Received: 24 May 2021 | Revised: 18 September 2021 | Accepted: 23 October 2021

DOI: 10.1111/sapm.12470

ORIGINAL ARTICLE

STUDIES IN APPLIED MATHEMATICS WILEY

Consensus-based sampling

J. A. Carrillo[1] | F. Hoffmann[2] | A. M. Stuart[3] | U. Vaes[4]

[1] Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK

[2] Hausdorff Center for Mathematics, Rheinische Friedrich-Wilhelms-Universität, Bonn 53115, Germany

[3] Department of Computing and Mathematical Sciences, Caltech, Pasadena, California 91125, USA

[4] MATHERIALS team, Inria Paris, Paris 75012, France

Correspondence
J. A. Carrillo, Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK.
Email: carrillo@maths.ox.ac.uk

Funding information
Advanced Grant Nonlocal-CPD (Nonlocal PDEs for Complex Particle Dynamics: Phase Transitions, Patterns and Synchronization) of the European Research Council Executive Agency (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 883363), EPSRC grant number EP/T022132/1, EPSRC grant number EP/P031587/1, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - GZ 2047/1, Projekt-ID 390685813, NSF (awards AGS-1835860 and DMS-1818977), the Office of Naval Research (award N00014-17-1-2079), and the Fondation Sciences Mathématiques de Paris (FSMP), through a postdoctoral fellowship in the "mathematical interactions" program.

Abstract

We propose a novel method for sampling and optimization tasks based on a stochastic interacting particle system. We explain how this method can be used for the following two goals: (i) generating approximate samples from a given target distribution and (ii) optimizing a given objective function. The approach is derivative-free and affine invariant, and is therefore well-suited for solving inverse problems defined by complex forward models: (i) allows generation of samples from the Bayesian posterior and (ii) allows determination of the maximum a posteriori estimator. We investigate the properties of the proposed family of methods in terms of various parameter choices, both analytically and by means of numerical simulations. The analysis and numerical simulation establish that the method has potential for general purpose optimization tasks over Euclidean space; contraction properties of the algorithm are established under suitable conditions, and computational experiments demonstrate wide basins of attraction for various specific problems. The analysis and experiments also demonstrate the potential for the sampling methodology in regimes in which the target distribution is unimodal and close to Gaussian; indeed we prove that the method recovers a Laplace approximation to the measure in certain parametric regimes and provide numerical evidence that this Laplace approximation attracts a large set of initial conditions in a number of examples.

KEYWORDS
optimization, sampling, stochastic interacting particle systems

1 INTRODUCTION

1.1 Background

We consider the inverse problem of finding θ from y where

$$y = G(\theta) + \eta. \tag{1}$$

Here $y \in \mathbf{R}^K$ is the *observation*, $\theta \in \mathbf{R}^d$ is the *unknown parameter*, $G : \mathbf{R}^d \to \mathbf{R}^K$ is the *forward model*, and η is the *observational noise*. We adopt the Bayesian approach to inversion[1] and assume that the parameter and the noise are independent and normally distributed: $\theta \sim \mathsf{N}(0, \Sigma)$ and $\eta \sim \mathsf{N}(0, \Gamma)$. By (1) and Bayes' formula, the posterior density (i.e., the conditional probability density function of θ given y) equals

$$\rho(\theta) = \frac{\exp(-f(\theta))}{\int_{\mathbf{R}^d} \exp(-f(\theta))\,\mathrm{d}\theta}, \tag{2}$$

where

$$f(\theta) := \Phi(\theta; y) + \frac{1}{2}|\theta|_\Sigma{}^2, \qquad \Phi(\theta; y) = \frac{1}{2}|y - G(\theta)|_\Gamma{}^2. \tag{3}$$

In the foregoing and in what follows, we adopt the following notation: for a positive definite matrix A,

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

For a matrix B, we denote by $\|B\|$ the operator norm induced by the Euclidean vector norm, and we also define the weighted matrix norm $\|B\|_A = \|A^{-1/2} B A^{-1/2}\|$ (noting that this is not the induced matrix norm from vector norm $|\bullet|_A$).

Solving inverse problems in the Bayesian framework can be prohibitively expensive because of the need to characterize an entire probability distribution. One approach to this is simply to seek the point of maximum posterior probability, the MAP point,[1,2] defined by

$$\theta_* = \operatorname{argmin}_\theta f(\theta). \tag{4}$$

However, this essentially reduces the solution of the inverse problem to a classical optimization approach[3] and fails to capture uncertainty. A compromise between a fully Bayesian approach and the classical optimization approach is to seek a Gaussian approximation of the measure.[4] By the Bernstein–von Mises theorem (and its extensions),[5] the posterior is expected to be well approximated by a Gaussian density in the large data limit, if the parameter is identifiable in the infinite data setting; a Gaussian approximation is also expected to be good if the forward map is close to linear. For these reasons, use of the Laplace method[6] to obtain a Gaussian approximation of the posterior density is often viewed as a useful approach in many application domains.

Many inverse problems arising in applications are defined by complex forward models G, often available only as a black box, and in particular adjoints and derivatives may not be readily available. Consensus-based approaches are proving to be interesting and viable derivative-free techniques for optimization.[7–9] The focus of this paper is on developing consensus-based sampling (CBS) of the posterior distribution for Bayesian inverse problems and, in particular, on the study of such methods in the context of Gaussian approximation of the posterior.

The computational methodology we introduce applies to arbitrary measures with negative log density f, and is not restricted to the choice in (3) resulting from the inverse problem (1). Some of our analysis, however, is specific to the inverse problem in the case where G is linear. The proposed methodology is potentially useful for the solution of complex problems for which the evaluation of f or G is expensive, and derivatives of f and G are not available, or noisy and not useable. In this sense the proposed methodology is competitive with state-of-the-art ensemble Kalman methods for inverse problems, which are also of particular value for derivative-free sampling when G is expensive to evaluate. The fact that the analysis of the accuracy of the proposed sampling method is confined to unimodal distributions, which are close to Gaussian, is also a limitation of ensemble Kalman methods. Our work thus provides impetus for further innovation in the analysis and design of particle-based, derivative-free sampling methods.

1.2 | Literature review

Systematic procedures to sample probability measures have their roots in statistical physics and the 1953 paper of Metropolis et al.[10] In 1970 Hastings recognized this work as a special case of what is now known as the Metropolis-Hastings methodology.[11] These methods in turn may be seen as part of the broader Markov chain Monte Carlo (MCMC) approach to sampling.[12] In 2006, sequential Monte Carlo (SMC) methods, based on creating a homotopy deforming the initial (simple to sample) measure into the desired target measure, were introduced[13]; in practice these methods work best when entwined with MCMC kernels. These SMC methods introduce the idea of using the evolution of a system of interacting particles to approximate the desired target measure; the large particle limit of this evolution captures the homotopy from the initial measure to the target measure. In a parallel development, the mathematical physics community has developed a large body of understanding of interacting particle systems, and their mean-field limits, initially primarily for models on a countable state space[14,15] and more recently for models in uncountable state space.[16–20] Studying interactions between sampling, collective dynamics of particles and mean-field limits holds considerable promise as a direction for finding improved sampling algorithms for specific classes of problems and is an active area of research.[21–24]

The focus of this work is on sampling measure (2), or optimizing objective function (4), by means of algorithms, which only involve black box evaluation of G. While some MCMC and SMC methods are of this type, the Metropolis algorithm being a primary example, the use of collective dynamics of particles opens the door to a wider range of methods to solve inverse problems in this setting. There are two primary classes of methods emerging in this context: those arising from consensus forming dynamics[9] and those arising from ensemble Kalman methods.[25]

Iterative ensemble Kalman methods for inverse problems were introduced in Refs. 26, 27. Similar ideas are also implicit in the work of Reich[22] who studies state estimation sequential data assimilation, rather than the inverse problem; however, what is termed the “analysis” step in sequential data assimilation corresponds to solving a Bayesian inverse problem. These iterative ensemble Kalman methods are similar to SMC in that they seek to map the prior to the posterior in finite continuous time or in a finite number of steps. Reich also introduced continuous time analysis of ensemble Kalman methods for state estimation in Refs. 28, 29, naming the resulting algorithm the ensemble Kalman Bucy filter (EnKBF); the ensemble Kalman approach to inverse problems introduced in Refs. 26, 27 may be studied using the EnKBF leading to a clear link with SMC methods in continuous time. An alternative Kalman methodology (ensemble Kalman inversion [EKI]) for the optimization approach to the inverse problem, which involves iteration to infinity, was introduced and studied in Refs. 30, 31 in discrete time and in Refs. 32, 33 in continuous time; the idea of using ensemble methods for optimization rather than sampling was anticipated in Ref. 22. The ensemble-based optimization approach was generalized to approximate sampling of the Bayesian posterior solution to the inverse problem in Ref. 34 (the ensemble Kalman sampler [EKS]), and studied further in Refs. 35–37.

The idea of consensus-based optimization (CBO) may be seen as a variation of particle swarm optimization methods,[38,39] which are themselves related to Cucker-Smale dynamics for collective behavior and opinion formation.[16,18,40–43] These dynamical systems model the tendency of the constituent particles to align (consensus in velocity) or to concentrate in certain variables modeling averaged quantities (consensus in position or opinion), and they have been extensively studied in terms of long time asymptotics leading to consensus.[42,44] CBO was introduced in Ref. 9 based on the following simple idea: particles are explorers in the landscape of the graph of the function $f(\theta)$ to be minimized, they are able to exchange information instantaneously, and they redirect their movement toward the location of a consensus position in parameter space that is a weighted average of the explorer's parameter values relative to the Gibbs measure associated to the function f, $\frac{1}{Z}e^{-f(\theta)}$. Noise is introduced for suitable exploration in parameter space but the strength of the noise is reduced according to the distance to the consensus parameter values. These effects lead to concentration in parameter space at the global minimum of the function, as proven in Ref. 7 for the mean-field limit Partial Differential Equation (PDE) and in Ref. 45 for the particle system under certain conditions on f and the parameters of the model . The original CBO method has been recently improved so as to be efficient for high-dimensional optimization problems,[8] such as those arising in machine learning, by adding coordinate-wise noise terms and introducing ideas from random batch methods[46] for computing stochastic particle systems efficiently. Furthermore, these ideas have been recently used to solve constraint problems on the sphere.[47–49] There are other approaches to the use of interacting particles system in optimization, including the use of individual gradient dynamics coupled through a graph Laplacian.[50–53]

The development of the EKI into the EKS suggests a parallel development of CBO into a sampling methodology. In this paper we pursue this idea and develop CBS. A key property of the EKS is that it is affine invariant[54] as shown in Ref. 36 where the Affine Invariant Interacting Langevin Dynamics (ALDI) algorithm is introduced; relatedly, in the mean-field limit, the rate of convergence to the posterior is the same for all Gaussian posterior distributions.[34] We will show identical properties for the CBS algorithm. Our focus is on unimodal distributions and obtaining Gaussian approximations to the target distribution. We note, however, that there are recent forays into the use of ensemble Kalman methods for the sampling of multimodal distributions.[55,56] Furthermore, there is also recent work extending ensemble Kalman methods to inverse problems beyond the setting of additive Gaussian noise; more complex loss functions, such as cross-entropy and those arising in logistic regression[57,58] are considered. And finally, recent work shows that ensemble methods automatically smooth noisy likelihood functions, essentially denoising rough energy landscapes.[59] Similar developments for the CBS methodology proposed here would also be of interest. Like the EKS, the CBS approach is only exact for Gaussian problems and in the mean-field limit. However, recently developed methods based on multiscale stochastic dynamics provide a refinable methodology for sampling from non-Gaussian distributions[60]; methods such as CBS or EKS may be used to precondition these multiscale stochastic dynamics algorithms, making them more efficient. Alternatively, the CBS method may be used in the calibration step employed within the calibrate-emulate-sample methodology introduced in Ref. 61. Thus, the methods developed in this paper potentially form an important component in an efficient and rigorously justifiable approach to solving Bayesian inverse problems.

1.3 | Our contributions

We introduce CBS as a method to approximate probability distributions of the form (2), or to find the MAP estimator (4). The method requires G only as a black-box (it is derivative-free) and hence is of potential use for large-scale inverse problems. We study the proposed algorithm in settings where the posterior is Gaussian or close to Gaussian. We reemphasize that the computational methodology does not require the specific choice of f in (3), it applies to arbitrary measures with negative log density f, up to an additive constant; however some of our analysis exploits the specific form in (3) in the case where G is linear. We show the following:

- in the case of linear G, and in the mean-field limit, parameters can be chosen in the algorithm so that, if initiated at a Gaussian, successive iterates remain Gaussian and converge to the Gaussian posterior (2);
- in the case of linear G, and in the mean-field limit, parameters can be chosen in the algorithm so that, if initiated at a Gaussian, successive iterates remain Gaussian and converge to a Dirac located at the MAP point θ_* given by (4);
- the CBS method is affine invariant and, in the case of linear G and in the mean-field limit, converges at the same rate across all linear inverse problems defined by (2); for linear G, we obtain sharp convergence rates that are explicit in terms of all parameters of the method;
- in the case of nonlinear G, and in the mean-field limit, parameters can be chosen in the algorithm so that it has a steady-state solution, which is Gaussian, close to the Laplace approximation of the posterior (2) and the algorithm is a local contraction mapping in the neighborhood of the steady state; we make explicit the dependence of this approximation, and its rate of attraction, on the parameters of the method;
- we present numerical results illustrating the foregoing theory and, more generally, demonstrating the viability of the CBS scheme for sampling posterior distributions and for finding MAP estimators.

