, 16].

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:

(2.14a)

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

(2.14b)
$$\theta^{(j)} = \theta + \sigma\sqrt{K}\xi^{(j)}, \qquad\qquad\qquad\qquad j = 1, \dots, J,$$

(2.14c)

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

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 (2.14a) 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 (2.14), 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 $\mathcal{N}(\sqrt{K^{-1}}m, \sqrt{K^{-1}}\Sigma\sqrt{K^{-1}})$. In view of this connection, Theorems 2.1 and 2.2 apply mutatis mutandis to the dynamics (2.14) with $D_K(\Xi) = K$.

We motivate in subsection 3.3 that when the forward model is linear, a good preconditioning matrix K is given by the covariance of the Bayesian posterior. In practice, we observed that preconditioning with the posterior covariance works well also for nonlinear forward models, provided that the posterior distribution is unimodal. 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 gradient-free ALDI (gfALDI) approach proposed in [31], which enables generating approximate samples from the Bayesian posterior. The ALDI method is based on the dynamics

$$\dot{\theta}^{(\ell)} = -\frac{1}{L}\sum_{k=1}^{L}\langle G(\theta^{(k)}) - \bar{G}, G(\theta^{(\ell)}) - y\rangle_\Gamma(\theta^{(k)} - \bar{\theta})$$

$$(2.15) \qquad\qquad -C(\Theta)\Sigma^{-1}(\theta^{(\ell)} - m) + \frac{d+1}{L}(\theta^\ell - \bar{\theta}) + \sqrt{2C(\Theta)}\,\dot{w}^{(\ell)}, \qquad \ell = 1, \dots, L,$$

where $\Theta = (\theta^{(1)}, \ldots, \theta^{(L)})$ 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 in $\mathbf{R}^L$, $C(\Theta) = \frac{1}{L} \sum_{\ell=1}^{L} (\theta^{(\ell)} - \bar{\theta}) \otimes (\theta^{(\ell)} - \bar{\theta})$ is the covariance matrix associated with the ensemble, and $\sqrt{C(\Theta)}$ is the $\mathbf{R}^{d \times L}$ matrix defined by

$$\sqrt{C(\Theta)} = \frac{1}{\sqrt{L}} \left[(\theta^{(1)} - \bar{\theta}) \quad \ldots \quad (\theta^{(L)} - \bar{\theta}) \right].$$

It holds that $C(\Theta) = \sqrt{C(\Theta)} \sqrt{C(\Theta)}^{\mathsf{T}}$, which motivates the square root notation. The first argument in the inner product on the right-hand side of (2.15) is related to consensus, whereas the second argument measures the mismatch with the observed data. In practice, the initial ensemble members are drawn independently from the Gaussian prior distribution. When the Bayesian posterior distribution is close to Gaussian, it is expected that each particle in the ensemble is an approximate sample from the posterior for sufficiently large times. Therefore, the expectation of an observable f with respect to π is approximated from ALDI iterates as

$$(2.16) \qquad \mathbf{E}_\pi f \approx \frac{1}{NL} \sum_{i=N_0+1}^{N_0+N} \sum_{\ell=1}^{L} \delta_{\hat{\theta}_i^{(\ell)}},$$

where the first N_0 iterations can be discarded for reducing the bias from initial conditions, and where $\{\hat{\theta}_i^{(\ell)}\}_{i \geq 0}$ is a discrete-time approximation of $\theta^{(\ell)}$ for $\ell \in \{1, \ldots, L\}$. The main difference between ALDI and EKS is the presence of the second-to-last term on the right-hand side of (2.15). It is shown by means of numerical experiments in [32] that this corrective drift term, which was first identified in [68], is crucial for accuracy when the number of particles L is of the same order of magnitude as the dimension of the state space d.

Although EKS and ALDI are self-preconditioned, these methods can suffer from stability issues for small times when the posterior distribution is far from or much more concentrated than the prior; this stiffness issue is discussed for the ensemble Kalman–Bucy filter in [4]. In practice, it is often useful to proceed in two steps: (i) first, run N_0 iterations with a small time step, until the ensemble reaches a region of high posterior probability; (ii) then use a larger step size for the rest of the simulation, and use only these iterations for the computation of averages with respect to the posterior distribution, as in (2.16). The matrix K could also be learned from the EKS, UKS, or CBS approaches to approximate sampling, rather than from ALDI.

Before closing this section, we note that the local preconditioning approach developed in [57, 79], based on localized covariance matrices, could potentially also be useful for improving the performance of our method. Once a rough approximation of the posterior has been calculated using (2.1), for example, self-preconditioning could be achieved through a localized covariance matrix constructed from all the previously generated samples that are in the vicinity of the distinguished particle. It may be worthwhile to explore this idea in future work, but in this paper we consider only preconditioning through a position-independent matrix K, as in (2.14).

3. Numerical experiments. In this section, we present numerical experiments demonstrating the performance and limitations of our method. Subsections 3.1 to 3.3 serve as proof of concept: in subsection 3.1 a toy inverse problem with low-dimensional parameter and data is presented, in subsection 3.2 an example with bimodal posterior is considered, and in subsection 3.3 preconditioning is exemplified. A more challenging simulated example, closer to inverse problems arising in real applications, is then considered in subsection 3.4.

3.1. Low-dimensional parameter space. We first consider the inverse problem with low-dimensional parameter space that was first presented in [29] and later employed as a test problem in [43, 31]. In this problem, the forward model maps the unknown $(u_1, u_2) \in \mathbf{R}^2$ to the observation $\big(p(x_1), p(x_2)\big) \in \mathbf{R}^2$, where $x_1 = 0.25$ and $x_2 = 0.75$ and where $p(x)$ denotes the solution to the boundary value problem

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

with boundary conditions $p(0) = 0$ and $p(1) = u_2$. This problem admits the following explicit solution [43]:

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

We employ the same parameters as in [31]: 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.
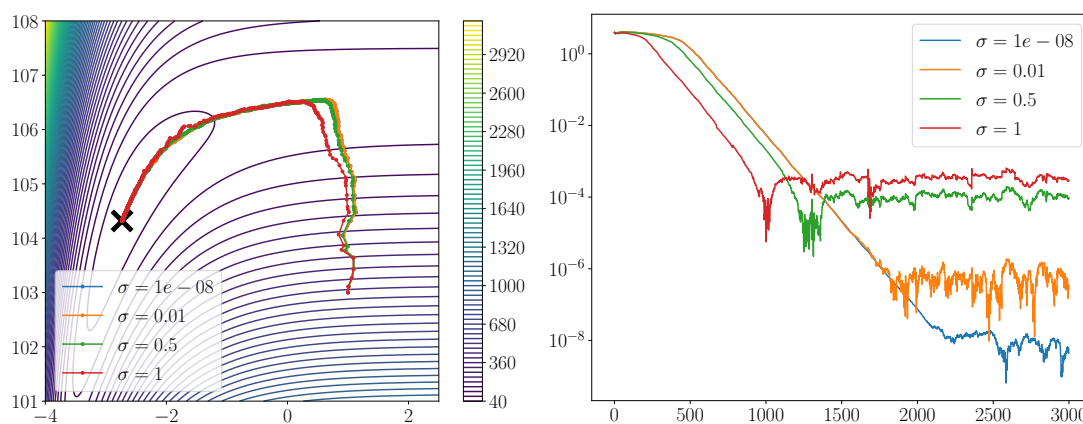
Figure 1. Left: *Trajectories of the numerical solution $\hat{\theta}_n$ for fixed $\delta = 10^{-8}$ and different values of σ.* Right: *Error $|\hat{\theta}_n - \theta_{\mathrm{MAP}}|_2$ along the trajectories, where θ_{MAP} is the MAP estimator.*
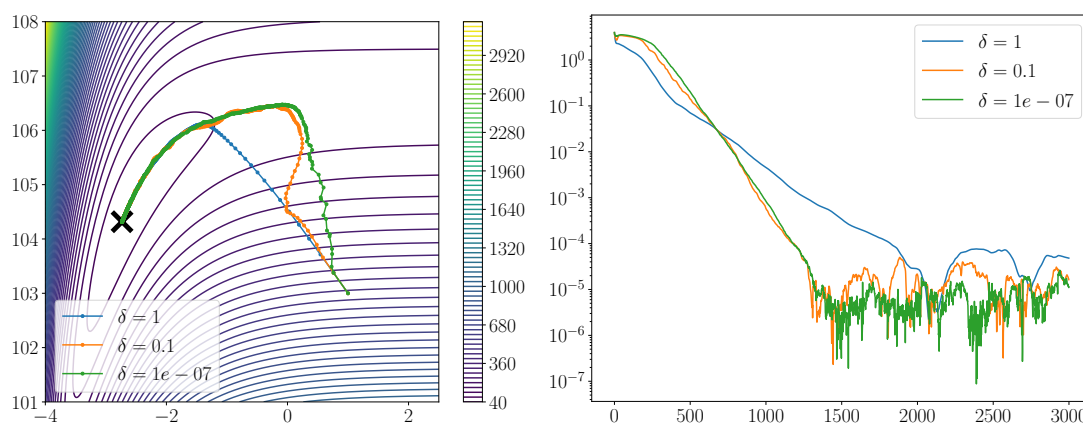

Figure 2. Left: *Trajectories of the numerical solution $\hat{\theta}_n$ for fixed $\sigma = 0.1$ and different values of δ.* Right: *Error $|\hat{\theta}_n - \theta_{\mathrm{MAP}}|_2$ along the trajectories, where θ_{MAP} is the MAP estimator.*

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 distributions, given, respectively, by

$$m_p = \begin{pmatrix} -2.714... \\ 104.346... \end{pmatrix}, \quad C_p = \begin{pmatrix} 0.0129... & 0.0288... \\ 0.0288... & 0.0808... \end{pmatrix},$$
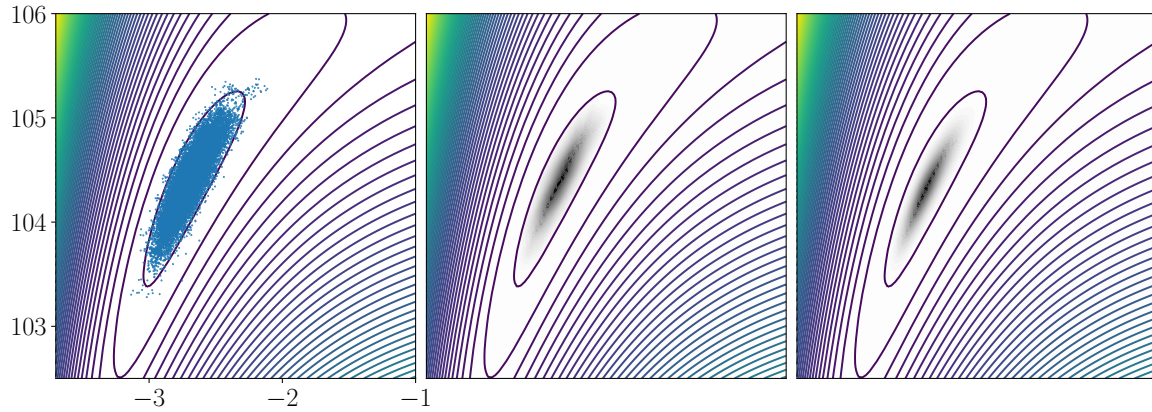
Figure 3. Left: *Iterates* 1,000 *to* 20,000. Middle: *Approximation of the Bayesian posterior based on these iterates, using kernel density estimation.* Right: *True Bayesian posterior.*

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. Two-dimensional bimodal example. In this section, we consider a bimodal example from [79], associated with forward model

$$G : \mathbf{R}^2 \ni \theta \mapsto |\theta_1 - \theta_2|^2 \in \mathbf{R},$$

noise distribution $\eta \sim \mathcal{N}(0, I_2)$, and prior distribution $\mathcal{N}(0, I_2)$. This example is employed in [79] for demonstrating the ability of an appropriately localized version of ALDI to sample from multimodal distributions. We consider two different values for the data: $y = 2$ and $y = 4.2297$, the latter value being the one used in [79]. In both cases, the posterior is bimodal and so ensemble Kalman methods fail to accurately capture the Bayesian posterior, but the energy barrier between the two modes is much higher when $y = 4.2297$, which makes this case challenging also for our method.