The results are in arbitrary dimension d, with the exception of the results concerning the Laplace approximation, which are restricted to $d = 1$. There are no intrinsic barriers to extending the Laplace approximation results to arbitrary dimension, but doing so will be technically involved and would lose the focus of the paper.

In Section 2 we introduce the method, including its continuous time limit, and mean-field limits in both discrete and continuous time; we establish its properties in the Gaussian setting. Section 3 contains analysis of the method beyond the Gaussian setting, deriving conditions for convergence to an approximation of the MAP estimator when in optimization mode, and for convergence to the Laplace approximation of the target measure when in sampling mode. In Section 4 we provide the numerical experiments. Proofs of most of the theoretical results in Sections 2 and 3 are presented in Section 5.

2 | PRESENTATION OF THE METHOD

We propose a novel method for sampling and optimization tasks based on a system of interacting particles. Our goals are the following:

(1) Sampling: to generate approximate samples from the posterior distribution (2); this allows to understand the distribution of parameters taking into account both model (1) and the available data y.
(2) Optimization: to find the minimizer of $f(\bullet)$, which corresponds to the MAP point (4), the most likely parameter θ given the data y and the model relating them.

To introduce the approach, we start by defining the mean-field limits of the algorithms, in discrete and continuous time; later we explain how particle approximations of the mean-field limit lead to implementable algorithms. We will be interested in the following McKean difference equation: given parameters $\lambda > 0$, $\beta > 0$, and $\alpha \in [0, 1)$,

$$\begin{cases} \theta_{n+1} = \mathcal{M}_\beta(\rho_n) + \alpha\big(\theta_n - \mathcal{M}_\beta(\rho_n)\big) + \sqrt{(1-\alpha^2)\,\lambda^{-1}\mathcal{C}_\beta(\rho_n)}\,\boldsymbol{\xi}_n, \\ \rho_n = \mathrm{Law}(\theta_n), \end{cases} \tag{5}$$

where $\boldsymbol{\xi}_n$, for $n \in \{0, 1, \dots\}$ are independent $\mathsf{N}(\mathbf{0}, I_d)$ random variables, and $\mathcal{M}_\beta, \mathcal{C}_\beta$ denote, respectively, the mean and variance for a suitable reweighting of measures:

$$\mathcal{M}_\beta : \rho \mapsto \mathcal{M}(L_\beta \rho), \quad \mathcal{C}_\beta : \rho \mapsto \mathcal{C}(L_\beta \rho), \quad L_\beta : \rho \mapsto \frac{\rho \mathrm{e}^{-\beta f}}{\int_{\mathbf{R}^d} \rho \mathrm{e}^{-\beta f}} \tag{6a}$$

$$\mathcal{M}(\mu) = \int_{\mathbf{R}^d} \theta \mu(d\theta), \quad \mathcal{C}(\mu) = \int_{\mathbf{R}^d} \big(\theta - \mathcal{M}(\mu)\big) \otimes \big(\theta - \mathcal{M}(\mu)\big)\, \mu(d\theta). \tag{6b}$$

Letting $\alpha = \exp(-\Delta t)$ and viewing θ_n as a discrete time approximation of a continuous time process $\theta(t)$ at time $t = n\Delta t$, we find that the $\Delta t \to 0$ continuous-time limit associated with these dynamics is the following McKean Stochastic Differential Equation (SDE):

$$\begin{cases} \mathrm{d}\theta_t = -\big(\theta_t - \mathcal{M}_\beta(\rho_t)\big)\,\mathrm{d}t + \sqrt{2\lambda^{-1}\mathcal{C}_\beta(\rho_t)}\,\mathrm{d}\mathbf{W}_t, \\ \rho_t = \mathrm{Law}(\theta_t), \end{cases} \tag{7}$$

where $\mathbf{W}_t$ denotes a standard Brownian motions in $\mathbf{R}^d$. We refer to the two families of methods as CBS methods, parameterized by α, β with the ranges $\alpha \in [0,1)$ corresponding to (5) and $\alpha = 1$ corresponding to (7). Recall that $\beta > 0$. We will focus on two choices of λ: (i) the choice $\lambda = 1$, when the method is used to minimize $f(\bullet)$, which will be referred to as CBS-O(α, β); and (ii) $\lambda = (1+\beta)^{-1}$ when the method is used for sampling the target distribution $e^{-f(\bullet)}$, which will be referred to as CBS(α, β).

In Section 2.1, we introduce the notation used throughout the paper. In Section 2.2 we give motivation for the mean-field stochastic dynamical systems (5) and (7). In Section 2.3 we describe key properties of the mean-field models, and in Section 2.4, we establish convergence to equilibrium for (5) and (7) in the setting where the forward model G is linear and the law of the initial condition is Gaussian. Section 2.5 introduces particle approximations to the mean-field limit.

2.1 | Notation

In what follows, we denote by $g(\bullet; \mathbf{m}, C)$ the density of the Gaussian random variable $\mathsf{N}(\mathbf{m}, C)$:

$$\mathrm{g}(\theta; \mathbf{m}, C) = \frac{1}{\sqrt{(2\pi)^d \det(C)}} \exp\left(-\frac{1}{2}|\theta - \mathbf{m}|_C^2\right). \tag{8}$$

We also use the shorthand notation

$$\mathbf{m}_\beta(\mathbf{m}, C) := \mathcal{M}_\beta\big(\mathrm{g}(\bullet; \mathbf{m}, C)\big), \qquad C_\beta(\mathbf{m}, C) := \mathcal{C}_\beta\big(\mathrm{g}(\bullet; \mathbf{m}, C)\big). \tag{9}$$

More generally, we frequently denote $\mathbf{m}_n = \mathcal{M}(\rho_n)$ and $C_n = \mathcal{C}(\rho_n)$ for the standard mean and covariance calculated with respect to a probability measure ρ_n. For a matrix $A \in \mathbf{R}^{d\times d}$, we denote by $\|A\|$ the operator norm induced by the Euclidean vector norm, and by $\|A\|_{\mathrm{F}}$ the Frobenius norm.[1] Sometimes, we will make use of the shorthand notation $\|A\|_B := \|B^{-1/2}AB^{-1/2}\|$ for a given invertible matrix $B \in \mathbf{R}^{d\times d}$. We let $\mathbf{N} := \{0, 1, 2, 3, \dots\}$ and $\mathbf{N}_{>0} := \{1, 2, 3, \dots\}$, and we denote by $\mathcal{S}_{++}^d$ the set of symmetric strictly positive definite matrices in $\mathbf{R}^{d\times d}$. For symmetric matrices X and Y, the notation $X \succcurlyeq Y$ (resp. $X \preccurlyeq Y$) means that $X - Y$ is positive semidefinite (resp. negative semidefinite).

2.2 | Motivation

The mean-field model (5) contains a number of tuneable parameters. In this section, we give intuition about the role of these parameters in effecting approximate sampling or optimization for the

[1] The Frobenius norm on matrices should not to be confused with the norm $|\mathbf{u}|_A := \langle \mathbf{u}, A^{-1}\mathbf{u}\rangle^{\frac{1}{2}}$ on vectors defined previously.

inverse problem defined by (1). We motivate sampling primarily through the discrete time mean-field model and optimization primarily through the continuous time mean-field model. However, both discrete and continuous time models apply to optimization and to sampling. In practice, the mean-field SDEs in this subsection can be made into algorithms by invoking finite particle approximations, as described in Section 2.5.

2.2.1 | Sampling

Let $G(\bullet) = G\bullet$ be a linear map so that the posterior distribution given by (2) is Gaussian, and denote this Gaussian by $\mathsf{N}(\mathbf{a}, A)$. The mean $\mathbf{a}$ and covariance A may be identified by completing the square in (2): f is of the form $\frac{1}{2}|\theta - \mathbf{a}|_A^2$.

To motivate the algorithms that are the object of study in this paper we describe parameter choices for which the iteration (5) has equilibrium distribution given by the Gaussian $\mathsf{N}(\mathbf{a}, A)$. For any choice of forward model G, it can be shown that the evolution of the first and second moments is given by

$$\mathcal{M}(\rho_{n+1}) = \alpha\mathcal{M}(\rho_n) + (1-\alpha)\mathcal{M}_\beta(\rho_n), \tag{10a}$$

$$\mathcal{C}(\rho_{n+1}) = \alpha^2\mathcal{C}(\rho_n) + \lambda^{-1}(1-\alpha^2)\mathcal{C}_\beta(\rho_n). \tag{10b}$$

From these identities it is clear that any fixed point of the mean and covariance is independent of α. Further, when the initial distribution ρ_0 is Gaussian the systems of Equations (5) for $\alpha \in [0,1)$ map Gaussians into Gaussians. Computing the relationship between the mean and covariance of the Gaussian ρ and the mean and covariance of the Gaussian $L_\beta\rho$ gives

$$\mathbf{m}_\beta(\mathbf{m}, C) = \left(C^{-1} + \beta A^{-1}\right)^{-1}\left(\beta A^{-1}\mathbf{a} + C^{-1}\mathbf{m}\right), \tag{11a}$$

$$C_\beta(\mathbf{m}, C) = \left(C^{-1} + \beta A^{-1}\right)^{-1}. \tag{11b}$$

Therefore, the mean and covariance of a nondegenerate Gaussian steady-state $g(\bullet; \mathbf{m}_\infty, C_\infty)$ for (5) satisfies

$$\mathbf{m}_\infty = \left(C_\infty^{-1} + \beta A^{-1}\right)^{-1}\left(\beta A^{-1}\mathbf{a} + C_\infty^{-1}\mathbf{m}_\infty\right),$$

$$C_\infty = \lambda^{-1}\left(C_\infty^{-1} + \beta A^{-1}\right)^{-1}.$$

This has solution

$$\mathbf{m}_\infty = \mathbf{a}, \qquad C_\infty = \frac{1-\lambda}{\lambda\beta}A.$$

Choosing $\lambda^{-1} = 1 + \beta$ delivers a steady state equal to the posterior distribution. This motivates our choice of λ in the sampling case. Furthermore, choosing $\lambda = 1$ is seen to be natural in the optimization setting: the fixed point of the iteration is then a Dirac at the MAP estimator $\mathbf{a}$. We will demonstrate that these two distinguished choices of λ work well for sampling and optimization, beyond the setting of a Gaussian posterior $\mathsf{N}(\mathbf{a}, A)$.

Remark 1 (Enlarging the Choice of Parameters). The mean-field dynamics (5) can be generalized to the form

$$\theta_{n+1} = p_1\theta_n + p_2\mathcal{M}(\rho_n) + p_3\mathcal{M}_\beta(\rho_n) + \sqrt{p_4\mathcal{C}(\rho_n) + p_5\mathcal{C}_\beta(\rho_n)}\,\boldsymbol{\xi}_n, \qquad \rho_n = \mathrm{Law}(\theta_n), \tag{12}$$

where $(\boldsymbol{\xi}_n)_{n=0,1,\ldots}$ are independent $\mathsf{N}(\mathbf{0}, I_d)$ random variables. Given β, one can ask the following question: for what values of the parameters $(p_1, p_2, p_3, p_4, p_5)$ does the dynamics (12) admit the Gaussian $\mathsf{N}(\mathbf{a}, A)$ as an equilibrium distribution? A calculation analogous to that above shows that $\mathsf{N}(\mathbf{a}, A)$ is a steady state of (12) if and only if

$$p_1 + p_2 + p_3 = 1, \tag{13a}$$

$$p_1^2 + p_4 + p_5(1+\beta)^{-1} = 1. \tag{13b}$$

Note that these constraints do not guarantee that $\mathsf{N}(\mathbf{a}, A)$ is the only steady state, and in fact, if $p_1 = 1$ and $p_2 = p_3 = p_4 = p_5 = 0$, then any distribution is a steady state. In this paper, we study only the dynamics (5), which corresponds to the special case where $p_2 = p_4 = 0$ and $p_1 = \alpha$, $p_3 = 1 - \alpha$ and $p_5 = \lambda^{-1}(1 - \alpha^2)$, but it is potentially useful to exploit this wider class of mean-field models.