The approximate posteriors for $y = 2$ and $y = 4.2297$, obtained from 10^6 iterations of our method discretized using (2.6) with a time step $\Delta = 10^{-2}$, parameters $\delta = \sigma = 10^{-5}$, and $J = 8$ auxiliary particles, are depicted in Figures 4 and 5, respectively. Whereas gfALDI without localized covariance fails for both values of y, our method gives a very good approximation of the true posterior distribution when $y = 2$, which illustrates the strength of our method compared to ensemble Kalman–based methods without localization.

For $y = 4.2297$, however, the posterior distribution appears to be well approximated by our method within each of the two high-density regions, but the probabilities of these regions are not accurately captured: only 33% of the iterates are such that $\theta_2 - \theta_1 \geq 0$, whereas this fraction is 50% under the true posterior probability. This discrepancy is due to the strong metastability of the dynamics caused by bimodality, and indeed the line $\theta_2 - \theta_1 = 0$ was crossed by the iterates only 88 times during the simulation. For comparison, for $y = 2$ the
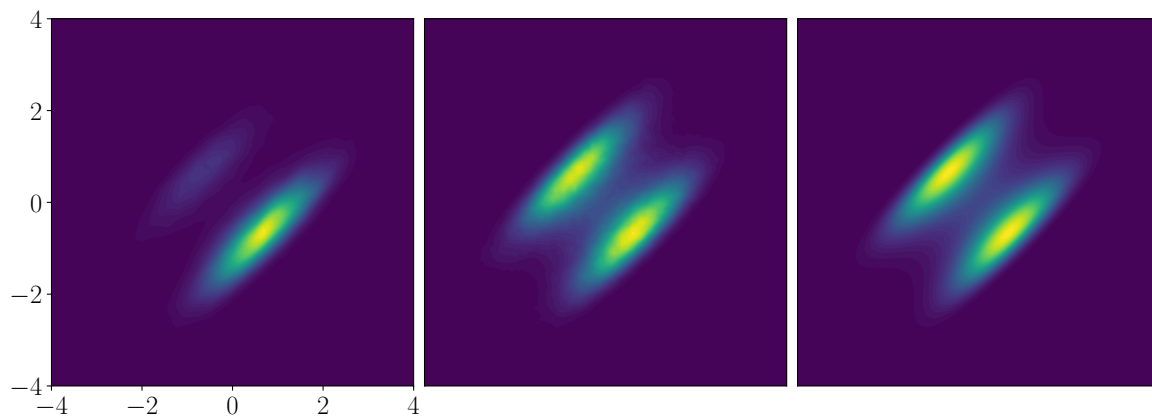
Figure 4. *All figures are for $y = 2$.* Left: *Approximation of the Bayesian posterior constructed by kernel density estimations from 10^6 samples generated with gfALDI.* Middle: *Approximation of the Bayesian posterior constructed from 10^6 iterations of the multiscale method.* Right: *True Bayesian posterior.*

Figure 5. *All figures are for $y = 4.2297$.* Left: *Approximation of the Bayesian posterior constructed by kernel density estimations from 10^6 samples generated with gfALDI.* Middle: *Approximation of the Bayesian posterior constructed from 10^6 iterations of the multiscale method.* Right: *True Bayesian posterior.*
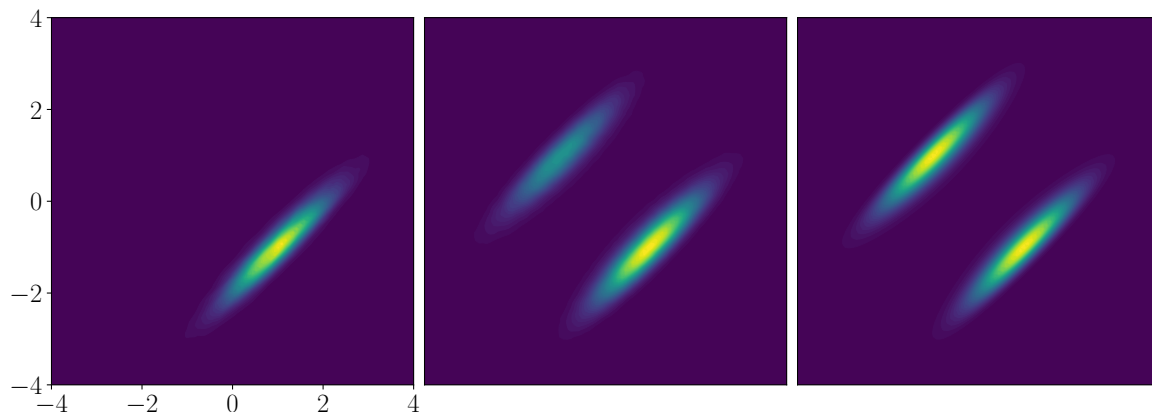
proportion of iterates such that $\theta_2 - \theta_1 \geq 0$ was 50.8%, and the line $\theta_2 - \theta_1 = 0$ was crossed 33,492 times. This example shows that strongly multimodal distributions are challenging for our method, which is not surprising given the connection with overdamped Langevin dynamics established in Theorem 2.1.

3.3. Toy example with preconditioning. We now illustrate the preconditioning methodology proposed in subsection 2.4 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: $y = (1, k, k^2)$, $\Sigma = +\infty I_3$ (that is, there is no prior regularization), and $\Gamma = I_3$, so that the MAP estimator is $\theta_{\mathrm{MAP}} = (1, 1, 1)$ and the

least-squares functional is given by

$$(3.2) \qquad \Phi(\theta; y) = \frac{1}{2} \left(|\theta_1 - y_1|^2 + k^2 |\theta_2 - y_2|^2 + k^4 |\theta_3 - y_3|^2 \right).$$

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 (3.2) 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 $\sigma \to 0$, 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 (2.14) in the limit as $\max(\delta, \sigma) \to 0$ is

$$(3.3) \qquad \mathrm{d}\theta_t = -(\theta_t - \theta_{\mathrm{MAP}}) \, \mathrm{d}t + \nu \sqrt{2K} \, \mathrm{d}w_t.$$

When employed for integrating this equation with $\nu = 0$, the explicit Euler scheme is stable for time steps satisfying $\Delta < 2$. In practice, we approximate the preconditioning parameter K from 100,000 iterations of gfALDI, run with 5 particles initialized independently as $\mathcal{N}(0, I_3)$ and discretized using the Euler–Maruyama method. The first 10,000 iterations are run with a small time step $\Delta = 10^{-3}$ to avoid stability issues, and they are discarded for the computation of the posterior covariance in order to reduce the bias originating from initial conditions. The remaining 90,000 iterations are run with a larger time step $\Delta = 10^{-2}$, which does not lead to stability issues in the later stages of the simulation thanks to self-preconditioning by the ensemble covariance. We employed $L = 5$ particles because it was shown in [32] that choosing $L > d + 1$ ensures that the continuous-time ALDI dynamics with gradients (which coincide with gradient-free ALDI dynamics in the case of a linear forward model) is ergodic with respect to the product measure $\pi \times \cdots \times \pi$, where π is the Bayesian posterior. Using this approach, we obtain a good approximation of the posterior covariance:

$$\widehat{K} = \begin{pmatrix} 1.00... & 0.00553... & -0.000302... \\ 0.00553... & 0.0399... & 6.48... \times 10^{-5} \\ -0.000302... & 6.48... \times 10^{-5} & 0.00154... \end{pmatrix}.$$

The effect of using this preconditioner is illustrated in Figure 6 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.
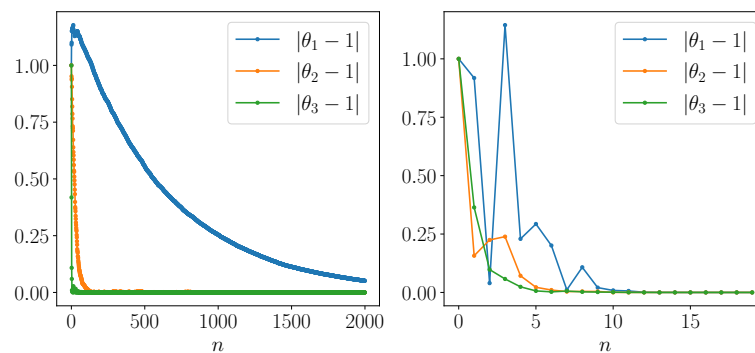
Figure 6. *Error between the iterates and the MAP estimator, without (left) and with (right) preconditioning. Here the label n of the x axis denotes the iteration index. We observe that a better approximation of the MAP estimator is obtained after 20 iterations with preconditioning, rather than after 2,000 iterations without preconditioning.*

In practice, the mean of the ensemble obtained after application of gfALDI is also useful; it can be employed as initial condition for (2.14). This is the approach taken in the next section.

3.4. Higher-dimensional parameter space. We now present an example from [31] 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 [23], 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 2.4.

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

$$-\nabla \cdot \left(e^{a(x)} \nabla p(x) \right) = f(x), \qquad\qquad x \in D, \tag{3.4a}$$

$$p(x) = 0, \qquad\qquad x \in \partial D. \tag{3.4b}$$

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

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

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 [31]: $\tau = 3$ and $\alpha = 2$. In this setting, it can be shown by reasoning as in [23, 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 (3.4) 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 $a(x)$ is almost surely in $L^2(D)$, we deduce that, almost surely,

$$(3.5) \qquad a = \sum_{\ell \in \mathbf{N}^2} (a, \psi_\ell)\,\psi_\ell =: \sum_{\ell \in \mathbf{N}^2} \sqrt{\lambda_\ell}\,\theta_\ell\,\psi_\ell,$$

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., [23] 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 taken by a finite number of continuous linear functionals $\ell_1, \dots, \ell_K$ over $H_0^1(D)$, when evaluated at the solution p to (3.4), perturbed by additive Gaussian noise. That is, the forward model maps the unknown permeability $a(\bullet)$ to

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

In practice, however, we consider that the data consist of pointwise measurements of the solution to (3.4), up to noise. We assume that these are taken at a finite number of equidistant points given by

$$(3.6) \qquad x_{ij} = \left(\frac{i}{M}, \frac{j}{M}\right), \qquad 1 \le i, j \le M - 1.$$

Since pointwise evaluation is not a continuous functional on $H_0^1(D)$, our example deviates here from the framework in [23]. As mentioned in [31], 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 (3.5) 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 (3.4) 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 $N = 7$, leading to a state space of dimension 64, and an FEM using quadratic elements over a regular mesh with 20 elements per direction, implemented with Gridap [5]. 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 using as many KL terms as in the numerical inference, i.e.,

$$(3.7) \qquad a^\dagger(x) = \sum_{|\ell|_\infty \leq N} \sqrt{\lambda_\ell} \theta_\ell^\dagger \psi_\ell, \qquad \theta_\ell^\dagger \sim \mathcal{N}(0,1).$$

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 7. The preconditioning matrix K is calculated using the methodology outlined in subsection 2.4 by running 200 iterations of gfALDI with an ensemble size equal to 512; the first 100 iterations are employed for transitioning from the prior to a rough approximation of the posterior, using the adaptive time-stepping scheme of [53], and then gfALDI is run with fixed time step for an additional 100 iterations, and only these iterations are used for the approximation of the posterior distribution.

The MAP estimator calculated from 300 iterations of the multiscale method (2.6), with a fixed time step equal to $\Delta = 0.02$, parameters $\delta = \sigma = 10^{-5}$, and $J = 8$ auxiliary processes, is illustrated in Figure 7. It appears from the figure that the MAP estimator obtained is close to the truth; denoting by θ^{MAP} the MAP estimator, we indeed calculate that

$$(3.8) \qquad \frac{\left\| a^\dagger - a^{\mathrm{MAP}} \right\|_{L^2(D)}}{\left\| a^\dagger \right\|_{L^2(D)}} = \frac{\sqrt{\sum_{|\ell|_\infty \leq N} \lambda_\ell \left| \theta_\ell^\dagger - \theta_\ell^{\mathrm{MAP}} \right|^2}}{\sqrt{\sum_{|\ell|_\infty \leq N} \lambda_\ell \left| \theta_\ell^\dagger \right|^2}} = 0.116...,$$

showing that the relative error is approximately 12%.

We now turn our attention to the problem of sampling from the Bayesian posterior. The marginals of the posterior associated with the first 16 KL coefficients, obtained by kernel density estimation from 20,000 iterations of the multiscale method (2.6) with $\nu = 1$ and all other parameters unchanged, are illustrated in Figure 8. In the same figure, the marginals of the approximate posterior distributions calculated using gfALDI and MCMC are depicted. The MCMC method employed is a variation on the preconditioned Crank–Nicolson (pCN) algorithm described in [21]. Specifically the proposal is based not on the prior Gaussian but on a Gaussian distribution $\mathcal{N}(m, \alpha K)$; that is, given θ_n the proposal for θ_{n+1} is

$$\theta_{n+1}^\star = m + \sqrt{1 - \beta^2}\,(\theta_n - m) + \beta \xi_n, \qquad \xi_n \sim \mathcal{N}(0, \alpha K).$$
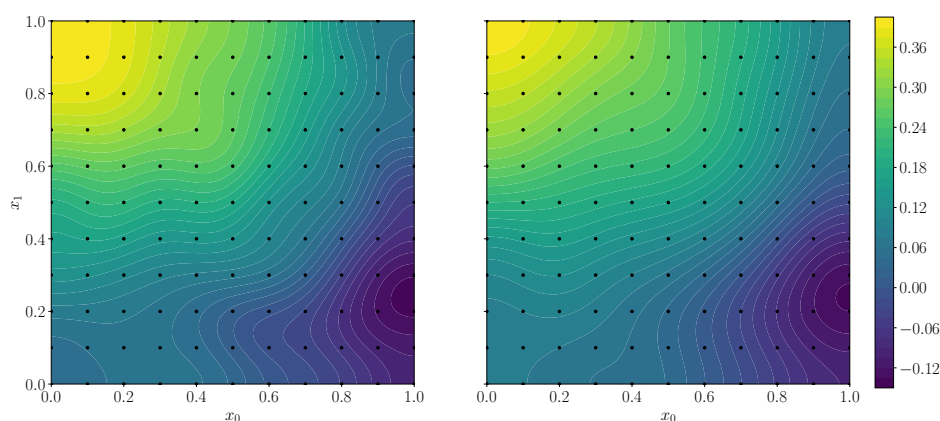
Figure 7. *Logarithms of true (left) and approximate (right) permeability profiles. The approximate permeability profile was constructed from the approximation of the MAP estimator provided by* (2.6) *with* $\nu = 0$. *The black dots are the observation points.*
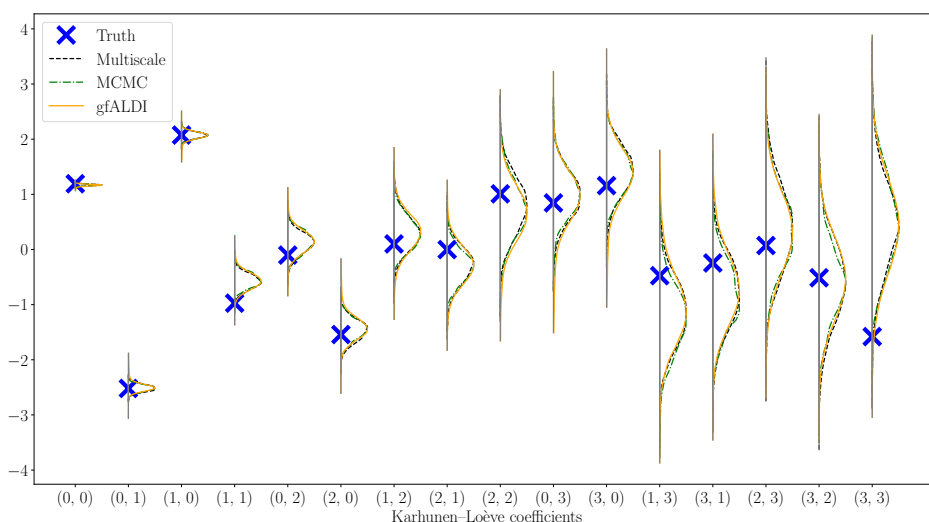

Figure 8. *Approximate posterior samples produced by* (2.6) *with* $\nu = 1$, *without model misspecification. From these posterior samples, the marginal distributions of the KL coefficients were approximated by kernel density estimation using Gaussian kernels; they are depicted (nonnormalized) in solid lines. The crosses are the true values of the coefficients, i.e., the values employed to generate the data.*

The parameters m and K are set to the mean and covariance of the posterior estimated by gfALDI, respectively, and a scaling factor $\alpha \geq 1$ is employed to ensure that the posterior distribution is absolutely continuous with respect to $\mathcal{N}(m, \alpha K)$. To generate the numerical results in Figure 8, we ran 20,000 iterations of this method with $\beta = 0.1$ and $\alpha = 4$. The agreement between the true parameter and the posterior samples is good overall, and the agreement between the approximate posteriors is also very good.

4. 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 MAP estimate. The method relies on a gradient approximation with a structure similar to that in the ensemble Kalman methods for sampling (EKS, ALDI) and inversion (EKI); a similar gradient structure was identified within the "analysis" step of the EnKF in the paper [78]. In contrast with these algorithms and other approximate sampling methods such as UKS and CBS, however, the method we propose is provably refinable over a bounded time interval: using tools from multi-scale analysis, we prove strong pathwise convergence to the gradient descent or overdamped Langevin dynamics, depending on whether it is employed for inversion or sampling, respectively. Although we show this result for the particular case where the state space is the d-dimensional torus, we believe that it should be possible, using results from [70], to extend our results to the case of an unbounded parameter space.

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 ALDI (or EKS, 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); the ideas developed in [6, 7, 66] might be useful in this regard. 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 [52], or semi-implicit discretizations (which preserve linearity of the updates) based on the formulation of ensemble Kalman methods in [53, section 4.3.3]; the diagonally semi-implicit modification of the forward Euler scheme used in [4] in the context of the ensemble Kalman–Bucy filter may also prove useful for developing efficient time-stepping schemes. Finally, one could study the mean field $J \to \infty$ and averaging $\delta \to 0$ limits for the alternative derivative-free formulation (2.5).

On the practical side, it will be important to determine how the method can be coupled to more efficient or less computationally expensive preconditioners than those computed from ALDI or EKS, or how these methods, 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. Our numerical experiments in this paper suggest that choosing $\delta > 0$ may not be advantageous for convergence, and if this is consistently observed, then the simpler discrete-time formulation (2.7) corresponding to the case $\delta = 0$ should be preferred over (2.6).

Appendix A. 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 multidimensional torus $\mathbf{T}^d$ with a uniform prior, so $\Phi_R = \Phi$. If the forward model satisfies $G \in C^3(\mathbf{T}^d, \mathbf{R}^K)$ then, by Taylor's formula, it holds for all $(\theta, \xi) \in \mathbf{T}^d \times \mathbf{R}^d$ that

(A.1)
$$
\begin{aligned}
&-\frac{1}{\sigma} \langle G(\theta + \sigma\xi) - G(\theta), G(\theta) - y \rangle_\Gamma \, \xi \\
&\qquad = -(\xi \otimes \xi) \nabla \Phi_R(\theta) - \frac{\sigma}{2} \sum_{k=1}^{K} \sum_{\ell=1}^{K} (\Gamma^{-1})_{k\ell} \Big((\xi \otimes \xi) : \mathrm{D}^2 G_k(\theta) \Big) \Big(G_\ell(\theta) - y_\ell \Big) \xi \\
&\qquad\quad - \frac{\sigma^2}{6} \sum_{k=1}^{K} \sum_{\ell=1}^{K} (\Gamma^{-1})_{k\ell} \Big((\xi \otimes \xi \otimes \xi) \mathbin{\vdots} \mathrm{D}^3 G_k(\theta + \alpha_k \sigma \xi) \Big) \Big(G_\ell(\theta) - y_\ell \Big) \xi
\end{aligned}
$$

for some $\alpha \in [0,1]^K$ depending on θ, ξ, and 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 $\mathrm{D}^2 G_k(\theta) \in \mathbf{R}^{d \times d}$ for the Hessian of G_k at θ, and the notation $\mathrm{D}^3 G_k(\theta) \in \mathbf{R}^{d \times d \times d}$ for the tensor of third derivatives of G_k at θ, i.e., the third-order tensor with components $(\mathrm{D}^3 G_k)_{ij\ell}(\theta) = \partial_i \partial_j \partial_\ell G_k(\theta)$. 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 $S \mathbin{\vdots} T = \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{k=1}^{d} S_{ijk} T_{ijk}$ for $S, T \in \mathbf{R}^{d \times d \times d}$.