2.2.2 | Optimization

We now discuss the algorithm in optimization mode, through the lens of the continuous time limit. Another starting point triggering the research in this paper is the use of systems of interacting particles for minimizing a target function $f(\theta)$. Refs. 7, 9 introduce the CBO technique for achieving this aim by means of particle approximations of the stochastic dynamical system

$$\dot{\theta} = -(\theta - \bar{\theta}) + \sigma|\theta - \bar{\theta}|\,\dot{\mathbf{W}}, \qquad \bar{\theta} = \mathcal{M}_\beta(\rho_t), \tag{14}$$

where $\mathbf{W}$ is a standard Brownian motion in $\mathbf{R}^d$, $\sigma > 0$ is the noise strength, and ρ_t is the law of θ. The idea behind the CBO method is to think about realizations of θ as explorers, in the landscape of the function $f(\theta)$, which can continuously exchange the evaluation of the function f at their position θ, through $\mathcal{M}_\beta(\rho_t)$. Then, the explorers compute a weighted average of their position in parameter space and direct their relaxation movement toward this average $\bar{\theta}$; this explains the first term on the right-hand side of (14). The role of the second term is to impose the property of noise strength decreasing proportionally to the distance of the explorer to the weighted average $\bar{\theta}$. The choice of the weighted average promotes the concentration toward parameter points θ leading to smaller values of f. The resulting law of the system converges as $t \to \infty$ toward a Dirac mass concentrated at the MAP point θ_*, the global minimizer of f, under certain conditions on f; see Refs. 7, 45. The weighted covariance $\bar{\mathcal{C}} = \mathcal{C}_\beta(\rho_t)$ provides an alternative to the cooling schedule in

(14) by way of using $\bar{C} = \mathcal{C}_\beta(\rho_t)$ as the modulation of the noise. In other words, one could propose as alternative to the CBO method (14), the following mean-field system

$$\dot{\theta} = -(\theta - \bar{\theta}) + \sqrt{2\bar{C}}\,\dot{\mathbf{W}}. \tag{15}$$

This gives (7) in the optimization mode $\lambda = 1$. We show in Proposition 3 for the quadratic case, and Proposition 6 for the one-dimensional convex case, that (15) converges precisely to the minimizer of f, whereas the CBO method usually concentrates to a point in the vicinity of the minimizer, with an error depending on β. On the other hand, while the CBO dynamics concentrates exponentially fast under rather general assumptions on f, including the multi-dimensional nonconvex setting,[7,8] the dynamics (15) converges algebraically in time and our proofs concern only simple settings, considering quadratic or one-dimensional convex functions f. Adapting the parameter β during the evolution is shown empirically to improve the rate of convergence for (15), see the discussions in Section 4; but analysis is needed to understand this property. Other differences between the methods are that, unlike CBO, the dynamics (15) is affine invariant (see Section 2.3.2) and satisfies the invariant subspace property (see Lemma 4), although further investigation is necessary to determine whether these two properties are useful in the context of optimization.

In terms of time complexity, one iteration of (the particle approximations of) either method requires the evaluation of f at all the particles; thus, in the context of Bayesian inverse problems where evaluating the forward model is the dominating computational expense, the methods have a similar computational cost per iteration. For problems where the dimension of the state space is very large and evaluation of f is cheap, however, the particle method corresponding to (15) is slightly more expensive than that of (14), as it requires calculating the square root of large matrices $\bar{C}$. We note, however, that employing a generalized square root as proposed in Ref. 36 for the ALDI method would help to mitigate this difficulty.

2.3 | Key properties of the mean-field limits

In this subsection, we summarize key properties of the stochastic dynamics (5) and (7). We consider, in turn: (i) the time evolution of the laws; (ii) the affine invariance; (iii) the steady states; (iv) the evolution of the first and second moments; and (v) propagation properties for Gaussian initial conditions.

2.3.1 | Evolution equations for the law of the mean-field dynamical systems

The time evolution of the law of the solution (5) is governed by the following discrete-time dynamics on probability densities:

$$\rho_{n+1}(\theta) = \int_{\mathbf{R}^d} g\big(\theta; \mathcal{M}_\beta(\rho_n) + \alpha\big(u - \mathcal{M}_\beta(\rho_n)\big), (1-\alpha^2)\,\lambda^{-1}\,\mathcal{C}_\beta(\rho_n)\big)\,\rho_n(u)\,\mathrm{d}u. \tag{16}$$

When $\alpha = 0$, the map (16) takes a particularly simple form (recalling notation 8 for a Gaussian):

$$\rho_{n+1} = g\big(\theta; \mathcal{M}_\beta(\rho_n), \lambda^{-1}\,\mathcal{C}_\beta(\rho_n)\big). \tag{17}$$

Likewise, the time evolution of the law of the solution to (7) is governed by the following nonlinear and nonlocal Fokker–Planck equation:

$$\frac{\partial \rho}{\partial t} = \nabla \cdot \Big(\big(\theta - \mathcal{M}_\beta(\rho)\big)\rho + \lambda^{-1} \mathcal{C}_\beta(\rho) \nabla \rho \Big). \tag{18}$$

Remark 2. We will not discuss here the question of existence and uniqueness of solutions to (18), and we assume from now on that there exists a unique strong solution to (18) for smooth initial data $\rho_0 \in \mathcal{P}_2(\mathbf{R}^d)$, implying in turn the existence and uniqueness of a solution to (7). Equation (18) will be analyzed in subsequent work.

2.3.2 | Affine invariance

A fundamental property of both (5) and (7) is that they are affine invariant, in the sense of Ref. 54; the utility of this concept has been established for MCMC methods in Ref. 62 and for Langevin-based dynamics through the ALDI algorithm in Ref. 36. For linear inverse problems with posterior $\mathsf{N}(\mathbf{a}, A)$ this has the consequence that the rate of convergence is independent of the conditioning of A. We study affine invariance of (5); a similar reasoning can be employed to show that the continuous-time mean-field dynamics (7) are also affine invariant.

To demonstrate affine invariance for (5), let $\{\theta_n\}_{n\in\mathbf{N}}$ denote the solution to (5) with initial condition $\theta_0 \sim \rho_0$, and let $\rho_n = \mathrm{Law}(\theta_n)$. Consider a vector $\mathbf{b} \in \mathbf{R}^d$ and an invertible matrix $B \in \mathbf{R}^{d\times d}$, which, together, define the affine transformation $\theta \mapsto B\theta + \mathbf{b}$. We introduce the following notation:

$$\widetilde{\theta}_n = B\theta_n + \mathbf{b}, \qquad \widetilde{f}(\widetilde{\theta}) = f\Big(B^{-1}(\widetilde{\theta} - \mathbf{b})\Big), \qquad \widetilde{L}_\beta : \mu \mapsto \frac{\mu\, \mathrm{e}^{-\beta \widetilde{f}}}{\int_{\mathbf{R}^d} \mathrm{e}^{-\beta \widetilde{f}}}.$$

We also introduce $\widetilde{\mathcal{M}}_\beta : \mu \mapsto \mathcal{M}(\widetilde{L}_\beta \mu)$ and $\widetilde{\mathcal{C}}_\beta : \mu \mapsto \mathcal{C}(\widetilde{L}_\beta \mu)$. To prove the affine invariance of the scheme (5), we must show that $\{\widetilde{\theta}_n\}_{n\in\mathbf{N}}$ is equal in law to the solution $\{\widehat{\theta}_n\}_{n\in\mathbf{N}}$ of

$$\widehat{\theta}_{n+1} = \alpha \widehat{\theta}_n + (1-\alpha)\widetilde{\mathcal{M}}_\beta(\widehat{\rho}_n) + \sqrt{(1-\alpha^2)\lambda^{-1}\widetilde{\mathcal{C}}_\beta(\widehat{\rho}_n)}\, \widehat{\boldsymbol{\xi}}_n, \qquad \widehat{\rho}_n = \mathrm{Law}(\widehat{\theta}_n), \tag{19}$$

with initial condition $\widehat{\theta}_0 = \widetilde{\theta}_0$ and where $\{\widehat{\boldsymbol{\xi}}_n\}_{n\in\mathbf{N}}$ are independent $\mathsf{N}(\mathbf{0}, I_d)$ random variables. To show this, we apply the affine transformation $\theta \mapsto B\theta + \mathbf{b}$ to both sides of (5), which leads to

$$\widetilde{\theta}_{n+1} = \alpha \widetilde{\theta}_n + (1-\alpha)\big(B\mathcal{M}_\beta(\rho_n) + \mathbf{b}\big) + B\sqrt{(1-\alpha^2)\lambda^{-1}\mathcal{C}_\beta(\rho_n)}\, \xi_n, \qquad \rho_n = \mathrm{Law}(\theta_n).$$

Now notice that $B\mathcal{M}_\beta(\rho_n) + \mathbf{b} = \widetilde{\mathcal{M}}(\widetilde{\rho}_n)$, where $\widetilde{\rho}_n = \mathrm{Law}(\widetilde{\theta}_n)$, that

$$B\sqrt{\mathcal{C}_\beta(\rho_n)}\,\boldsymbol{\xi}_n = \sqrt{B\mathcal{C}_\beta(\rho_n)B^\top}\,\boldsymbol{\xi}_n \quad \text{in law},$$

and that $B\mathcal{C}_\beta(\rho_n)B^\top = \widetilde{\mathcal{C}}_\beta(\widetilde{\rho}_n)$, which implies that $\{\widetilde{\theta}_n\}_{n\in\mathbf{N}}$ is indeed a solution to (19).

2.3.3 | Steady states

The steady states of (5) and (7) coincide, if they exist, and they are necessarily Gaussian. Recall the notation (8). We have:

Lemma 1. *Let probability distribution ρ_∞ have finite second moment and be a steady-state solution of (16) or (18). Then*

$$\rho_\infty(\bullet) = \mathrm{g}\big(\bullet; \mathcal{M}_\beta(\rho_\infty), \lambda^{-1}\mathcal{C}_\beta(\rho_\infty)\big). \tag{20}$$

Conversely, all probability distributions solving (20) are steady states of (16) and (18). In particular, all steady states are Gaussian (with the limiting case of Diracs included in the definition) and all Dirac masses are steady states.

Proof. If ρ_∞ is an invariant measure for the law of (7), then ρ_∞ must be an invariant measure of the following SDE:

$$\mathrm{d}\theta_t = -\big(\theta_t - \mathcal{M}_\beta(\rho_\infty)\big)\,\mathrm{d}t + \sqrt{2\lambda^{-1}\mathcal{C}_\beta(\rho_\infty)}\,\mathrm{d}\mathbf{W}_t. \tag{21}$$

Because this is just the Ornstein–Uhlenbeck process, we deduce (20).

Similarly, if ρ_∞ is an invariant measure for the law of the discrete-time dynamics (5), then ρ_∞ is the invariant measure of the following equation:

$$X_{n+1} = \mathcal{M}_\beta(\rho_\infty) + \alpha\big(X_n - \mathcal{M}_\beta(\rho_\infty)\big) + \sqrt{(1-\alpha^2)\lambda^{-1}\mathcal{C}_\beta(\rho_\infty)}\boldsymbol{\xi}_n,$$

where $(\boldsymbol{\xi}_n)_{n=0,1,\dots}$ are independent $\mathsf{N}(\mathbf{0}, I_d)$ random variables. Because this equation is an exact discretization of (21), we deduce that (20) holds. ■

2.3.4 | Equations for the moments

The evolution equations for the moments given in (10) hold regardless of whether ρ_n is Gaussian but they define closed equations characterizing ρ_n completely in settings where ρ_0 is Gaussian. The evolution of the moments can also be written for the limiting continuous time stochastic dynamical system (7) obtained when $\alpha \to 1$:

$$\partial_t(\mathcal{M}(\rho)) = -\mathcal{M}(\rho) + \mathcal{M}_\beta(\rho), \tag{22a}$$

$$\partial_t(\mathcal{C}(\rho)) = -2\mathcal{C}(\rho) + 2\lambda^{-1}\mathcal{C}_\beta(\rho). \tag{22b}$$

2.3.5 | Propagation of Gaussians

We show that Gaussianity is preserved along the flow, both in discrete and continuous time.

Lemma 2. *Let $\lambda \in (0, 1]$ and $\beta > 0$.*

(i) Discrete time $\alpha = 0$. The law of (5) is Gaussian for all $n \in \mathbf{N}$.
(ii) Discrete time $\alpha \in (0, 1)$. If the initial law ρ_0 for (5) is Gaussian, then so is the law for any $n \in \mathbf{N}_{>0}$, and the time evolution of the moments $(\mathbf{m}_n, C_n)$ of ρ_n is governed by the recurrence relation

$$\mathbf{m}_{n+1} = \alpha \mathbf{m}_n + (1 - \alpha)\mathbf{m}_\beta(\mathbf{m}_n, C_n), \tag{23a}$$

$$C_{n+1} = \alpha^2 C_n + \lambda^{-1}(1 - \alpha^2) C_\beta(\mathbf{m}_n, C_n), \tag{23b}$$

with $\mathbf{m}_\beta$, C_β given by (9).
(iii) Continuous time $\alpha \to 1$. If the initial law ρ_0 for (7) is Gaussian, then so is the corresponding law for any $t > 0$. The time evolution of the moments $(\mathbf{m}(t), C(t))$ of the solution is governed by the equation

$$\dot{\mathbf{m}} = -\mathbf{m} + \mathbf{m}_\beta(\mathbf{m}, C), \tag{24a}$$

$$\dot{C} = -2C + 2\lambda^{-1} C_\beta(\mathbf{m}, C). \tag{24b}$$

Proof. For the discrete-time dynamics in setting (i), this follows directly from (17). For (ii) note that, if $\theta_n \sim \mathsf{N}(\mathbf{m}_n, C_n)$, then θ_{n+1}, being the sum of Gaussian random variables as given in (5), is also normally distributed.