Equation (A.1) motivates the following notation:

$$F_0(\theta, \Xi) = -C(\Xi) \nabla \Phi_R(\theta),$$

$$F_1(\theta, \Xi) = -\frac{1}{2J} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{\ell=1}^{K} (\Gamma^{-1})_{k\ell} \Big(\big(\xi^{(j)} \otimes \xi^{(j)} \big) : \mathrm{D}^2 G_k(\theta) \Big) \Big(G_\ell(\theta) - y_\ell \Big) \xi^{(j)},$$

$$F_2(\theta, \Xi) = \frac{1}{\sigma^2} \big(F^\sigma(\theta, \Xi) - F_0(\theta, \Xi) - \sigma F_1(\theta, \Xi) \big).$$

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

$$\dot{\theta} = F_0(\theta, \Xi) + \sigma F_1(\theta, \Xi) + \sigma^2 F_2(\theta, \Xi) + \nu\sqrt{2}\dot{w},$$

$$\dot{\xi}^{(j)} = -\frac{1}{\delta^2}\xi^{(j)} + \sqrt{\frac{2}{\delta^2}}\,\dot{w}^{(j)}, \qquad j = 1, \dots, J.$$

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

$$(A.2) \qquad \forall(\theta, \Xi) \in \mathbf{T}^d \times (\mathbf{R}^d)^J, \qquad |F^\sigma(\theta, \Xi) - F_0(\theta, \Xi)| \le C\sigma \sum_{j=1}^{J} |\xi^{(j)}|^3$$

for a constant C independent of σ. Likewise, if $G \in C^3(\mathbf{T}^d, \mathbf{R}^K)$, then by (A.1) it holds that

$$(A.3) \qquad \forall(\theta, \Xi) \in \mathbf{T}^d \times (\mathbf{R}^d)^J, \qquad |F^\sigma(\theta, \Xi) - F_0(\theta, \Xi) - \sigma F_1(\theta, \Xi)| \le C\sigma^2 \sum_{j=1}^{J} |\xi^{(j)}|^4$$

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 [74, 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

$$\mathcal{L}_0 = \sum_{j=1}^{J} -\xi^{(j)} \cdot \nabla_{\xi^{(j)}} + \triangle_{\xi^{(j)}},$$

$$\mathcal{L}_1 = -C(\Xi)\,\nabla\Phi_R(\theta) \cdot \nabla_\theta + \big(F^\sigma(\theta, \Xi) - F_0(\theta, \Xi)\big) \cdot \nabla_\theta + \nu\triangle_\theta.$$

Let Φ_R^ε denote a mollification with parameter ε of Φ_R, as defined in Lemma B.2. Since $\Phi \in C^2(\mathbf{T}^d, \mathbf{R})$ by the assumption that $G \in C^2(\mathbf{T}^d, \mathbf{R}^K)$, it holds by the standard properties of mollifiers that

$$(A.4) \qquad \forall i \in \{1, \dots, d\}, \qquad \|\partial_{\theta_i}\Phi_R^\varepsilon\|_{L^\infty(\mathbf{T}^d)} \le \|\partial_{\theta_i}\Phi_R\|_{L^\infty(\mathbf{T}^d)},$$

and similarly for the second derivatives. 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

$$(A.5) \qquad A(\Xi) = \frac{1}{2J}\sum_{j=1}^{J}(\xi^{(j)} \otimes \xi^{(j)} - I_d).$$

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

$$\psi(\theta_t, \Xi_t) - \psi(\theta_0, \Xi_0) = \frac{1}{\delta^2} \int_0^t \left(I_d - C(\Xi_s)\right) \nabla \Phi_R^\varepsilon(\theta_s) \, \mathrm{d}s + \int_0^t \left(F^\sigma(\theta_s, \Xi_s) \cdot \nabla_\theta + \nu \triangle_\theta\right) \psi(\theta_s, \Xi_s) \mathrm{d}s$$

$$+ \sqrt{\frac{2}{\delta^2}} \sum_{j=1}^J \int_0^t (\mathrm{d}w_s^{(j)} \cdot \nabla_{\xi^{(j)}}) \psi(\theta_s, \Xi_s) + \nu \sqrt{2} \int_0^t (\mathrm{d}w_s \cdot \nabla_\theta) \psi(\theta_s, \Xi_s).$$

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

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

and so we deduce

$$\theta_t - \theta_0 = -\int_0^t \nabla \Phi_R(\theta_s) \, \mathrm{d}s + \int_0^t \left(I_d - C(\Xi_s)\right)\left(\nabla \Phi_R(\theta_s) - \nabla \Phi_R^\varepsilon(\theta_s)\right) \mathrm{d}s$$

$$+ \int_0^t \left(F^\sigma(\theta_s, \Xi_s) - F_0(\theta_s, \Xi_s)\right) \mathrm{d}s + \nu\sqrt{2}w_t + \delta^2 \left(\psi(\theta_t, \Xi_t) - \psi(\theta_0, \Xi_0)\right)$$

$$- \delta^2 \int_0^t \left(F^\sigma(\theta_s, \Xi_s) \cdot \nabla_\theta + \nu \triangle_\theta\right) \psi(\theta_s, \Xi_s) \, \mathrm{d}s - \delta M_t - \delta^2 N_t,$$

where M_t and N_t are the martingale terms:

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

$$N_t = \nu\sqrt{2} \int_0^t \left(\mathrm{d}w_s \cdot \nabla_\theta\right) \psi(\theta_s, \Xi_s).$$

Using the fact that ϑ solves the averaged equation (2.8) with the same initial condition and Brownian motion, we deduce

$$\theta_t - \vartheta_t = -\int_0^t \left(\nabla \Phi_R(\theta_s) - \nabla \Phi_R(\vartheta_s)\right) \mathrm{d}s + \int_0^t \left(I_d - C(\Xi_s)\right)\left(\nabla \Phi_R(\theta_s) - \nabla \Phi_R^\varepsilon(\theta_s)\right) \mathrm{d}s$$

$$+ \int_0^t \left(F^\sigma(\theta_s, \Xi_s) - F_0(\theta_s, \Xi_s)\right) \mathrm{d}s + \delta^2 \left(\psi(\theta_t, \Xi_t) - \psi(\theta_0, \Xi_0)\right)$$

$$- \delta^2 \int_0^t \left(F^\sigma(\theta_s, \Xi_s) \cdot \nabla_\theta + \nu \triangle_\theta\right) \psi(\theta_s, \Xi_s) \, \mathrm{d}s - \delta M_t - \delta^2 N_t.$$

Let $e_t = \theta_t - \vartheta_t$. Using the Lipschitz continuity of $\nabla \Phi_R$ in order to bound the first integral on the right-hand side, the bound (B.3) for the second, the bound (A.2) for the third, the simple inequality (A.4) for the fourth, and the bound (B.2) for the fifth, noticing that this bound implies the inequality

$$\forall (\theta, \Xi) \in \mathbf{T}^d \times (\mathbf{R}^d)^J, \qquad \max\left\{\left|\nabla_\theta \psi_i(\theta, \Xi)\right|, \left|\mathrm{D}_\theta^2 \psi_i(\theta, \Xi)\right|_\mathrm{F}\right\} \le C\varepsilon^{-1}\left(1 + \sum_{j=1}^J |\xi^{(j)}|^2\right)$$

for all $i \in \{1, \ldots, d\}$, we obtain

$$|e_t| \leq L \int_0^t |e_s| \, \mathrm{d}s + C\varepsilon \int_0^t \left(1 + \sum_{j=1}^J |\xi_s^{(j)}|^2 \right) \mathrm{d}s + C\sigma \sum_{j=1}^J \int_0^t |\xi_s^{(j)}|^3 \, \mathrm{d}s$$

$$+ C\delta^2 \left(\sum_{j=1}^J \left| \xi_0^{(j)} \right|^2 + \sum_{j=1}^J \left| \xi_t^{(j)} \right|^2 \right)$$

$$+ C\delta^2 \varepsilon^{-1} \int_0^t \left(1 + \sum_{j=1}^J \left| \xi_s^{(j)} \right|^5 \right) \mathrm{d}s + \delta \, |M_t| + \delta^2 \, |N_t| \, .$$

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

$$\mathbf{E} \left(\sup_{0 \leq s \leq t} |e_s|^p \right) \leq CT^{p-1} \int_0^t \mathbf{E} \, |e_s|^p \, \mathrm{d}s + C\delta^p T^{p-1} \int_0^t \left(1 + \sum_{j=1}^J \mathbf{E} |\xi_s^{(j)}|^{2p} \right) \mathrm{d}s$$

$$+ \sigma^p T^{p-1} \sum_{j=1}^J \int_0^t \mathbf{E} |\xi_s^{(j)}|^{3p} \, \mathrm{d}s + C\delta^{2p} \left(\sum_{j=1}^J \mathbf{E} |\xi_0^{(j)}|^{2p} + \sum_{j=1}^J \mathbf{E} \left(\sup_{0 \leq s \leq t} |\xi_s^{(j)}|^{2p} \right) \right)$$

$$+ C\delta^p T^{p-1} \int_0^t \left(1 + \sum_{j=1}^J \mathbf{E} \left| \xi_s^{(j)} \right|^{5p} \right) \mathrm{d}s + C\delta^p \mathbf{E} \left(\sup_{0 \leq s \leq t} |M_t|^p \right) + C\delta^{2p} \mathbf{E} \left(\sup_{0 \leq s \leq t} |N_t|^p \right)$$

for all $t \in [0, T]$. Since $\xi^{(1)}, \ldots, \xi^{(J)}$ are identically distributed stationary stochastic processes, their moments are constant in time and they coincide, so we deduce

$$\mathbf{E} \left(\sup_{0 \leq s \leq t} |e_s|^p \right) \leq C \int_0^t \mathbf{E} \left(\sup_{0 \leq u \leq s} |e_u|^p \right) \mathrm{d}s + C\sigma^p + C\delta^p + C\delta^{2p} \mathbf{E} \left(\sup_{0 \leq s \leq t} |\xi_s^{(1)}|^{2p} \right)$$

$$+ C\delta^p \mathbf{E} \left(\sup_{0 \leq s \leq t} |M_t|^p \right) + C\delta^{2p} \mathbf{E} \left(\sup_{0 \leq s \leq t} |N_t|^p \right) \qquad \forall t \in [0, T].$$

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

(A.6) $$\qquad \forall \delta > 0, \qquad \mathbf{E} \left(\sup_{0 \leq t \leq T} \left| \xi_t^{(1)} \right|^p \right) \leq C \left(1 + \log \left(1 + \frac{T}{\delta^2} \right) \right)^{p/2} .$$

Therefore, since

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

it holds that

(A.7) $$\qquad \forall \delta \in (0, 1], \qquad \delta^p \mathbf{E} \left(\sup_{0 \leq t \leq T} |\xi_1|^{2p} \right) \leq C.$$