To show (iii), we consider a solution $(\mathbf{m}(t), C(t))$ to the moment Equations (24). Then, $g(\theta; \mathbf{m}(t), C(t))$ solves (18). To see this, one can verify that general Gaussians $g(\theta; \mathbf{m}, C)$ satisfy the relations

$$\nabla_\theta g = -\nabla_{\mathbf{m}} g, \qquad x^T(\mathrm{D}_\theta^2\, g)y = 2D_C g : (x \otimes y),$$

for any $x, y \in \mathbf{R}^d$; see similar computations in Refs. 34, 35. The first identity can be checked directly, and the second identity follows, e.g., from equations (57) and (61) in Ref. 63. Then

$$\begin{aligned} \frac{\partial}{\partial t}\big(g(\theta, \mathbf{m}(t), C(t))\big) &= \nabla_{\mathbf{m}} g \cdot \dot{\mathbf{m}} + D_C g : \dot{C} \\ &= -\nabla_{\mathbf{m}} g \cdot \big(\mathbf{m} - \mathbf{m}_\beta\big) + 2D_C g : \big(\lambda^{-1} C_\beta - C\big) \\ &= \nabla_\theta g \cdot \big(\mathbf{m} - \mathbf{m}_\beta\big) + \nabla_\theta \cdot (-C \nabla_\theta g) + \lambda^{-1} D_\theta^2 g : C_\beta \\ &= \nabla_\theta \cdot \big((\theta - \mathbf{m}_\beta) g + \lambda^{-1}\, C_\beta\, \nabla_\theta g\big), \end{aligned}$$

where we used the explicit expression of $C\nabla_\theta g$ in the last equation. ■

TABLE 1 Convergence rates for CBS in sampling and optimization modes, in the case of a Gaussian target distribution and a Gaussian initial condition with $C_0 \in \mathcal{S}^d_{++}$. This table summarizes the results in Propositions 1 to 3. All rates are sharp, see Remark 4

	Sampling		**Optimization**	
	Mean	**Covariance**	**Mean**	**Covariance**
$\alpha = 0$	$\left(\frac{1}{1+\beta}\right)^n$	$\left(\frac{1}{1+\beta}\right)^n$	$\frac{k_0}{k_0+\beta n}$	$\frac{k_0}{k_0+\beta n}$
$\alpha \in (0, 1)$	$\left(\frac{1+\alpha\beta}{1+\beta}\right)^n$	$\left(\frac{1+\alpha^2\beta}{1+\beta}\right)^n$	$\left(\frac{k_0+\beta}{k_0+\beta+\beta(1-\alpha^2)n}\right)^{\frac{1}{1+\alpha}}$	$\frac{k_0+\beta}{k_0+\beta+\beta(1-\alpha^2)n}$
$\alpha = 1$	$\mathrm{e}^{-\left(\frac{\beta}{1+\beta}\right)t}$	$\mathrm{e}^{-\left(\frac{2\beta}{1+\beta}\right)t}$	$\left(\frac{k_0+\beta}{k_0+\beta+2\beta t}\right)^{\frac{1}{2}}$	$\frac{k_0+\beta}{k_0+\beta+2\beta t}$

2.4 | Convergence for Gaussian targets

In this subsection, we consider the case of a linear forward map in (1), leading to the posterior distribution being a Gaussian $\mathsf{N}(\mathbf{a}, A)$ where, throughout, we assume that A is strictly positive definite, $A \in \mathcal{S}^d_{++}$. The corresponding potential $f(\cdot)$. The corresponding potential $f(\cdot)$ is given by the quadratic function $f(\theta) = \frac{1}{2}|\theta - \mathbf{a}|^2_A$. Recall the shorthand notation $\|B\|_A = \|A^{-1/2}BA^{-1/2}\|$. Throughout this section, we denote

$$k_0 = \|C_0^{-1}\|_{A^{-1}} = \|A^{1/2}C_0^{-1}A^{1/2}\|.$$

The main convergence results of this subsection, Propositions 1 to 3, establish the convergence of the moments of the solutions to (5) and (7), respectively, in the case of Gaussian initial conditions. All results show algebraic convergence in optimization mode ($\lambda = 1$) and exponential convergence in sampling mode ($\lambda = (1 + \beta)^{-1}$); this is analogous to what is known about the EKI[32] and the EKS[34] methods. We provide in Table 1 an overview of the results we obtain. Most proofs of the results presented in the rest of this subsection are given in Section 5.1.

We draw a number of conclusions from these results. First, in the discrete time setting, smaller choices of α provide a faster rate of convergence, and choosing $\alpha = 0$ is therefore the most favorable choice in this regard. Second, larger choices of β increase the speed of convergence, without limit as $\beta \to \infty$ for $\alpha = 0$; in the case $\alpha > 0$, increasing β is favorable but does not give rates, which increase without limit.

2.4.1 | Convergence analysis for the discrete-time dynamics

Using the explicit expression of the weighted moments in the Gaussian case (11), we can rewrite the right-hand sides of Equation (23) as

$$(\mathbf{m}_{n+1} - \mathbf{a}) = \left[\alpha I_d + (1-\alpha)A(A + \beta C_n)^{-1}\right](\mathbf{m}_n - \mathbf{a}),$$
$$C_{n+1} = \left[\alpha^2 I_d + (1-\alpha^2)\lambda^{-1}A(A + \beta C_n)^{-1}\right]C_n\,.$$

Letting $\widetilde{\mathbf{m}}_n := A^{-1/2}(\mathbf{m}_n - \mathbf{a})$ and $\widetilde{C}_n := \beta A^{-1/2} C_n A^{-1/2}$, we can verify that $(\widetilde{\mathbf{m}}_n, \widetilde{C}_n)_{n \in \mathbf{N}}$ solves the following recurrence relation:

$$\widetilde{\mathbf{m}}_{n+1} = \left[\alpha I_d + (1-\alpha)(I_d + \widetilde{C}_n)^{-1}\right] \widetilde{\mathbf{m}}_n \,, \tag{25a}$$

$$\widetilde{C}_{n+1} = \left[\alpha^2 I_d + (1-\alpha^2)\lambda^{-1}(I_d + \widetilde{C}_n)^{-1}\right] \widetilde{C}_n \,. \tag{25b}$$

This is a recurrence relation uniquely solvable given initial conditions $(\widetilde{\mathbf{m}}_0, \widetilde{C}_0)$. We begin by studying the easier case $\alpha = 0$, where the convergence of the scheme can be computed explicitly by a direct argument.

Lemma 3. *Consider the iterative scheme (17) with $\alpha = 0$ and initial conditions $(\mathbf{m}_0, C_0) \in \mathbf{R}^d \times \mathcal{S}^d_{++}$. Then, for any $\lambda \in (0,1]$ and $\beta > 0$, we have*

$$\mathbf{m}_n = \mathbf{a} + \lambda^n C_n C_0^{-1}(\mathbf{m}_0 - \mathbf{a}), \qquad C_n^{-1} = \begin{cases} \lambda^n C_0^{-1} + (1-\lambda^n) C_\infty^{-1} & \text{if } \lambda \neq 1, \\ C_0^{-1} + n\beta A^{-1} & \text{if } \lambda = 1. \end{cases}$$

Proof. When $\alpha = 0$, the evolution Equations (25) for the moments simplify to

$$\widetilde{\mathbf{m}}_{n+1} = (I_d + \widetilde{C}_n)^{-1} \widetilde{\mathbf{m}}_n \,, \qquad \widetilde{C}_{n+1}^{-1} = \lambda\left(\widetilde{C}_n^{-1} + I_d\right).$$

For $\lambda = 1$, the result for the covariance matrix is easily obtained by solving the second equation explicitly for $\widetilde{C}_n^{-1}$. Next, consider the case $\lambda \neq 1$. We have

$$\widetilde{C}_n^{-1} = \lambda^n \widetilde{C}_0^{-1} + (\lambda + \cdots + \lambda^n) I_d = \lambda^n \widetilde{C}_0^{-1} + \lambda \left(\frac{1-\lambda^n}{1-\lambda}\right) I_d \,.$$

For the evolution of the mean, notice that

$$\widetilde{C}_{n+1}^{-1} \widetilde{\mathbf{m}}_{n+1} = \lambda\left(\widetilde{C}_n^{-1} + I_d\right)(I_d + \widetilde{C}_n)^{-1} \widetilde{\mathbf{m}}_n = \lambda \widetilde{C}_n^{-1} \widetilde{\mathbf{m}}_n \,.$$

Hence, $\widetilde{\mathbf{m}}_n = \lambda^n \widetilde{C}_n \widetilde{C}_0^{-1} \widetilde{\mathbf{m}}_0$ and the result follows. ■

We deduce from this result a convergence estimate for the mean and the covariance of the iterates.

Proposition 1. *Consider the iterative scheme (17) with $\alpha = 0$ and initial conditions $(\mathbf{m}_0, C_0) \in \mathbf{R}^d \times \mathcal{S}^d_{++}$. Then the following statements hold:*

(i) Sampling mode $\lambda = (1+\beta)^{-1}$. For all $n \in \mathbf{N}$, it holds that

$$|\mathbf{m}_n - \mathbf{a}|_A \leqslant \max(1, k_0) \lambda^n |\mathbf{m}_0 - \mathbf{a}|_A \,,$$
$$\|C_n - A\|_A \leqslant \max(1, k_0) \lambda^n \|C_0 - A\|_A \,.$$

(ii) Optimization mode $\lambda = 1$. For all $n \in \mathbf{N}$, it holds that

$$|\mathbf{m}_n - \mathbf{a}|_A \leqslant \left(\frac{k_0}{k_0 + \beta n}\right)|\mathbf{m}_0 - \mathbf{a}|_A\,, \qquad C_n \preccurlyeq \left(\frac{k_0}{k_0 + \beta n}\right)C_0.$$

To study the convergence in the general case $\alpha \in (0, 1)$, we will reduce the evolution of the moments (25) to the scalar case,

$$u_{n+1} = \left[\alpha + (1-\alpha)(1+v_n)^{-1}\right]u_n, \tag{26a}$$

$$v_{n+1} = \left[\alpha^2 + (1-\alpha^2)\lambda^{-1}(1+v_n)^{-1}\right]v_n \tag{26b}$$

by diagonalization. Then, using Lemma A.1, the asymptotic behavior of the moments can be summarized as follows.

Proposition 2. *Consider the iterative scheme (5) with $\alpha \in (0, 1)$ and initial conditions $(\mathbf{m}_0, C_0) \in \mathbf{R}^d \times \mathcal{S}^d_{++}$. Then the following statements hold:*

(i) Sampling mode $\lambda = (1+\beta)^{-1}$. For all $n \in \mathbf{N}$,

$$|\mathbf{m}_n - \mathbf{a}|_A \leqslant \max(1, k_0)^{\frac{1}{1+\alpha}}\big((1-\alpha)\lambda + \alpha\big)^n|\mathbf{m}_0 - \mathbf{a}|_A\,,$$

$$\|C_n - A\|_A \leqslant \max(1, k_0)\big((1-\alpha^2)\lambda + \alpha^2\big)^n\|C_0 - A\|_A\,.$$

(ii) Optimization mode $\lambda = 1$. For all $n \in \mathbf{N}$, it holds that

$$|\mathbf{m}_n - \mathbf{a}|_A \leqslant \left(\frac{k_0 + \beta}{k_0 + \beta + \beta(1-\alpha^2)n}\right)^{\frac{1}{1+\alpha}}|\mathbf{m}_0 - \mathbf{a}|_A\,,$$

$$C_n \preccurlyeq \left(\frac{k_0 + \beta}{k_0 + \beta + \beta(1-\alpha^2)n}\right)C_0\,.$$

2.4.2 | Convergence analysis for the continuous-time dynamics

Next, we consider the limiting case $\alpha \to 1$. Rewriting the right-hand side of (24a) and (24b) using (11), we obtain for any $\lambda \in (0, 1]$ and $\beta > 0$,

$$\dot{\mathbf{m}} = -\beta C(A + \beta C)^{-1}(\mathbf{m} - \mathbf{a}), \tag{27a}$$

$$\dot{C} = -2\beta\, C\,(A + \beta C)^{-1}\left(C - \left(\frac{1-\lambda}{\beta\lambda}\right)A\right). \tag{27b}$$

Proposition 3. *Let $(\mathbf{m}(t), C(t))$ denote the solution to Equations (27) with initial conditions $(\mathbf{m}_0, C_0) \in \mathbf{R}^d \times \mathcal{S}^d_{++}$. Then, the following statements hold:*

(i) Sampling mode $\lambda = (1+\beta)^{-1}$. For all $t > 0$,

$$|\mathbf{m}(t) - \mathbf{a}|_A \leqslant \max\left(1, k_0^{\lambda/2}\right) \mathrm{e}^{-(1-\lambda)t} |\mathbf{m}_0 - \mathbf{a}|_A,$$

$$\|C(t) - A\|_A \leqslant \max\left(1, k_0^{\lambda}\right) \mathrm{e}^{-2(1-\lambda)t} \|C_0 - A\|_A .$$

(ii) Optimization mode $\lambda = 1$. For all $t \geqslant 0$, it holds

$$|\mathbf{m}(t) - \mathbf{a}|_A \leqslant \left(\frac{k_0 + \beta}{k_0 + \beta + 2t\beta}\right)^{\frac{1}{2}} |\mathbf{m}_0 - \mathbf{a}|_A ,$$

$$C(t) \preccurlyeq \left(\frac{k_0 + \beta}{k_0 + \beta + 2t\beta}\right) C_0 .$$

Remark 3 (Discrete to Continuum). Notice that, by letting $\alpha = \mathrm{e}^{-t/n}$ in the convergence results obtained for $\alpha \in (0,1)$ in Proposition 2 and taking the limit $n \to \infty$, we recover the convergence results of the continuous-time setting, up to the constant prefactor.