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 [64, Theorem 7.2], which is valid for $p \geq 2$, we obtain

$$\mathbf{E}\left(\sup_{0 \leq t \leq T} |M_t|^p\right) \leq \mathbf{E}\left(J^{p-1} \sum_{j=1}^{J} \sup_{0 \leq t \leq T} \left|M_t^{(j)}\right|^p\right) = J^p \mathbf{E}\left(\sup_{0 \leq t \leq T} \left|M_t^{(1)}\right|^p\right)$$

$$\leq J^p \left(\frac{p^3}{p-1}\right)^{\frac{p}{2}} T^{\frac{p-2}{p}} \int_0^T \mathbf{E}\left|\nabla_{\xi^{(1)}} \psi(\theta_t, \Xi_t)\right|_{\mathrm{F}}^p \, \mathrm{d}t,$$

where $|\bullet|_{\mathrm{F}}$ denotes the Frobenius norm. Since $\nabla \Phi_R$ is bounded on $\mathbf{T}^d$ and the moments of $\xi_t^{(1)}, \ldots, \xi_t^{(J)}$ are constant in time, we obtain the bound

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

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

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

Similarly, for fixed T it holds that

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

Using these bounds together with (A.7), we obtain

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

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

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

For the sake of simplicity, we consider only the case where $\nu = 0$, in which case a regularization in the same spirit as (B.1) is not necessary. Since the right-hand side is a cubic polynomial in $\xi^{(1)}, \ldots, \xi^{(J)}$ 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 $\xi^{(1)}, \ldots, \xi^{(J)}$. Indeed, the eigenfunctions of $\mathcal{L}_0$ are given by tensor products of Hermite polynomials; see, for example, [71, section 4.4] and [1]. Therefore, after applying

Itô's formula to the function $\varphi(\theta, \Xi) := A(\Xi)\,\nabla\Phi_R(\theta) + \phi(\Xi;\theta)$, we obtain

$$\theta_t - \theta_0 = -\int_0^t \nabla\Phi_R(\theta_s)\,\mathrm{d}s + \sqrt{2}\,w_t + \int_0^t \big(F^\sigma(\theta_s, \Xi_s) - F_0(\theta_s, \Xi_s) - \sigma F_1(\theta_s, \Xi_s)\big)\,\mathrm{d}s$$

$$+ \delta^2\big(\varphi(\theta_t, \Xi_t) - \varphi(\theta_0, \Xi_0)\big) - \delta^2 \int_0^t F^\sigma(\theta_s, \Xi_s) \cdot \nabla_\theta\varphi(\theta, \Xi)\,\mathrm{d}s$$

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

By (A.3), the last term on the first line leads to a bound scaling as σ^{2p}. 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 $\xi^{(1)}, \ldots, \xi^{(J)}$. ∎

A.1. Analysis of the discrete-time numerical method. Before showing Theorem 2.2, we show a preparatory result.

Proposition A.1. *Assume that $G \in C^2(\mathbf{T}^d, \mathbf{R}^K)$ and let $\hat{\theta}_n$ be the solution obtained by (2.7). Then there exists a constant $C = C(T, J)$ such that*

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

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 [27, 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

$$(A.8) \qquad \hat{\vartheta}_{n+1} = \hat{\vartheta}_n + \nabla\Phi_R(\hat{\vartheta}_n)\,\Delta + \nu\sqrt{2\Delta}\,x_n, \qquad \hat{\vartheta}_0 = \hat{\theta}_0.$$

By the standard theory of numerical methods for SDEs [35, 52], the difference between $\vartheta_{n\Delta}$ and its approximation $\hat{\vartheta}_n$ satisfies the bound

$$(A.9) \qquad \mathbf{E}\left(\sup_{0 \leq n \leq T/\Delta}\left|\hat{\vartheta}_n - \vartheta_{n\Delta}\right|^2\right) \leq C\Delta.$$

Here and below, C denotes a constant independent of σ and Δ that is allowed to change from occurrence to occurrence. Subtracting (A.8) 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

$$e_{n+1} = e_n + \left(F^\sigma(\hat{\theta}_n) + C(\hat{\Xi}_n)\nabla\Phi_R(\hat{\theta}_n)\right)\Delta + \left(\nabla\Phi_R(\hat{\vartheta}_n) - \nabla\Phi_R(\hat{\theta}_n)\right)\Delta$$

$$+ \left(\nabla\Phi_R(\hat{\theta}_n) - C(\hat{\Xi}_n)\nabla\Phi_R(\hat{\theta}_n)\right)\Delta$$

$$=: e_n + X_n\Delta + Y_n\Delta + Z_n\Delta.$$

Therefore

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

Using the tower property of conditional expectation, it holds that

$$|\mathbf{E}\,(e_n\cdot Z_n)| = \left|\mathbf{E}\left(\mathbf{E}\big(e_n\cdot Z_n\,|\,\hat\theta_n,\hat\vartheta_n\big)\right)\right| = \left|\mathbf{E}\left(e_n\cdot\mathbf{E}\big(Z_n\,|\,\hat\theta_n,\hat\vartheta_n\big)\right)\right| = 0,$$

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 that $\mathbf{E}\big(Z_n\,|\,\hat\theta_n,\hat\vartheta_n\big) = 0$. Thus, since by Young's inequality $(a+b)^2 \le (1+\varepsilon)a^2+\big(1+\frac{1}{\varepsilon}\big)b^2$ for any $a,b\in\mathbf{R}$,

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

$$(A.10)\qquad \le (1+\Delta)\mathbf{E}\,|e_n|^2 + \big(2\Delta + 3\Delta^2\big)\,\big(\mathbf{E}\,|X_n|^2 + \mathbf{E}\,|Y_n|^2\big) + 3\Delta^2\mathbf{E}\,|Z_n|^2\,.$$

We now bound the terms one by one.

- By (A.2), it holds that

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

  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 that $\mathbf{E}\,|X_n|^2 \le C\sigma^2$. If G is three times differentiable, we can carry out the Taylor expansion to the next order as in (A.3), leading to the refined bound $\mathbf{E}\,|X_n|^2 \le C\sigma^4$.

- The expectation of $|Y_n|^2$ can be bounded by using the Lipschitz continuity of Φ_R:

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

- To bound $\mathbf{E}\,|Z_n|^2$, 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:

$$(A.11)\qquad \mathbf{E}\,|Z_n|^2 \le 2\mathbf{E}\left|\nabla\Phi_R(\hat\theta_n)\right|^2 + 2\mathbf{E}\left|C(\hat\Xi_n)\nabla\Phi_R(\hat\theta_n)\right|^2 \le C.$$

Going back to (A.10) and combining the bounds, we deduce

$$\forall\delta\le 1,\qquad \mathbf{E}\,|e_{n+1}|^2 \le (1+\Delta)\mathbf{E}\,|e_n|^2 + C\Delta(\sigma^{2\beta} + \mathbf{E}\,|e_n|^2) + C\Delta^2$$

$$\le (1+C\Delta)\mathbf{E}\,|e_n|^2 + C\Delta\left(\sigma^{2\beta} + \Delta\right).$$

Let $\varepsilon = \sigma^{2\beta} + \Delta$. Applying the previous bound recursively,

$$\mathbf{E}\,|e_{n+1}|^2 \le (1+C\Delta)\big((1+C\Delta)\mathbf{E}\,|e_{n-1}|^2 + C\Delta\varepsilon\big) + C\Delta\varepsilon$$

$$\le \cdots \le (1+C\Delta)^{n+1}\,\mathbf{E}\,|e_0|^2 + C\Delta\varepsilon\sum_{i=0}^n(1+C\Delta)^i$$

$$\le \mathbf{E}\,|e_0|^2\,\mathrm{e}^{CT} + (n\Delta)\,\mathrm{e}^{CT}\,C\varepsilon \le C(\mathbf{E}\,|e_0|^2 + \varepsilon).$$

Since $\mathbf{E}\,|e_0|^2 = 0$, and in view of (A.9), this concludes the proof.   ■

We now prove the more general Theorem 2.2, which concerns the discrete-time dynamics (2.6). To this end, it is useful to introduce, for $j = 1, \ldots, J$ and $n = 0, \ldots, N$, an approximation $\hat{v}_n^{(j)}$ of $\hat{\xi}_n^{(j)}$ in (2.6b) 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), the following 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

$$(A.12) \qquad \hat{v}_n^{(j)} = \begin{cases} \hat{\xi}_n^{(j)} & \text{if } n \leq M, \\ \sqrt{1 - \mathrm{e}^{-2\frac{\Delta}{\delta^2}}} \sum_{m=n-M}^{n} x_{m-1}^{(j)} \, \mathrm{e}^{-(n-m)\frac{\Delta}{\delta^2}} & \text{if } n > M. \end{cases}$$

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

Lemma A.2. *Let $\{\hat{v}_n^{(j)}\}_{n=0}^{N}$ and $\{\hat{\xi}_n^{(j)}\}_{n=0}^{N}$, for $j = 1, \ldots, J$, be the discrete-time processes obtained by (2.6b) and (A.12). Then the following bound holds for a constant C independent of δ:*

$$(A.13) \qquad \forall j \in \{1, \ldots, J\}, \quad \forall n \in \{0, \ldots, N\}, \qquad \mathbf{E} \left| \hat{\xi}_n^{(j)} - \hat{v}_n^{(j)} \right|^4 \leq C \, \mathrm{e}^{-4(M+1)\frac{\Delta}{\delta^2}} \, .$$

Consequently, it holds that

$$(A.14) \qquad \forall n \in \{0, \ldots, N\}, \qquad \mathbf{E} \left| C(\hat{\Upsilon}_n) - C(\hat{\Xi}_n) \right|^2 \leq C \, \mathrm{e}^{-2(M+1)\frac{\Delta}{\delta^2}},$$

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 (A.8). In view of (A.9), it is sufficient to obtain a bound on e_n with the same right-hand side as in (2.12). Using the same definition for X_n and Y_n as in the proof of Proposition A.1, we have

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

now with

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

$$W_n = \nabla\Phi_R(\hat{\theta}_n) - C(\hat{\Upsilon}_n)\nabla\Phi_R(\hat{\theta}_n).$$

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

$$
|e_n|^2 \le 4\Delta^2 \left| \sum_{m=0}^{n-1} X_m \right|^2 + 4\Delta^2 \left| \sum_{m=0}^{n-1} Y_m \right|^2 + 4\Delta^2 \left| \sum_{m=0}^{n-1} Z_m \right|^2 + 4\Delta^2 \left| \sum_{m=0}^{n-1} W_m \right|^2
$$

$$
\le 4n\Delta^2 \sum_{m=0}^{n-1} |X_m|^2 + 4n\Delta^2 \sum_{m=0}^{n-1} |Y_m|^2 + 4n\Delta^2 \sum_{m=0}^{n-1} |Z_m|^2 + 4\Delta^2 \left| \sum_{m=0}^{n-1} W_m \right|^2 .
$$

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

(A.15)

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

The first term is bounded as in the proof of Proposition A.1. For the second term, we use the Lipschitz continuity of Φ_R to deduce

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

For the third term, we use (A.14), 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 (A.15), we calculate

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

$$
= \sum_{\ell=0}^{m-1} \mathbf{E} \, |W_\ell|^2 + 2 \sum_{\ell=0}^{m-1} \sum_{k=\ell+1}^{(\ell+M)\wedge(m-1)} \mathbf{E}(W_\ell \cdot W_k) + 2 \sum_{\ell=0}^{m-1} \sum_{k=\ell+M+1}^{m-1} \mathbf{E}(W_\ell \cdot W_k)
$$

$$
\le \sum_{\ell=0}^{m-1} \mathbf{E} \, |W_\ell|^2 + 2M \sum_{\ell=0}^{m-1} \mathbf{E} \, |W_\ell|^2 + 2 \sum_{\ell=0}^{m-1} \sum_{k=\ell+M+1}^{m-1} \mathbf{E}(W_\ell \cdot W_k).
$$

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

(A.16)

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

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

$$
|\mathbf{E}(W_\ell \cdot W_k)| = \left| \mathbf{E}\big(\mathbf{E}(W_\ell \cdot W_k \,|\, \hat{\theta}_\ell, \hat{\Upsilon}_\ell) \big) \right| = \left| \mathbf{E}\big(W_\ell \cdot \mathbf{E}(W_k \,|\, \hat{\theta}_\ell, \hat{\Upsilon}_\ell) \big) \right|
$$

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

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

$$
\begin{aligned}
\mathbf{E}_\ell(W_k) &= \frac{1}{J} \sum_{j=1}^{J} \mathbf{E}_\ell\left(\left(I_d - \hat{v}_k^{(j)} \otimes \hat{v}_k^{(j)} \right) \nabla \Phi_R(\hat{\theta}_k) \right) \\
&= \frac{1}{J} \sum_{j=1}^{J} \mathbf{E}_\ell\left(\mathbf{E}_\ell\left(\left(I_d - \hat{v}_k^{(j)} \otimes \hat{v}_k^{(j)} \right) \nabla \Phi_R(\hat{\theta}_k) \,|\, \hat{v}_k^{(j)} \right) \right) \\
&= \frac{1}{J} \sum_{j=1}^{J} \mathbf{E}_\ell\left(\left(I_d - \hat{v}_k^{(j)} \otimes \hat{v}_k^{(j)} \right) \mathbf{E}_\ell\left(\nabla \Phi_R(\hat{\theta}_k) \,|\, \hat{v}_k^{(j)} \right) \right) \\
&= \frac{1}{J} \sum_{j=1}^{J} \mathbf{E}_\ell\left(I_d - \hat{v}_k^{(j)} \otimes \hat{v}_k^{(j)} \right) \mathbf{E}_\ell\left(\nabla \Phi_R(\hat{\theta}_k) \right) \\
&\quad + \frac{1}{J} \sum_{j=1}^{J} \mathbf{E}_\ell\left(\left(I_d - \hat{v}_k^{(j)} \otimes \hat{v}_k^{(j)} \right) \left(\mathbf{E}_\ell\left(\nabla \Phi_R(\hat{\theta}_k) \,|\, \hat{v}_k^{(j)} \right) - \mathbf{E}_\ell\left(\nabla \Phi_R(\hat{\theta}_k) \right) \right) \right).
\end{aligned}
\tag{A.17}
$$

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

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

Consequently, we can calculate the first expectation in (A.17) exactly. Using Hölder's inequality for the other term, we obtain

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

Therefore, employing Jensen's inequality, we deduce

$$
\begin{aligned}
\mathbf{E}\left| \mathbf{E}_\ell(W_k) \right|^2 &\leq C \, \mathrm{e}^{-4(M+1)\frac{\Delta}{\delta^2}} + C \sum_{j=1}^{J} \mathbf{E} \left| \mathbf{E}_\ell\left(\nabla \Phi_R(\hat{\theta}_k) \,|\, v_k^{(j)} \right) - \mathbf{E}_\ell\left(\nabla \Phi_R(\hat{\theta}_k) \right) \right|^4 \\
&=: C \, \mathrm{e}^{-4(M+1)\frac{\Delta}{\delta^2}} + C \sum_{j=1}^{J} \mathbf{E}(B_j).
\end{aligned}
\tag{A.18}
$$

Since $\mathbf{E}_\ell\left(\nabla \Phi_R(\hat{\theta}_{k-M}) \,|\, \hat{v}_k^{(j)} \right) = \mathbf{E}_\ell\left(\nabla \Phi_R(\hat{\theta}_{k-M}) \right)$, it holds that

$$
\begin{aligned}
B_j &\leq 2^3 \left| \mathbf{E}_\ell\left(\nabla \Phi_R(\hat{\theta}_k) \,|\, \hat{v}_k^{(j)} \right) - \mathbf{E}_\ell\left(\nabla \Phi_R(\hat{\theta}_{k-M}) \,|\, \hat{v}_k^{(j)} \right) \right|^4 \\
&\quad + 2^3 \left| \mathbf{E}_\ell\left(\nabla \Phi_R(\hat{\theta}_{k-M}) \right) - \mathbf{E}_\ell\left(\nabla \Phi_R(\hat{\theta}_k) \right) \right|^4 \\
&= 2^3 \left| \mathbf{E}_\ell\left(\nabla \Phi_R(\hat{\theta}_k) - \nabla \Phi_R(\hat{\theta}_{k-M}) \,|\, \hat{v}_k^{(j)} \right) \right|^4 + 2^3 \left| \mathbf{E}_\ell\left(\nabla \Phi_R(\hat{\theta}_{k-M}) - \nabla \Phi_R(\hat{\theta}_k) \right) \right|^4.
\end{aligned}
$$

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

$$B_j \leq C \mathbf{E}_\ell \left(\left| \hat{\theta}_k - \hat{\theta}_{k-M} \right|^4 \mid \hat{v}_k^{(j)} \right) + C \mathbf{E}_\ell \left| \hat{\theta}_{k-M} - \hat{\theta}_k \right|^4,$$

so $\mathbf{E}(B_j) \leq C \mathbf{E} \left| \hat{\theta}_{k-M} - \hat{\theta}_k \right|^4$. Using the bound

$$\mathbf{E} \left| \hat{\theta}_k - \hat{\theta}_{k-M} \right|^4 = \mathbf{E} \left| \sum_{i=k-M}^{k-1} F^\sigma(\hat{\theta}_i, \hat{\Xi}_i)\Delta + \sum_{i=k-M}^{k-1} \nu \sqrt{2\Delta} x_i \right|^4 \leq C \left(M^4 \Delta^4 + M^2 \Delta^2 \right),$$

which is justified because $F^\sigma(\theta, \Xi) \leq C \big(1 + \sum_{j=1}^J \left| \xi^{(j)} \right|^3 \big)$ by (A.2), we deduce by going back to (A.18) that

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

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

$$\sqrt{\mathbf{E} \left| \mathbf{E}(W_k \mid \hat{\theta}_\ell, \hat{\Upsilon}_\ell) \right|^2} \leq C \, \mathrm{e}^{-2(M+1)\frac{\Delta}{\delta^2}} + C M^2 \Delta^2 + C M \Delta.$$

Employing this bound in (A.16), we deduce

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

$$\leq C \left(\Delta + M \Delta + M^2 \Delta^2 + \mathrm{e}^{-2(M+1)\frac{\Delta}{\delta^2}} \right).$$

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

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

(A.19) $$\leq C \left(\Delta + \log(1 + \delta^{-1})\delta^2 \right) \qquad \forall \delta \in (0, 1].$$

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 (A.19) vanishes in the limit $\delta \to 0$, i.e., when $\hat{\xi}_n^{(j)}$ are drawn independently from $\mathcal{N}(0, I_d)$ at each iteration. In this case, we recover the statement of Proposition A.1.

Combining everything in (A.15), we obtain

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

By the discrete Grönwall lemma, we deduce

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

which concludes the proof. ∎

Appendix B. Auxiliary results for Theorem 2.1.

Lemma B.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 that

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

for a constant C independent of T.

Proof. The strategy of the proof parallels that in [73, Theorem A.1], so here we give only a sketch. We use the notation $W(t)$ 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 the process $Z_t = \mathrm{e}^{-t} W(\mathrm{e}^{2t})$; see, for example, [71, Chapter 1]. Therefore, it holds that

$$\mathbf{E}\left(\sup_{0 \le t \le T} |Y_t|^p \right) = \mathbf{E}\left(\sup_{0 \le s \le \mathrm{e}^{2T}} \left| \frac{W_s}{\sqrt{s}} \right|^p \right)$$

$$\le \left(\log\log(2 + \mathrm{e}^{2T}) \right)^{p/2} \mathbf{E}\left(\sup_{0 \le s \le \mathrm{e}^{2T}} \left| \frac{W_s}{\sqrt{s \log\log(2 + s)}} \right|^p \right)$$

$$= \left(\log\log(2 + \mathrm{e}^{2T}) \right)^{p/2} \mathbf{E}\left| N\left(\mathrm{e}^{2T}\right) \right|^p,$$

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 $\log(e^{2T} + 2) \leq \log(e^{2T}) + 2 = 2T + 2$, so $\log \log(2 + e^{2T}) \leq 1 + \log(1 + T)$. It is shown in the proof of [73, Theorem A.1], and in the references therein, that there are constants C and σ such that

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

Therefore

$$\mathbf{E}\,\big|N(e^{2T})\big|^p = \int_0^\infty \mathbf{P}\big(\big|N(e^{2T})\big|^p > \lambda\big)\,\mathrm{d}\lambda$$

$$= \int_0^\infty \mathbf{P}\big(N(e^{2T}) > \lambda^{1/p}\big)\,\mathrm{d}\lambda \leq \int_0^\infty C\, e^{-\frac{\lambda^{2/p}}{4\sigma^2}}\,\mathrm{d}\lambda \leq K$$

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

Lemma B.2. *Assume* $\Phi_R \in C^2(\mathbf{T}^d, \mathbf{R})$, *and let* Φ_R^ε *denote a mollification with parameter* ε *of* Φ_R, *that is,*

(B.1)

$$\Phi_R^\varepsilon = \varrho_\varepsilon * \Phi_R, \qquad \varrho_\varepsilon = \varepsilon^{-d}\varrho\big(\varepsilon^{-1}\theta\big), \qquad \varrho: \mathbf{R}^d \to \mathbf{R}; \theta \mapsto \begin{cases} k\exp\left(-\frac{1}{1-|\theta|^2}\right) & \text{if } |\theta| \leq 1, \\ 0 & \text{if } |\theta| > 1, \end{cases}$$

where k *is a constant such that* ϱ *integrates to* 1 *over* $\mathbf{R}^d$ *and* $*$ *is the usual convolution of functions on* $\mathbf{R}^d$ *(identifying* Φ_R *with a 1-periodic function over* $\mathbf{R}^d$*). Then there is* C *independent of* ε *such that*