Remark 4 (Sharpness). It is possible to show, using the lower bounds on the trend to equilibrium provided by Lemmas A.1 and A.2, that the convergence rates we obtained in Propositions 2 and 3 are all sharp with respect to n and t, respectively. Note that the argument leading to Proposition 2 also applies to the case $\alpha = 0$. However, the upper bounds we obtain in Proposition 1 are stronger than those we would be able to obtain by applying Lemma A.1 for $\alpha = 0$. Lower bounds for the sampling mode in the case $\alpha = 0$ can be obtained the same way as for $\alpha \in (0,1)$. In optimization mode ($\lambda = 1$), we can derive lower bounds explicitly using the expression from Lemma 3 as follows: for $\widetilde{C}_n := \beta A^{-1/2} C_n A^{-1/2}$, we have $\tilde{C}_0 \preccurlyeq \|\tilde{C}_0\| I_d$, so

$$\tilde{C}_n^{-1} = \tilde{C}_0^{-1} + n I_d \preccurlyeq \left(1 + n\|\tilde{C}_0\|\right) \tilde{C}_0^{-1} \quad \Rightarrow \quad C_n \succcurlyeq \left(\frac{1}{1 + \beta n \|C_0\|_A}\right) C_0.$$

The conclusion from the above observations is that all rates provided in Table 1 are sharp.

Remark 5 (Attractor). As a consequence of the above convergence results for linear objective functions f, the steady-state $(\mathbf{a}, A)$ is the unique attractor of the moment Equations (23) and (24) when taking an initial condition with $C_0 \in \mathcal{S}_{++}^d$. Therefore, while the mean-field dynamics (16) and (18) admit infinitely many steady states given by all Dirac distributions in addition to the Gaussian steady-state $\mathsf{N}(\mathbf{a}, A)$, the solutions to the mean-field dynamics always converge to the desired target measure $\mathsf{N}(\mathbf{a}, A)$ when initialized at Gaussian initial conditions with $C_0 \in \mathcal{S}_{++}^d$, avoiding the manifold of Diracs along the evolution.

2.5 | Particle approximations

In this subsection we describe particle approximations of the mean-field dynamics (5) and (7). This leads to the implementable algorithms used in Section 4. The following is a discrete-time system of interacting particles in $\mathbf{R}^d$ with mean-field limit given by (5):

$$\theta_{n+1}^{(j)} = \mathcal{M}_\beta(\rho_n^J) + \alpha\Big(\theta_n^{(j)} - \mathcal{M}_\beta(\rho_n^J)\Big) + \sqrt{(1-\alpha^2)\,\lambda^{-1}\mathcal{C}_\beta(\rho_n^J)}\,\boldsymbol{\xi}_n^{(j)}, \qquad j = 1, \dots, J. \tag{30}$$

Here $\boldsymbol{\xi}_n^{(j)}$, for $j \in \{1, \dots, J\}$ and $n \in \mathbf{N}$, are independent $\mathsf{N}(\mathbf{0}, I_d)$ random variables, and ρ_n^J is the empirical measure associated with the particle system at iteration n,

$$\rho_n^J := \frac{1}{J}\sum_{j=1}^{J} \delta_{\theta_n^{(j)}}.$$

We note that

$$\mathcal{M}_\beta(\rho_n^J) = \frac{\sum_{j=1}^J \mathrm{e}^{-\beta f(\theta_n^{(j)})}\,\theta_n^{(j)}}{\sum_{j=1}^J \mathrm{e}^{-\beta f(\theta_n^{(j)})}}, \tag{31a}$$

$$\mathcal{C}_\beta(\rho_n^J) = \frac{\sum_{j=1}^J \Big(\Big(\theta_n^{(j)} - \mathcal{M}_\beta(\rho_n^J)\Big) \otimes \Big(\theta_n^{(j)} - \mathcal{M}_\beta(\rho_n^J)\Big)\Big)\,\mathrm{e}^{-\beta f(\theta_n^{(j)})}}{\sum_{j=1}^J \mathrm{e}^{-\beta f\left(\theta_n^{(j)}\right)}}. \tag{31b}$$

The limit cases $\alpha = 0$ and $\alpha \to 1$ for fixed $\lambda > 0$ and $\beta > 0$ reduce to simpler systems. Indeed, in the case where $\alpha = 0$, the method simplifies to

$$\theta_{n+1}^{(j)} = \mathcal{M}_\beta(\rho_n^J) + \sqrt{\lambda^{-1}\mathcal{C}_\beta(\rho_n^J)}\,\boldsymbol{\xi}_n^{(j)}, \qquad j = 1, \dots, J.$$

On the other hand, when $\alpha \approx 1$, the particle evolution Equation (30) may be viewed as a time discretization with timestep $\Delta t = -\log\alpha$ of the following continuous-time interacting particle system, in which we generalize the notation (31) to continuous time in the obvious way:

$$\dot{\theta}^{(j)} = -\big(\theta^{(j)} - \mathcal{M}_\beta(\rho_t^J)\big) + \sqrt{2\lambda^{-1}\,\mathcal{C}_\beta(\rho_t^J)}\,\dot{\mathbf{W}}^{(j)}, \qquad j = 1, \dots, J, \tag{32}$$

where $\{\mathbf{W}^{(j)}\}_{j=1}^J$ are independent standard Brownian motions in $\mathbf{R}^d$. The formal mean-field limit of this equation is given by (7).

We note that the finite-dimensional particle systems (30) and (32) are both affine invariant; the proof is similar to that given for the mean-field limit. In addition, like ensemble Kalman–based methods for inverse problems ,[64] the particle systems (30) and (32) both satisfy the following invariant subspace property.

Lemma 4. *Let $\mathcal{S}$ denote the linear span of $\{\theta_0^{(j)}\}_{j=1}^J$. Then $\theta_n^{(j)} \in \mathcal{S}$ for all $(j, n) \in \{1, \dots, J\} \times \mathbf{N}$ and $\theta_t^{(j)} \in \mathcal{S}$ for all $(j, t) \in \{1, \dots, J\} \times [0, \infty)$.*

Proof. We prove only the first claim, which follows from a simple recursion. Let us assume the claim is true for $(j, n) \in \{1, \dots, J\} \times \{0, \dots, N\}$ and prove that it is then also true for $n = N + 1$. Let

$\mathbf{a} \in \mathcal{S}^{\perp}$, where $\mathcal{S}^{\perp}$ is the orthogonal complement of $\mathcal{S}$ in $\mathbf{R}^d$. Taking the inner product of both sides of (30) with $\mathbf{a}$, we obtain for all $j \in \{1, \dots, J\}$ that

$$\begin{aligned}\mathbf{a}^{\top}\theta_{N+1}^{(j)} &= \mathbf{a}^{\top}\mathcal{M}_{\beta}(\rho_n^J) + \alpha\mathbf{a}^{\top}\left(\theta_N^{(j)} - \mathcal{M}_{\beta}(\rho_N^J)\right) + \mathbf{a}^{\top}\sqrt{(1-\alpha^2)\lambda^{-1}\mathcal{C}_{\beta}(\rho_N^J)}\,\boldsymbol{\xi}_N^{(j)} \\ &= 0 + 0 + \mathbf{a}^{\top}\sqrt{(1-\alpha^2)\lambda^{-1}\mathcal{C}_{\beta}(\rho_N^J)}\,\boldsymbol{\xi}_N^{(j)},\end{aligned}$$

and so by the Cauchy–Schwarz inequality,

$$\left|\mathbf{a}^{\top}\theta_{N+1}^{(j)}\right|^2 \leqslant (1-\alpha^2)\lambda^{-1}\left|\boldsymbol{\xi}_N^{(j)}\right|^2 \left|\sqrt{\mathcal{C}_{\beta}(\rho_N^J)}\mathbf{a}\right|^2 = 0,$$

because $\mathcal{C}_{\beta}(\rho_N^J)\mathbf{a} = 0$ by the formula (31b) for the weighted covariance $\mathcal{C}_{\beta}(\rho_N^J)$. Because $\mathbf{a}$ was arbitrary in $\mathcal{S}^{\perp}$, the proof is complete. ■

Remark 6 (Cooling schedule). To improve algorithmic implementations it will be of value to develop a rigorous understanding of the relationship between the number of particles J and the parameter β needed to establish good performance of the method. Relatedly, it will also be useful to investigate theoretically the rate of convergence to equilibrium in the setting where a cooling schedule is employed for β. See Section 4 for numerical investigations in this direction.

3 | ANALYSIS BEYOND THE GAUSSIAN SETTING

In this section, we study the proposed method (5) in the case where the function f is not necessarily quadratic, and so the target probability distribution may be non-Gaussian. We begin, in Section 3.1, by presenting preliminary bounds on $\mathbf{m}_{\beta}(\mathbf{m}, C)$ and $C_{\beta}(\mathbf{m}, C)$ defined in (9), and then we analyze the optimization ($\lambda = 1$) and sampling ($\lambda = (1+\beta)^{-1}$) methods in Sections 3.2 and 3.3, respectively. The proofs of all results are presented in Section 5, with the exception of Theorem 2, which is presented in text.

The results in this section are based on the following two assumptions.

Assumption 1 (Convexity of the potential). The function f satisfies $f \in C^2(\mathbf{R}^d)$ and $\mathrm{D}^2 f(\theta) \succcurlyeq L \succcurlyeq \ell I_d$ for all $\theta \in \mathbf{R}^d$, for some $L \in \mathcal{S}_{++}^d$ and some $\ell > 0$.

Assumption 1 guarantees the existence of a unique global minimizer for f, which we will denote throughout this section by

$$\theta_* := \mathop{\arg\min}_{\theta \in \mathbf{R}^d} f(\theta).$$

Assumption 2 (Bound from above on the Hessian). The function f satisfies $f \in C^2(\mathbf{R}^d)$ and $\mathrm{D}^2 f(\theta) \preccurlyeq U \preccurlyeq u I_d$ for all $\theta \in \mathbf{R}^d$, for some $U \in \mathcal{S}_{++}^d$ and some $u > 0$.

These assumptions are very similar to the ones made in Ref. 7 to show the convergence of the CBO method[9] for global optimization. The convergence results we present in this section are summarized in Table 2.

TABLE 2 Sharp upper bounds on the convergence rates for CBS in sampling and optimization modes, in the case of a non-Gaussian target distribution and a Gaussian initial condition with strictly positive definite covariance matrix C_0. Here k is a positive constant independent of n, t, α and β, and $\widetilde{k}_0 := \|L^{1/2} C_0^{-1} L^{1/2}\|$, where L is the symmetric positive definite matrix from Assumption 1, and q is any constant strictly greater than $2\max(2, u/\ell)$, where ℓ and u are the constants from Assumption 1 and 2, respectively. Obtaining sharp convergence rates for the mean in the non-Gaussian case for $\alpha \neq 0$ in optimization mode is an open problem

	Sampling		**Optimization**	
	Mean ($d = 1$)	**Covariance ($d = 1$)**	**Mean ($d = 1$)**	**Covariance (any d)**
$\alpha = 0$	$\left(\frac{k}{\beta}\right)^n$	$\left(\frac{k}{\beta}\right)^n$	$\lesssim \frac{\log(n)}{n}$	$\frac{\widetilde{k}_0}{\widetilde{k}_0+\beta n}$
$\alpha \in (0,1)$	$\left(\alpha + (1-\alpha^2)\frac{k}{\beta}\right)^n$	$\left(\alpha + (1-\alpha^2)\frac{k}{\beta}\right)^n$	$\lesssim n^{-1/q}$ (not optimal)	$\frac{\widetilde{k}_0+\beta}{\widetilde{k}_0+\beta+\beta(1-\alpha^2)n}$
$\alpha = 1$	$\mathrm{e}^{-\left(1-\frac{2k}{\beta}\right)t}$	$\mathrm{e}^{-\left(1-\frac{2k}{\beta}\right)t}$	$\lesssim t^{-1/q}$ (not optimal)	$\frac{\widetilde{k}_0+\beta}{\widetilde{k}_0+\beta+2\beta t}$

3.1 | Preliminary bounds

We first obtain sharp bounds on $\mathcal{C}_\beta$, which, in the special case when f is quadratic, enable to recover (11b). The first bound relies on a logarithmic Sobolev inequality for the probability measure $\frac{1}{Z_\beta}\mathrm{e}^{-\beta f}$, where Z_β is the normalization constant.