(B.2) $\qquad \forall(i, j, k) \in \{1, \ldots, d\}^3, \qquad \|\partial_{\theta_i}\partial_{\theta_j}\partial_{\theta_k}\Phi_R^\varepsilon(\theta)\|_{L^\infty(\mathbf{T}^d)} \leq C\varepsilon^{-1}\sup_{\theta \in \mathbf{T}^d}\big|\mathrm{D}^2\Phi_R(\theta)\big|_{\mathrm{F}},$

where $|\bullet|_{\mathrm{F}}$ *denotes the Frobenius norm, and*

(B.3) $$\|\nabla\Phi_R^\varepsilon - \nabla\Phi\|_{L^\infty(\mathbf{T}^d)} \leq C\varepsilon.$$

Proof. By the standard properties of mollifiers, it holds that

(B.4) $\qquad \forall(i, j, k) \in \{1, \ldots, d\}^3, \qquad \partial_{\theta_i}\partial_{\theta_j}\partial_{\theta_k}\Phi_R^\varepsilon = \partial_{\theta_i}\varrho_\varepsilon * \partial_{\theta_j}\partial_{\theta_k}\Phi_R.$

We calculate

$$\partial_{\theta_i}\varrho_\varepsilon(\theta) = -2\varepsilon^{-(d+1)}\left(\varrho(\varepsilon^{-1}\theta)\frac{\varepsilon^{-1}\theta_i}{(1-|\varepsilon^{-1}\theta|^2)^2}\right) =: -2\varepsilon^{-(d+1)}g_i(\varepsilon^{-1}\theta).$$

The function g_i, for any $i \in \{1, \ldots, d\}$, is smooth and supported in the closed ball of radius 1, and so there is some constant M (independent of ε) such that $|g_i(\tau)| \leq M\varrho(\tau/2)$ for all $\tau \in \mathbf{R}^d$. Therefore

$$\forall\theta \in \mathbf{T}^d, \qquad \big|\partial_{\theta_i}\varrho_\varepsilon(\theta)\big| = 2\varepsilon^{-(d+1)}\big|g_i(\varepsilon^{-1}\theta)\big| \leq 2M\varepsilon^{-(d+1)}\varrho(\varepsilon^{-1}\theta/2) = 2^{d+1}M\varepsilon^{-1}\varrho_{2\varepsilon}(\theta),$$

and so, going back to (B.4), it is simple to bound the third derivatives of Φ_R^ε using the second derivatives of Φ_R, at the expense of a large factor ε^{-1} on the right-hand side:

$$\forall \theta \in \mathbf{T}^d, \qquad |\partial_{\theta_i} \partial_{\theta_j} \partial_{\theta_k} \Phi_R^\varepsilon(\theta)| \le C\varepsilon^{-1} \sup_{\theta \in \mathbf{T}^d} \left| \mathrm{D}^2 \Phi_R(\theta) \right|_{\mathrm{F}}.$$

This proves (B.2). For the second claim, note that the following inequality holds for any 1-periodic Lipschitz continuous f with Lipschitz constant L:

$$\forall \theta \in \mathbf{T}^d, \qquad |f(\theta) - \varrho_\varepsilon * f(\theta)| = \left| \int_{\mathbf{R}^d} \big(f(\theta) - f(\theta + \tau) \big) \varrho_\varepsilon(\tau) \, \mathrm{d}\tau \right|$$
$$\le \int_{\mathbf{R}^d} |f(\theta) - f(\theta + \tau)| \varrho_\varepsilon(\tau) \, \mathrm{d}\tau \le L\varepsilon.$$

In particular, since $\nabla \Phi_R$ is Lipschitz continuous by the assumption, we have (B.3). ∎

Appendix C. Proof of Lemma A.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 that

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

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

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

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

It remains to prove that the second term in parentheses is bounded from above independently of δ. Since it holds that $|a - b|^4 \le C \sum_{i=1}^d |a_i - b_i|^4$ for all $a, b \in \mathbf{R}^d$, we assume without loss of generality that $d = 1$ in order to establish (A.13). For simplicity of notation, let $v_m = x_{n-M-2-m}^{(j)}$, so that

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

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

$$
\mathbf{E} \left| \sum_{m=0}^{n-M-2} v_m \, \mathrm{e}^{-m\frac{\Delta}{\delta^2}} \right|^4 = \sum_{m=0}^{n-M-2} \mathbf{E} \left| v_m \right|^4 \mathrm{e}^{-4m\frac{\Delta}{\delta^2}} + 6 \sum_{m=0}^{n-M-2} \sum_{\ell=m+1}^{n-M-2} \mathbf{E} \left| v_m \right|^2 \mathbf{E} \left| v_\ell \right|^2 \mathrm{e}^{-2(m+\ell)\frac{\Delta}{\delta^2}}
$$

$$
\leq 3 \sum_{m=0}^{\infty} \mathrm{e}^{-4m\frac{\Delta}{\delta^2}} + 6 \sum_{m=0}^{\infty} \sum_{\ell=m+1}^{\infty} \mathrm{e}^{-2(m+\ell)\frac{\Delta}{\delta^2}}
$$

$$
\leq \frac{3}{1-\mathrm{e}^{-4\frac{\Delta}{\delta^2}}} + \frac{6}{\left(1-\mathrm{e}^{-2\frac{\Delta}{\delta^2}}\right)^2} \leq \frac{9}{\left(1-\mathrm{e}^{-2\frac{\Delta}{\delta^2}}\right)^2}.
$$

Therefore we deduce (A.13), because the denominator cancels out with the factor of the second term in parentheses in (C.1). In order to derive (A.14), we use Hölder's inequality and (A.13),

$$
\mathbf{E} \left| C(\hat{\Upsilon}_n) - C(\hat{\Xi}_n) \right|^2 \leq C \sum_{j=1}^{J} \mathbf{E} \left| \hat{v}_n^{(j)} \otimes \hat{v}_n^{(j)} - \hat{\xi}_n^{(j)} \otimes \hat{\xi}_n^{(j)} \right|^2
$$

$$
\leq C \sum_{j=1}^{J} \mathbf{E} \left| \hat{v}_n^{(j)} \otimes (\hat{v}_n^{(j)} - \hat{\xi}_n^{(j)}) - (\hat{\xi}_n^{(j)} - \hat{v}_n^{(j)}) \otimes \hat{\xi}_n^{(j)} \right|^2
$$

$$
\leq C \sum_{j=1}^{J} \mathbf{E} \left(\left(\left| \hat{v}_n^{(j)} \right|^2 + \left| \hat{\xi}_n^{(j)} \right|^2 \right) \left| \hat{v}_n^{(j)} - \hat{\xi}_n^{(j)} \right|^2 \right)
$$

$$
\leq C \sum_{j=1}^{J} \mathbf{E} \left(\left| \hat{v}_n^{(j)} - \hat{\xi}_n^{(j)} \right|^4 \right)^{1/2} \leq C \, \mathrm{e}^{-2(M+1)\frac{\Delta}{\delta^2}},
$$

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 (2017), pp. 720–761, https://doi.org/10.1137/16M1094117.

[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 (2018), 045002.

[3] V. AKCELIK, G. BIROS, AND O. GHATTAS, *Parallel multiscale Gauss–Newton–Krylov methods for inverse wave propagation*, in Proceedings of the 2002 ACM/IEEE Conference on Supercomputing (SC'02), IEEE, 2002, pp. 41–41.

[4] J. AMEZCUA, K. IDE, E. KALNAY, AND S. REICH, *Ensemble transform Kalman–Bucy filters*, Quart. J. Roy. Meteorol. Soc., 140 (2014), pp. 995–1004.

[5] S. BADIA AND F. VERDUGO, *Gridap: An extensible finite element toolbox in Julia*, J. Open Source Softw., 5 (2020), 2520.

[6] V. BALLY AND D. TALAY, *The law of the Euler scheme for stochastic differential equations. I. Convergence rate of the distribution function*, Probab. Theory Related Fields, 104 (1996), pp. 43–60.

[7] V. BALLY AND D. TALAY, *The law of the Euler scheme for stochastic differential equations. II. Convergence rate of the density*, Monte Carlo Methods Appl., 2 (1996), pp. 93–128.

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

[9] M. Bocquet and P. Sakov, *An iterative ensemble Kalman smoother*, Quart. J. Roy. Meteorol. Soc., 140 (2014), pp. 1521–1535.

[10] A. Borovykh, N. Kantas, P. Parpas, and G. A. Pavliotis, *To interact or not? The convergence properties of interacting stochastic mirror descent*, in Beyond First Order Methods in Machine Learning Systems Workshop, 36th International Conference on Machine Learning (ICML), PMLR, 2020.

[11] A. Borovykh, N. Kantas, P. Parpas, and G.A. Pavliotis, *On stochastic mirror descent with interacting particles: Convergence properties and variance reduction*, Phys. D, 418 (2021), 132844.

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

[13] K. M. Brown and J. E. Dennis, Jr., *Derivative free analogues of the Levenberg-Marquardt and Gauss algorithms for nonlinear least squares approximation*, Numer. Math., 18 (1971/1972), pp. 289–297.

[14] 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 (2018), pp. 1037–1066.

[15] J. A. Carrillo, F. Hoffmann, A. M. Stuart, and U. Vaes, *Consensus Based Sampling*, preprint, https://arxiv.org/abs/2106.02519, 2021.

[16] J. A. Carrillo and U. Vaes, *Wasserstein stability estimates for covariance-preconditioned Fokker-Planck equations*, Nonlinearity, 34 (2021), pp. 2275–2295.

[17] 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 (2018), 30.

[18] Y. Chen and D. S. Oliver, *Ensemble randomized maximum likelihood method as an iterative ensemble smoother*, Math. Geosci., 44 (2012), pp. 1–26.

[19] C. Clason, T. Helin, R. Kretschmann, and P. Piiroinen, *Generalized modes in Bayesian inverse problems*, SIAM/ASA J. Uncertain. Quantif., 7 (2019), pp. 652–684, https://doi.org/10.1137/18M1191804.

[20] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to Derivative-Free Optimization*, MPS/SIAM Ser. Optim. 8, SIAM, Philadelphia, 2009, https://doi.org/10.1137/1.9780898718768.

[21] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White, *MCMC methods for functions: Modifying old algorithms to make them faster*, Statist. Sci., 28 (2013), pp. 424–446.

[22] 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 (2013), 095017.

[23] M. Dashti and A. M. Stuart, *The Bayesian approach to inverse problems*, in Handbook of Uncertainty Quantification, Vol. 1, Springer, Cham, 2017, pp. 311–428.

[24] P. Del Moral, *Mean Field Simulation for Monte Carlo Integration*, Monogr. Statist. Appl. Probab. 126, CRC Press, Boca Raton, FL, 2013.

[25] P. Del Moral, A. Doucet, and A. Jasra, *Sequential Monte Carlo samplers*, J. R. Stat. Soc. Ser. B Stat. Methodol., 68 (2006), pp. 411–436.

[26] O. R. A. Dunbar, A. Garbuno-Inigo, T. Schneider, and A. M. Stuart, *Calibration and uncertainty quantification of convective parameters in an idealized GCM*, J. Adv. Model. Earth Syst., 13 (2021), e2020MS002454.

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

[28] A. A. Emerick and A. C. Reynolds, *Investigation of the sampling performance of ensemble-based methods with a simple reservoir model*, Comput. Geosci., 17 (2013), pp. 325–350.

[29] O. G. Ernst, B. Sprungk, and H.-J. Starkloff, *Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems*, SIAM/ASA J. Uncertain. Quantif., 3 (2015), pp. 823–851, https://doi.org/10.1137/140981319.

[30] G. Evensen, *Data Assimilation: The Ensemble Kalman Filter*, 2nd ed., Springer-Verlag, Berlin, 2009.

[31] 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 (2020), pp. 412–441, https://doi.org/10.1137/19M1251655.

[32] A. Garbuno-Inigo, N. Nüsken, and S. Reich, *Affine invariant interacting Langevin dynamics for Bayesian inference*, SIAM J. Appl. Dyn. Syst., 19 (2020), pp. 1633–1658, https://doi.org/10.1137/

19M1304891.

[33] P. E. GILL, W. MURRAY, M. A. SAUNDERS, AND M. H. WRIGHT, *Computing forward-difference intervals for numerical optimization*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 310–321, https://doi.org/10.1137/0904025.

[34] J. GOODMAN AND J. WEARE, *Ensemble samplers with affine invariance*, Commun. Appl. Math. Comput. Sci., 5 (2010), pp. 65–80.

[35] C. GRAHAM AND D. TALAY, *Stochastic Simulation and Monte Carlo Methods: Mathematical Foundations of Stochastic Simulation*, Stoch. Model. Appl. Probab. 68, Springer, Heidelberg, 2013.

[36] S. GRATTON, A. S. LAWLESS, AND N. K. NICHOLS, *Approximate Gauss–Newton methods for nonlinear least squares problems*, SIAM J. Optim., 18 (2007), pp. 106–132, https://doi.org/10.1137/050624935.

[37] P. GREENGARD, *An Ensemblized Metropolized Langevin Sampler*, Master's thesis, Courant Institute, New York University, 2015.

[38] 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*, preprint, https://arxiv.org/abs/1805.08034, 2018.

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

[40] W. K. HASTINGS, *Monte Carlo sampling methods using Markov chains and their applications*, Biometrika, 57 (1970), pp. 97–109.

[41] T. HELIN AND M. BURGER, *Maximum a posteriori probability estimates in infinite-dimensional Bayesian inverse problems*, Inverse Problems, 31 (2015), 085009.

[42] T. HELIN AND M. LASSAS, *Hierarchical models in statistical inverse problems and the Mumford-Shah functional*, Inverse Problems, 27 (2011), 015008.

[43] M. HERTY AND G. VISCONTI, *Kinetic methods for inverse problems*, Kinet. Relat. Models, 12 (2019), pp. 1109–1130.

[44] M. R. HESTENES AND E. STIEFEL, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.

[45] D. Z. HUANG, T. SCHNEIDER, AND A. M. STUART, *Unscented Kalman Inversion*, preprint, https://arxiv.org/abs/2102.01580, 2021.

[46] M. A. IGLESIAS, K. J. H. LAW, AND A. M. STUART, *Ensemble Kalman methods for inverse problems*, Inverse Problems, 29 (2013), 045001.

[47] 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 (2015), pp. 2263–2300.

[48] J. KAIPIO AND E. SOMERSALO, *Statistical and Computational Inverse Problems*, Appl. Math. Sci. 160, Springer-Verlag, New York, 2005.

[49] N. KANTAS, P. PARPAS, AND G. A. PAVLIOTIS, *The Sharp, the Flat and the Shallow: Can Weakly Interacting Agents Learn to Escape Bad Minima?*, preprint, https://arxiv.org/abs/1905.04121, 2019.

[50] J. KENNEDY AND R. EBERHART, *Particle swarm optimization*, in Proceedings of ICNN'95-International Conference on Neural Networks, Vol. 4, IEEE, 1995, pp. 1942–1948.

[51] S. KIRKPATRICK, C. D. GELATT, JR., AND M. P. VECCHI, *Optimization by simulated annealing*, Science, 220 (1983), pp. 671–680.

[52] P. E. KLOEDEN AND E. PLATEN, *Numerical Solution of Stochastic Differential Equations*, Appl. Math. (N.Y.) 23, Springer-Verlag, Berlin, 1992.

[53] N. B. KOVACHKI AND A. M. STUART, *Ensemble Kalman inversion: A derivative-free technique for machine learning tasks*, Inverse Problems, 35 (2019), 095005.

[54] H. J. KUSHNER AND G. G. YIN, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed., Appl. Math. (N.Y.) 35, Stoch. Model. Appl. Probab., Springer-Verlag, New York, 2003.

[55] J. LARSON, M. MENICKELLY, AND S. M. WILD, *Derivative-free optimization methods*, Acta Numer., 28 (2019), pp. 287–404.

[56] B. LEIMKUHLER AND C. MATTHEWS, *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*, Interdiscip. Appl. Math. 39, Springer, Cham, 2015.

[57] B. LEIMKUHLER, C. MATTHEWS, AND G. STOLTZ, *The computation of averages from equilibrium and nonequilibrium Langevin molecular dynamics*, IMA J. Numer. Anal., 36 (2016), pp. 13–79.

[58] B. LEIMKUHLER, C. MATTHEWS, AND J. WEARE, *Ensemble preconditioning for Markov chain Monte*

Carlo simulation, Stat. Comput., 28 (2018), pp. 277–290.

[59] B. LEIMKUHLER AND M. SACHS, *Ergodic properties of quasi-Markovian generalized Langevin equations with configuration dependent noise and non-conservative force*, in Stochastic Dynamics Out of Equilibrium, G. Giacomin, S. Olla, E. Saada, H. Spohn, and G. Stoltz, eds., Springer Proc. Math. Stat. 282, Springer, Cham, 2019, pp. 282–330.

[60] T. LELIÈVRE AND G. STOLTZ, *Partial differential equations and stochastic methods in molecular dynamics*, Acta Numer., 25 (2016), pp. 681–880.

[61] K. LEVENBERG, *A method for the solution of certain non-linear problems in least squares*, Quart. Appl. Math., 2 (1944), pp. 164–168.

[62] Q. LIU AND D. WANG, *Stein variational gradient descent: A general purpose Bayesian inference algorithm*, in Adv. Neural Inform. Process. Syst., 2016, pp. 2378–2386.

[63] J. LU, Y. LU, AND J. NOLEN, *Scaling limit of the Stein variational gradient descent: The mean field regime*, SIAM J. Math. Anal., 51 (2019), pp. 648–671, https://doi.org/10.1137/18M1187611.

[64] X. MAO, *Stochastic Differential Equations and Applications*, 2nd ed., Horwood, Chichester, UK, 2008.

[65] D. W. MARQUARDT, *An algorithm for least-squares estimation of nonlinear parameters*, J. Soc. Indust. Appl. Math., 11 (1963), pp. 431–441, https://doi.org/10.1137/0111030.

[66] 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 (2010), pp. 552–577, https://doi.org/10.1137/090770527.

[67] 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 (1953), pp. 1087–1092.

[68] N. NÜSKEN AND S. REICH, *Note on Interacting Langevin Diffusions: Gradient Structure and Ensemble Kalman Sampler by Garbuno-Inigo, Hoffmann, Li and Stuart*, preprint, https://arxiv.org/abs/1908.10890, 2019.

[69] M. OTTOBRE AND G. A. PAVLIOTIS, *Asymptotic analysis for the generalized Langevin equation*, Nonlinearity, 24 (2011), pp. 1629–1653.

[70] E. PARDOUX AND A. YU. VERETENNIKOV, *On the Poisson equation and diffusion approximation.* I, Ann. Probab., 29 (2001), pp. 1061–1085.

[71] G. A. PAVLIOTIS, *Stochastic Processes and Applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*, Texts Appl. Math. 60, Springer, New York, 2014.

[72] G. A. PAVLIOTIS, G. STOLTZ, AND U. VAES, *Scaling limits for the generalized Langevin equation*, J. Nonlinear Sci., 31 (2021), 8.

[73] G. A. PAVLIOTIS AND A. M. STUART, *White noise limits for inertial particles in a random field*, Multiscale Model. Simul., 1 (2003), pp. 527–553, https://doi.org/10.1137/S1540345903421076.

[74] G. A. PAVLIOTIS AND A. M. STUART, *Multiscale Methods: Averaging and Homogenization*, Texts Appl. Math. 53, Springer, New York, 2008.

[75] 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 (2014), pp. A1525–A1555, https://doi.org/10.1137/130934805.

[76] 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 (2017), pp. 183–204.

[77] J. QUER, *Importance Sampling for Metastable Dynamical Systems in Molecular Dynamics*, Ph.D. thesis, Feie Universität Berlin, 2018.

[78] S. REICH, *A dynamical systems framework for intermittent data assimilation*, BIT, 51 (2011), pp. 235–249.

[79] S. REICH AND S. WEISSMANN, *Fokker–Planck particle systems for Bayesian inference: Computational approaches*, SIAM/ASA J. Uncertain. Quantif., 9 (2021), pp. 446–482, https://doi.org/10.1137/19M1303162.

[80] H. ROBBINS AND S. MONRO, *A stochastic approximation method*, Ann. Math. Statistics, 22 (1951), pp. 400–407.

[81] G. O. ROBERTS, A. GELMAN, AND W. R. GILKS, *Weak convergence and optimal scaling of random walk Metropolis algorithms*, Ann. Appl. Probab., 7 (1997), pp. 110–120.

[82] G. O. ROBERTS AND J. S. ROSENTHAL, *Optimal scaling for various Metropolis-Hastings algorithms*, Statist. Sci., 16 (2001), pp. 351–367.

[83] G. O. ROBERTS AND R. L. TWEEDIE, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, 2 (1996), pp. 341–363.

[84] P. SAKOV, D. S. OLIVER, AND L. BERTINO, *An iterative EnKF for strongly nonlinear systems*, Mon. Weather Rev., 140 (2012), pp. 1988–2004.

[85] C. SCHILLINGS AND A. M. STUART, *Analysis of the ensemble Kalman filter for inverse problems*, SIAM J. Numer. Anal., 55 (2017), pp. 1264–1290, https://doi.org/10.1137/16M105959X.

[86] C. SCHILLINGS AND A. M. STUART, *Convergence analysis of ensemble Kalman inversion: The linear, noisy case*, Appl. Anal., 97 (2018), pp. 107–123.

[87] A. M. STUART, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.

[88] D. B. WOODARD, S. C. SCHMIDLER, AND M. HUBER, *Sufficient conditions for torpid mixing of parallel and simulated tempering*, Electron. J. Probab., 14 (2009), pp. 780–804.

[89] Z. WU, *The effective energy transformation scheme as a special continuation approach to global optimization with application to molecular conformation*, SIAM J. Optim., 6 (1996), pp. 748–768, https://doi.org/10.1137/S1052623493254698.

[90] S. ZHANG, A. E. CHOROMANSKA, AND Y. LECUN, *Deep learning with elastic averaging SGD*, in Adv. Neural Inform. Process. Syst. 28, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds., Curran Associates, 2015, pp. 685–693.