Lemma 5 (Upper bound on weighted covariance). *If Assumption 1 holds, then*

$$\forall (\mathbf{m}, C) \in \mathbf{R}^d \times \mathcal{S}^d_{++}, \qquad \mathcal{C}_\beta(\mathbf{m}, C) \preccurlyeq \left(C^{-1} + \beta L\right)^{-1}.$$

Remark 7. We note that, by the standard Holley–Stroock result, see, e.g., Ref. [65, Theorem 2.11], a similar bound could be obtained when f is of the type $f_c + f_b$, where f_c satisfies the convexity property Assumption 1 and f_b is a bounded function.

The next lemma provides a bound from below on $\mathcal{C}_\beta$.

Lemma 6 (Lower bound on weighted covariance). *If Assumption 2 holds, then*

$$\forall (\mathbf{m}, C) \in \mathbf{R}^d \times \mathcal{S}^d_{++}, \qquad \mathcal{C}_\beta(\mathbf{m}, C) \succcurlyeq \left(C^{-1} + \beta U\right)^{-1}.$$

We now obtain a crude bound on the weighted first moment $\mathbf{m}_\beta(\mathbf{m}, C)$, which will be our starting point for establishing the existence of a steady state for the sampling scheme. This bound is useful because it shows that $\mathbf{m}_\beta(\mathbf{m}, C) \xrightarrow[\beta\to\infty]{} \theta_*$ for any fixed $\mathbf{m}$ and $C > 0$.

Lemma 7 (Bound on weighted mean). *If Assumptions 1 and 2 hold, then there exists a positive constant $k = k(\ell, u, d)$ such that,*

$$\forall (\mathbf{m}, C, \beta) \in \mathbf{R}^d \times \mathcal{S}^d_{++} \times \mathbf{R}_{>0}, \qquad |\mathbf{m}_\beta(\mathbf{m}, C) - \theta_*| \leqslant \sqrt{\frac{\|C^{-1}\|}{\ell\beta}}|\mathbf{m} - \theta_*| + k\left(\frac{1}{\|C\|} + \beta\ell\right)^{-1/2}.$$

Unfortunately, this bound degenerates in the limit $C \to 0$. In spatial dimension one, we will obtain, in the proof of Proposition 5, a finer bound on the weighted mean that can be used for proving convergence of the optimization scheme.

3.2 | Analysis of the optimization scheme

In this subsection, we are concerned with the large-time convergence of the law of the solutions to the mean-field evolution Equations (5) and (7) when $\lambda = 1$ and under the following assumption on the initial condition:

Assumption 3 (Nondegenerate Gaussian initial conditions). The initial condition for the mean-field evolution (16) (or (18), in the continuous time setting) is Gaussian with strictly positive definite covariance matrix.

Under this assumption, following Lemma 2, the solutions are normally distributed for all (discrete or continuous) times with the first and second moments evolving according to Equations (23) and (24), respectively. We will show that, under appropriate assumptions, the mean converges to θ_* and the covariance to zero.

Throughout this subsection, we denote by $\{(\mathbf{m}_n, C_n)\}_{n\in\mathbf{N}}$ a solution to (23) with $C_0 \succcurlyeq 0$, and by $\{(\mathbf{m}(t), C(t))\}_{t\in[0,\infty)}$ a solution to (24) with $C(0) \succcurlyeq 0$. We also denote by ρ_n and ρ_t solutions to (16) and (18), respectively.

We begin by showing that the covariance matrices decrease to zero with rates matching those obtained in the case of quadratic f in Section 2.4, up to constant prefactors.

Proposition 4 (Collapse of the ensemble in optimization mode). *Let $\lambda = 1$ and $\beta > 0$ and assume that Assumption 1 holds. Then we have*

(i) Discrete time $\alpha = 0$. If $C_0 \in \mathcal{S}^d_{++}$, then

$$C_n \preccurlyeq \left(\frac{\|C_0^{-1}\|_L}{\|C_0^{-1}\|_L + \beta n}\right) C_0. \tag{33}$$

(ii) Discrete time $\alpha \in (0,1)$. If $C_0 \in \mathcal{S}^d_{++}$, then

$$C_n \preccurlyeq \left(\frac{\|C_0^{-1}\|_L + \beta}{\|C_0^{-1}\|_L + \beta + \beta(1-\alpha^2)n}\right) C_0. \tag{34}$$

(iii) Continuous time $\alpha = 1$. If $C(0) \in \mathcal{S}^d_{++}$, then

$$C(t) \preccurlyeq \left(\frac{\|C(0)^{-1}\|_L + \beta}{\|C(0)^{-1}\|_L + \beta + 2\beta t}\right) C(0). \tag{35}$$

Ideally, we would like to show that $\mathbf{m}_n \xrightarrow[n\to\infty]{} \theta_*$ and $\mathbf{m}(t) \xrightarrow[t\to\infty]{} \theta_*$; however, we were able to show this result only in the one-dimensional setting. In the multidimensional case, we establish the following weaker result.

Theorem 1. *Let $\lambda = 1$, $\beta > 0$, $C_0 \in \mathcal{S}^d_{++}$, and suppose that Assumptions 1 and 2 hold. If there exists $\hat\theta \in \mathbf{R}^d$ such that $\mathbf{m}_n \xrightarrow[n\to\infty]{} \hat\theta$ for some $\alpha \in [0,1)$ or $\mathbf{m}(t) \xrightarrow[t\to\infty]{} \hat\theta$ for $\alpha = 1$, then $\hat\theta = \theta_*$ is the minimizer of f.*

It follows from the identity

$$\forall \mu \in \mathcal{P}_2(\mathbf{R}^d), \qquad W_2(\mu, \delta_{\theta_*})^2 = |\mathcal{M}(\mu) - \theta_*|^2 + \operatorname{tr}(\mathcal{C}(\mu)), \tag{36}$$

where $W_2(\bullet,\bullet)$ denotes the quadratic Wasserstein distance, that Proposition 4 and Theorem 1 can be combined to obtain convergence results for the solutions to the mean-field systems (16) and (18). For example, the following result holds in the discrete-time case.

Corollary 1. *Suppose that Assumptions 1 to 3 hold. If there exists $\hat\theta$ such that $\mathcal{M}(\rho_n) \xrightarrow[n\to\infty]{} \hat\theta$, then $W_2(\rho_n, \delta_{\theta_*}) \xrightarrow[n\to\infty]{} 0$.*

In the one-dimensional case, it is possible to prove the convergence of m_n and $m(t)$ to the minimizer θ_* without the a priori assumption that m_n and $m(t)$ have a limit.

Proposition 5 (Convergence in the one-dimensional case)**.** *Let $d = 1$, $\lambda = 1$, $\beta > 0$, $C_0 \in \mathcal{S}^d_{++}$, and suppose that Assumptions 1 and 2 are satisfied. Then it holds that $m_n \xrightarrow[n\to\infty]{} \theta_*$ for $\alpha \in [0,1)$ and, likewise, $m(t) \xrightarrow[t\to\infty]{} \theta_*$ for $\alpha = 1$.*

As above, this result can be combined with Proposition 4 to obtain a convergence result in Euclidean Wasserstein distance for the solution to (16) and (18), under Assumptions 1 to 3. When deriving this convergence result, we obtain nonoptimal rates of order $n^{-1/r}$ for the case $\alpha = 0$, $n^{-1/2r}$ for $\alpha \in (0,1)$ and $t^{-1/2r}$ for $\alpha = 1$, with $r = r(u,l) > 2$.

To conclude this section, we present a convergence result for m_n with an explicit sharp rate in the particular case $\alpha = 0$.

Proposition 6 (Rate of convergence)**.** *Let $d = 1$, $\lambda = 1$, $\beta > 0$, $\alpha = 0$, $C_0 \in \mathcal{S}^d_{++}$ and suppose that Assumptions 1 and 2 are satisfied. Suppose additionally that $\mathrm{e}^{-\beta f}$ is, together with all its derivatives, bounded from above uniformly in $\mathbf{R}$. Then there exists a positive constant $k = k(m_0, C_0)$ such that, for sufficiently large n,*

$$|m_n - \theta_*| \leqslant k\left(\frac{\log n}{n}\right).$$

The rate of convergence obtained in Proposition 6 is almost optimal in view of the fact shown in Section 2.4 that $|m_n - \theta_*|$ scales with n as $\mathcal{O}(1/n)$ in the case when f is quadratic. We expect the result to extend to other values of α and to the continuous-time solution to (24), but we focus on the case $\alpha = 0$ to avoid overly lengthy and technical proofs. We point out that, already in the Gaussian case, the argument to obtain an optimal decay rate for $\alpha \in (0,1]$ is quite technical.

Finding a simplified argument to prove optimal rates in the optimization setting is an interesting open problem, which we leave for future work.

3.3 | Analysis of the sampling scheme

In this subsection, we investigate the existence of steady states and convergence for the mean-field dynamics associated with the consensus-based samplers, that is when used with $\lambda = (1+\beta)^{-1}$. We consider both the iteration (16) (in the case $\alpha \in [0,1)$) and the nonlocal, nonlinear Fokker–Planck equation (18) (in the case $\alpha = 1$).

We begin by stating an existence result in the multidimensional setting. Because the corresponding proof is very short, we include it in this section.

Theorem 2 (Existence of steady states). *Let $\lambda = (1+\beta)^{-1}$, $\beta > 0$ and $\alpha \in [0,1]$. Suppose Assumptions 1 and 2 are satisfied. Then there exists $\underline{\beta}$ such that, for all $\beta \geqslant \underline{\beta}$, the dynamics (16) and (18) admit a Gaussian steady-state $g(\bullet; \mathbf{m}_\infty(\beta), C_\infty(\beta))$ satisfying*

$$U^{-1} \preccurlyeq C_\infty(\beta) \preccurlyeq L^{-1} \quad \text{and} \quad |\mathbf{m}_\infty(\beta) - \theta_*| = \mathcal{O}\left(\frac{1}{\sqrt{\beta}}\right).$$

Proof. By Lemma 1, a Gaussian $g(\bullet; \mathbf{m}_\infty, C_\infty)$ is a steady state if and only if

$$\mathbf{m}_\infty = \mathbf{m}_\beta(\mathbf{m}_\infty, C_\infty) \qquad \text{and} \qquad C_\infty = \lambda^{-1} C_\beta(\mathbf{m}_\infty, C_\infty),$$

i.e., if and only if $(\mathbf{m}_\infty(\beta), C_\infty(\beta))$ is a fixed point of the map

$$\Phi_\beta : (\mathbf{m}, C) \mapsto \big(\mathbf{m}_\beta(\mathbf{m}, C), (1+\beta) C_\beta(\mathbf{m}, C)\big).$$

To prove the result, we show that $\Phi_\beta(S_\beta) \subset S_\beta$ for all β sufficiently large, where

$$S_\beta = \left\{ (\mathbf{m}, C) : |\mathbf{m} - \theta_*| \leqslant R\beta^{-1/2} \text{ and } U^{-1} \preccurlyeq C \preccurlyeq L^{-1} \right\}$$

and $R = 2k/\sqrt{\ell}$, with $k = k(\ell, u, d)$ the constant from Lemma 7. Because Φ_β is continuous, the result then follows from Brouwer's fixed point theorem. By Lemmas 5 and 6, it holds that $U^{-1} \preccurlyeq (1+\beta) C_\beta(\mathbf{m}, C) \preccurlyeq L^{-1}$ for any $(\mathbf{m}, C) \in S_\beta$, so we have to show only that there exist $\underline{\beta}$ such that

$$\forall \beta \geqslant \underline{\beta}, \quad \forall (\mathbf{m}, C) \in S_\beta, \qquad |\mathbf{m}_\beta(\mathbf{m}, C) - \theta_*| \leqslant R\beta^{-1/2}.$$

If $(\mathbf{m}, C) \in S_\beta$, then by Lemma 7 there exists $k = k(\ell, u, d)$ such that

$$\begin{aligned} \forall \beta > 0, \qquad |\mathbf{m}_\beta(\mathbf{m}, C) - \theta_*| &\leqslant \frac{R}{\beta} \sqrt{\frac{u}{\ell}} + k(\ell + \beta\ell)^{-1/2} \\ &\leqslant R\beta^{-1/2} \left(\sqrt{\frac{u}{\beta\ell}} + \frac{k}{R} \sqrt{\frac{1}{\ell}} \right) = R\beta^{-1/2} \left(\sqrt{\frac{u}{\beta\ell}} + \frac{1}{2} \right) \end{aligned}$$

from where the statement follows easily with $\underline{\beta} = \frac{4u}{\ell}$. ■

This result shows that the sampling scheme admits a steady state whose mean is close to the minimizer of f for large β, but it does not provide much information on the covariance of the Gaussian steady state. In the one-dimensional setting, we can show that the steady state is in fact unique and arbitrarily close to the Laplace approximation of the target distribution provided that β is sufficiently large. By the Laplace approximation $\hat{\rho}$ of the target distribution, we mean the Gaussian probability distribution $\mathrm{g}(\,\cdot\,;\theta_*, \mathrm{D}^2 f(\theta_*)^{-1})$, that is,

$$\hat{\rho}(\theta) := \frac{\mathrm{e}^{-\hat{f}(\theta)}}{\int_{\mathbf{R}^d} \mathrm{e}^{-\hat{f}(\theta)}\, \mathrm{d}\theta}, \qquad \hat{f}(\theta) := f(\theta_*) + \frac{1}{2}\big((\theta - \theta_*) \otimes (\theta - \theta_*)\big) : \mathrm{D}^2 f(\theta_*).$$

(Note that $\hat{\rho}$ coincides with the target distribution when f is quadratic.) To establish results in the one-dimensional setting, we make the following additional assumption on f.

Assumption 4. Let $d = 1$. The function f is smooth and, together with all its derivatives, it is bounded from above by the reciprocal of a Gaussian, in the sense that for all $i \in \{0, 1, \dots\}$ there exists $\lambda_i \in \mathbf{R}$ such that

$$\|\mathrm{e}^{-\lambda_i t^2} f^{(i)}(t)\|_\infty < \infty.$$

We let $C_* := 1/f''(\theta_*)$ and denote by $B_R(m_*, C_*)$ the closed ball of radius R around (m_*, C_*).

Theorem 3 (Convergence to the steady state). *Let $d = 1$ and $\lambda = (1+\beta)^{-1}$, and suppose Assumptions 1 and 4 hold. For any $R \in (0, C_*)$, there exists $\underline{\beta} = \underline{\beta}(f, R)$ and $k = k(f, R)$ such that the following statements hold for all $\beta \geqslant \underline{\beta}$:*

- ***Steady state.*** *There exists a pair $(m_\infty(\beta), C_\infty(\beta))$, unique in $B_R(\theta_*, C_*)$, such that the Gaussian density $\rho_\infty = \mathrm{g}(\,\cdot\,; m_\infty, C_\infty)$ satisfies (20), and this pair satisfies*

$$\left| \begin{pmatrix} m_\infty(\beta) \\ C_\infty(\beta) \end{pmatrix} - \begin{pmatrix} m_* \\ C_0 \end{pmatrix} \right| \leqslant \frac{k}{\beta}.$$

  By Lemma 1, the density ρ_∞ is a steady state of both the iterative scheme (16) with any $\alpha \in [0, 1)$ and the nonlinear Fokker–Planck equation (18), corresponding to $\alpha = 1$.
- ***Discrete time*** $\alpha \in [0, 1)$. *If Assumption 3 holds and the moments of the initial (Gaussian) law satisfy $(m_0, C_0) \in B_R(\theta_*, C_*)$, then the solution to the iterative scheme Equation (16) converges geometrically to the steady-state ρ_∞ provided that $\alpha + (1 - \alpha^2)\frac{k}{\beta} < 1$. More precisely,*

$$\forall n \in \mathbf{N}, \qquad \left| \begin{pmatrix} m_n \\ C_n \end{pmatrix} - \begin{pmatrix} m_\infty(\beta) \\ C_\infty(\beta) \end{pmatrix} \right| \leqslant \left(\alpha + (1 - \alpha^2)\frac{k}{\beta} \right)^n \left| \begin{pmatrix} m_0 \\ C_0 \end{pmatrix} - \begin{pmatrix} m_\infty(\beta) \\ C_\infty(\beta) \end{pmatrix} \right|.$$

- ***Continuous time*** $\alpha = 1$. *If Assumption 3 holds and the moments of the initial (Gaussian) law satisfy $(m_0, C_0) \in B_R(\theta_*, C_*)$, then the solution to the mean-field Fokker Planck equation (18)* *converges exponentially to the steady-state* ρ_∞ *provided that* $1 - \frac{2k}{\beta} > 0$. *More precisely,*

$$\forall t \geq 0, \qquad \left|\begin{pmatrix} m(t) \\ C(t) \end{pmatrix} - \begin{pmatrix} m_\infty(\beta) \\ C_\infty(\beta) \end{pmatrix}\right| \leqslant \exp\!\left(-\left(1-\frac{2k}{\beta}\right)t\right)\left|\begin{pmatrix} m_0 \\ C_0 \end{pmatrix} - \begin{pmatrix} m_\infty(\beta) \\ C_\infty(\beta) \end{pmatrix}\right|.$$

There is no conceptual obstruction to generalizing this result to the multidimensional setting, but the associated calculations involving the Laplace's method, on which the proof of Theorrem 3 relies, are significantly more technical than in the one-dimensional setting, so we focus here on the one-dimensional case only.

4 | NUMERICAL EXPERIMENTS

In this section, we present numerical experiments illustrating our method. The performance of CBS in optimization mode is studied in Section 4.1. We then illustrate the efficacy of the method for sampling in Section 4.2, where a simple inverse problem with low-dimensional parameter and data is considered, and in Section 4.3, where a more realistic and challenging example is examined. Video animations associated with the numerical experiments presented in this section are freely available online.[66]

4.1 | General-purpose optimization

In this subsection, we study the efficacy of our method for solving optimization problems that do not necessarily originate from a Bayesian context. We also show empirically how the convergence of the algorithm can be improved by adapting the parameter β appropriately during the simulation. Throughout the subsection, we consider the same nonconvex test functions as those taken in[9]: the translated Ackley function, defined for $x \in \mathbf{R}^d$ by

$$f_A(x) = -20\exp\left(-\frac{1}{5}\sqrt{\frac{1}{d}\sum_{i=1}^{d}|x_i - b|^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos\big(2\pi(x_i - b)\big)\right) + \mathrm{e} + 20, \tag{37}$$

and the Rastrigin function, defined by

$$f_R(x) = \sum_{i=1}^{d}\Big((x_i - b)^2 - 10\cos\big(2\pi(x_i - b)\big) + 10\Big). \tag{38}$$

Both functions are minimized at $x_* = (b, \ldots, b)$, where $b \in \mathbf{R}$ is a translation parameter. They are depicted in Figure 1.

In all simulations presented below, the initial particle ensemble members are drawn independently from $\mathsf{N}(0, 3I_d)$, and the simulation is stopped when $\|C(\rho_n^J)\|_{\mathrm{F}} < 10^{-12}$ for the first time; here ρ_n^J denotes the empirical measure associated with the ensemble at iteration n.

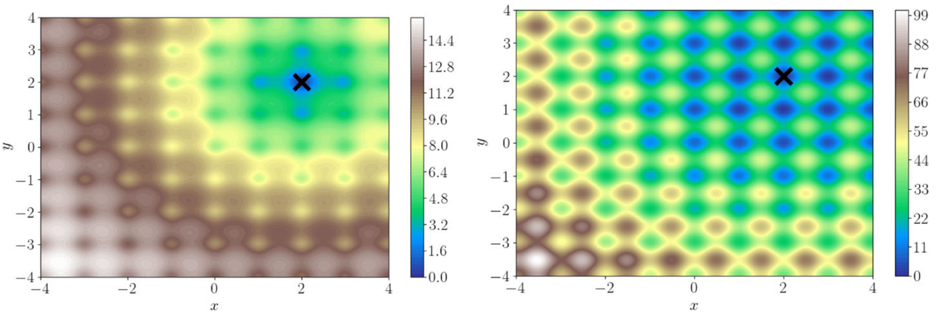

FIGURE 1 Ackley (left) and Rastrigin (right) functions for $d = 2$ and $b = 2$; see (37) and (38)

4.1.1 | Dynamic adaptation of β

In this paragraph, we show numerically that adapting β dynamically during a simulation can be advantageous for convergence. We consider the following simple adaptation scheme with parameter $\eta \in \left(\frac{1}{J}, 1\right)$: denoting by $\{\theta_n^{(j)}\}_{j=1}^J$ the ensemble at step n, the parameter β employed for the next iteration is obtained as the positive solution to the following equation:

$$J_{\text{eff}}(\beta) := \frac{\left(\sum_{j=1}^J \omega_j\right)^2}{\sum_{j=1}^J |\omega_j|^2} = \eta J, \qquad \omega_j := \mathrm{e}^{-\beta f(\theta_n^{(j)})}. \tag{39}$$

Employing the notation $f_j = f(\theta_n^{(j)})$, we calculate

$$J'_{\text{eff}}(\beta) = -2\beta \frac{\left(\sum_{j=1}^J \omega_j\right)\left(\sum_{j=1}^J f_j \omega_j\right) - \left(\sum_{j=1}^J f_j |\omega_j|^2\right)}{\left(\sum_{j=1}^J |\omega_j|^2\right)^2} \leqslant 0,$$

so J_{eff} is a continuous, nonincreasing function with $J_{\text{eff}}(0) = J$ and $\lim_{\beta\to\infty} J_{\text{eff}}(\beta) = 1$. Consequently, Equation (39) admits a unique solution in $(0, \infty)$. The left-hand side of (39) is known in statistics as an *effective sample size*, which motivates the notation J_{eff}. When this approach is employed, the parameter β is generally small in the early stage of the simulation as long as the initial ensemble has large enough spread, and it increases progressively as the simulation advances and the ensemble spread decreases. In other words, this cooling schedule for β ensures that roughly always the same proportion η of particles contribute to the weighted sums in the scheme. This adaptation approach is useful for a two primary reasons:

- On the one hand, provided that η and J are sufficiently large, adapting β according to (39) ensures that situations where the ensemble quickly collapses to a very narrow distribution do not arise. An early collapse of the ensemble is not desirable as the scheme may then get stuck in local minima of the objective function f, or in the case when the collapse is not complete, the convergence is slowed down considerably. This issue is especially critical when the scheme (30) is employed with $\alpha = 0$: in this case, if β is not sufficiently small at the beginning of the

TABLE 3 Performance of the CBS in optimization mode for the Ackley function in spatial dimension $d = 2$, without and with adaptive β. The three data presented in each cell are respectively the success rate of the method, the average number of iteration until the stopping criterion is met, and the average (over the successful runs) error at the final iteration, computed as the infinity norm between the minimizer and the ensemble mean. Our definition of the success rate is very similar to that used in Ref. 9: a run is considered successful if the ensemble mean is within 0.25, in infinity norm, of the minimizer at the final iteration

Adapt?	α	$J = 50$	$J = 100$	$J = 200$
No	0	100% | 511 | 8.73×10^{-3}	100% | 966 | 4.34×10^{-3}	100% | 1767 | 2.5×10^{-3}
No	0.5	100% | 611 | 1.22×10^{-2}	100% | 1191 | 6.87×10^{-3}	100% | 2141 | 3.38×10^{-3}
No	0.9	100% | 2028 | 1.6×10^{-2}	100% | 3693 | 8.31×10^{-3}	100% | 7259 | 5.22×10^{-3}
Yes	0	100% | 31 | 1.86×10^{-7}	100% | 31 | 1.09×10^{-7}	100% | 31 | 8.44×10^{-8}
Yes	0.5	100% | 49 | 2.86×10^{-7}	100% | 48 | 2.0×10^{-7}	100% | 48 | 1.43×10^{-7}
Yes	0.9	100% | 251 | 2.27×10^{-6}	100% | 242 | 4.36×10^{-7}	100% | 238 | 2.87×10^{-7}

simulation, it is often the case that the weighted covariance of the initial ensemble is very close to zero, in which case the ensemble collapses nearly to a point in a single step.

- On the other hand, increasing β in the later stage of the simulation significantly accelerates convergence to the minimizer. Indeed, when a fixed value of β is employed, the weights $\{\omega_j\}_{j=1}^J$ all converge to the same value as the simulation progresses and the ensemble collapses, and so the influence of the objective function on the dynamics diminishes. By increasing β dynamically, we strengthen the bias of the dynamics towards areas of small f, thereby accelerating convergence.

In the remainder of this section, we consider for simplicity only the choice $\eta = \frac{1}{2}$. A more detailed analysis of the efficiency of this approach, through both theoretical and numerical means, is left for future work. More generally, an interesting open question is whether it is possible to determine an optimal cooling schedule for β taking the above considerations into account. We illustrate in Table 3 the performance of CBS in optimization mode, with both fixed and adaptive β, for finding the minimizer of the Ackley function with $b = 0$ in dimension 2. The data presented in each cell are calculated from 100 independent runs of the method. For all the values of J and α considered, using the adaptive strategy based on (39) provides a significant advantage, in terms of both the number of iterations required for convergence and the accuracy of the approximate minimizer.

4.1.2 | Low-dimensional optimization problem: $d = 2$

The performance of CBS in optimization mode is illustrated in Tables 4 and 5, for the Ackley and Rastrigin functions respectively, in spatial dimension $d = 2$. We make a few observations:

- *Influence of* α: The simulations corresponding to $\alpha = 0$ consistently require fewer iterations to converge than those corresponding to $\alpha = \frac{1}{2}$, and they have a better success rate for the Rastrigin function.
- *Influence of* J: For the Rastrigin function, a high number of particles, i.e. a large value of J, correlates with a better success rate. With only 50 particles, the method often converges to the wrong

TABLE 4 Performance of the CBS in optimization mode for the Ackley function in spatial dimension $d = 2$. See the caption of Table 3 for a description of the data presented

b	α	$J = 50$	$J = 100$	$J = 200$
0	0	100% | 31 | 1.86×10^{-7}	100% | 31 | 1.09×10^{-7}	100% | 31 | 8.44×10^{-8}
0	0.5	100% | 49 | 2.86×10^{-7}	100% | 48 | 2.0×10^{-7}	100% | 48 | 1.43×10^{-7}
1	0	100% | 31 | 1.83×10^{-7}	100% | 31 | 1.16×10^{-7}	100% | 31 | 7.91×10^{-8}
1	0.5	100% | 49 | 3.23×10^{-7}	100% | 49 | 2.05×10^{-7}	100% | 49 | 1.47×10^{-7}
2	0	100% | 31 | 1.86×10^{-7}	100% | 32 | 1.1×10^{-7}	100% | 32 | 8.61×10^{-8}
2	0.5	100% | 51 | 3.03×10^{-7}	100% | 50 | 1.92×10^{-7}	100% | 50 | 1.38×10^{-7}

TABLE 5 Performance of the CBS in optimization mode for the Rastrigin function in spatial dimension $d = 2$. See the caption of Table 3 for a description of the data presented

b	α	$J = 50$	$J = 100$	$J = 200$
0	0	83% | 41 | 1.73×10^{-7}	99% | 45 | 1.19×10^{-7}	100% | 45 | 8.43×10^{-8}
0	0.5	77% | 74 | 3.39×10^{-4}	98% | 69 | 2.21×10^{-7}	100% | 66 | 1.56×10^{-7}
1	0	84% | 42 | 1.85×10^{-7}	99% | 44 | 1.03×10^{-7}	100% | 45 | 7.8×10^{-8}
1	0.5	72% | 68 | 6.03×10^{-7}	91% | 68 | 2.23×10^{-7}	100% | 68 | 1.56×10^{-7}
2	0	79% | 42 | 1.84×10^{-7}	96% | 44 | 1.12×10^{-7}	100% | 45 | 7.78×10^{-8}
2	0.5	58% | 80 | 4.14×10^{-4}	74% | 75 | 3.52×10^{-5}	96% | 74 | 1.54×10^{-7}

local minimizer, but with 200 particles the ensemble almost always collapses at the global minimizer.

- *Influence of b*: For the Rastrigin function, a low value of b correlates with better performance. This behavior, which was observed also for CBO in Ref. 9, is not surprising because, when $b = 0$, the minimizer is centered with respect to the initial ensemble.

We also note that, like CBO,[9] our method performs markedly better for the Ackley function than for the Rastrigin function. Snapshots of the particles are presented in Figure 2 for the parameters $\alpha = 0$ and $J = 100$.

4.1.3 | Higher-dimensional optimization problem: $d = 10$

In this paragraph, we repeat the numerical experiments of the previous section in higher dimension $d = 10$. We employ an adaptive β in all the simulations, as this approach was shown in the previous subsection to perform much better. The associated results are presented in Tables 6 and 7, which show that the method performs better for small α and large J for this case as well. Overall, the method seems to require a larger ensemble size than CBO to guarantee a similar success rate. A fair comparison of the computational expenses required by both methods is difficult, however, because the number of time steps employed in CBO is not documented in Ref. 9.

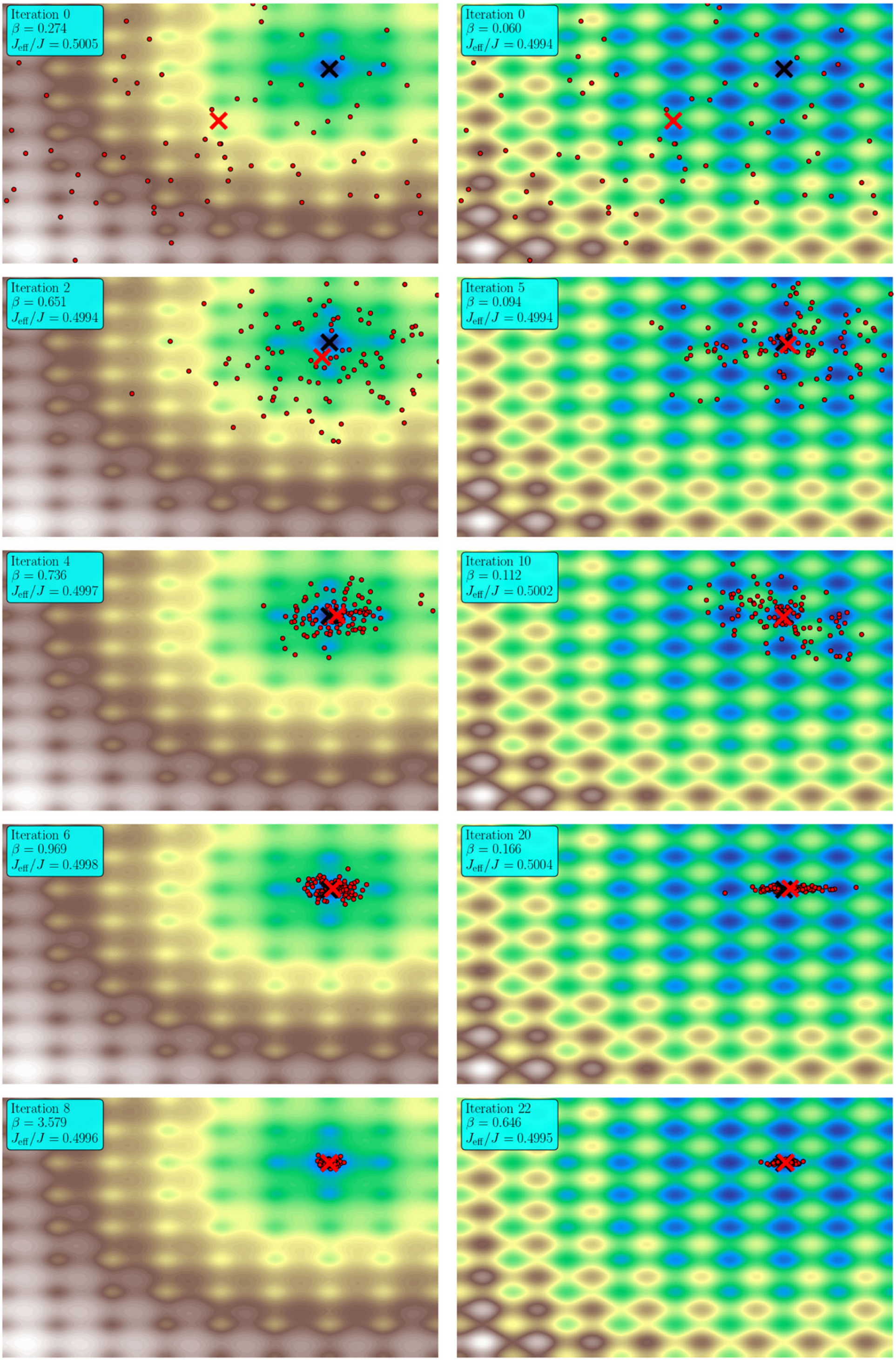

FIGURE 2 Illustration of the convergence of CBS in optimization mode for the Ackley (left) and Rastrigin (right) functions in dimension 2, for the parameters $J = 100$, $\alpha = 0$, and with adaptive β. The black cross denotes the unique global minimizer, and the red cross shows the ensemble mean

TABLE 6 Performance of the CBS in optimization mode for the Ackley function in dimension 10. See the caption of Table 3 for a description of the data presented

b	α	$J = 100$	$J = 500$	$J = 1000$
0	0	100% | 95 | 4.19×10^{-4}	100% | 77 | 9.81×10^{-8}	100% | 78 | 6.97×10^{-8}
0	0.5	100% | 248 | 1.27×10^{-2}	100% | 109 | 1.71×10^{-7}	100% | 110 | 1.13×10^{-7}
1	0	100% | 100 | 1.34×10^{-3}	100% | 78 | 1.04×10^{-7}	100% | 78 | 6.79×10^{-8}
1	0.5	98% | 278 | 3.27×10^{-2}	100% | 111 | 1.72×10^{-7}	100% | 111 | 1.13×10^{-7}
2	0	98% | 125 | 7.72×10^{-3}	100% | 78 | 9.71×10^{-8}	100% | 79 | 6.85×10^{-8}
2	0.5	65% | 306 | 6.53×10^{-2}	100% | 113 | 1.7×10^{-7}	100% | 113 | 1.13×10^{-7}

TABLE 7 Performance of the CBS in optimization mode for the Rastrigin function in dimension 10. See the caption of Table 3 for a description of the data presented

b	α	$J = 100$	$J = 500$	$J = 1000$
0	0	6% | 222 | 2.1×10^{-2}	95% | 107 | 9.69×10^{-8}	100% | 111 | 6.62×10^{-8}
0	0.5	10% | 331 | 6.68×10^{-2}	99% | 150 | 1.88×10^{-7}	100% | 155 | 1.14×10^{-7}
1	0	4% | 224 | 4.61×10^{-2}	94% | 108 | 9.66×10^{-8}	100% | 111 | 6.97×10^{-8}
1	0.5	0% | 334 | –	74% | 165 | 5.75×10^{-7}	99% | 162 | 1.18×10^{-7}
2	0	0% | 224 | –	74% | 113 | 9.82×10^{-8}	99% | 114 | 7.07×10^{-8}
2	0.5	0% | 333 | –	19% | 190 | 1.17×10^{-4}	69% | 189 | 1.24×10^{-7}

4.2 | Sampling: Low-dimensional parameter space

We first consider an inverse problem with low-dimensional parameter space that was first presented in Ref. 67 and later employed as a test problem in Refs. 34,68. In this problem, the forward model maps the unknown $(u_1, u_2) \in \mathbf{R}^2$ to the observation $(p(x_1), p(x_2)) \in \mathbf{R}^2$, where $x_1 = 0.25$ and $x_2 = 0.75$ and where $p(x)$ denotes the solution to the boundary value problem

$$-\,\mathrm{e}^{u_1}\,p'' = 1, \qquad x \in [0,1], \tag{40}$$

with boundary conditions $p(0) = 0$ and $p(1) = u_2$. This problem admits the following explicit solution[68]:

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

We employ the same parameters as in Ref. 34: the prior distribution is $\mathsf{N}(0, \sigma^2 I_2)$ with $\sigma = 10$, and the noise distribution is $\mathsf{N}(0, \gamma^2 I_2)$ with $\gamma = 0.1$. The observed data are $y = (27.5, 79.7)$.

We now investigate the efficiency of (30) for sampling from the posterior distribution. To this end, we use the parameters $\alpha = \beta = \frac{1}{2}$ and $J = 1000$ particles. The ensemble after 100 iterations is depicted in Figure 3, together with the true posterior. It appears from the figure that the Gaussian approximation of the posterior provided by scheme (30) is close to the true posterior, and indeed we can verify that the mean and covariance of the true and approximate posterior distributions,

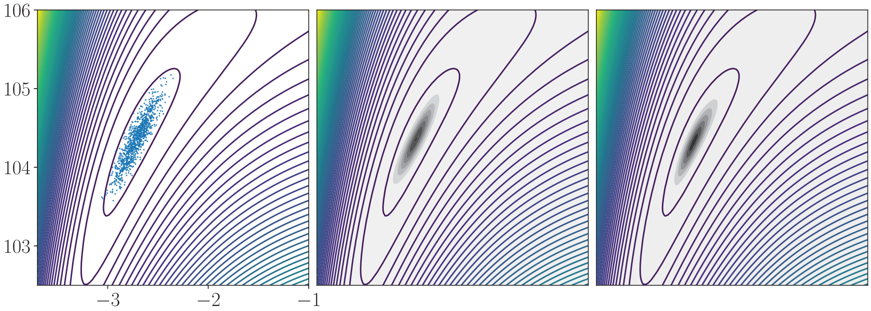

FIGURE 3 Left: Particles at iteration $n = 100$ for fixed $\alpha = \beta = \frac{1}{2}$. Middle: Gaussian density with the same mean and covariance as the empirical distribution associated with these particles. Right: True Bayesian posterior

which are given, respectively, by

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

and

$$\widetilde{m}_p \begin{pmatrix} -2.712\ldots \\ 104.356\ldots \end{pmatrix} \quad \widetilde{C}_p = \begin{pmatrix} 0.0135\ldots & 0.0302\ldots \\ 0.0302\ldots & 0.0829\ldots \end{pmatrix},$$

are fairly close.

4.3 | Sampling: Higher dimensional parameter space

In this section, we consider the more challenging inverse problem of finding the permeability field of a porous medium from noisy pressure measurements in a Darcy flow; for other methods applied to this problem, see Refs. 2,34,60. Assuming Dirichlet boundary conditions and scalar permeability for simplicity, we consider the forward model mapping the logarithm of the permeability, denoted by $a(x)$, to the solution of the PDE

$$-\nabla \cdot \left(\mathrm{e}^{a(x)}\, \nabla p(x)\right) = f(x), \qquad x \in D, \tag{41a}$$

$$p(x) = 0, x \in \partial D. \tag{41b}$$

Here $D = [0,1]^2$ is the domain and $f(x) = 50$ represents a source of fluid. We assume that noisy pointwise measurements of $p(x)$ are taken at a finite number equispaced points in D, given by

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

and that these measurements are perturbed by Gaussian noise with distribution $\mathcal{N}(0, \gamma^2 I_K)$, where $\gamma = 0.01$ and $K = (M-1)^2$. For the prior distribution, we employ a Gaussian measure on $L^2(D)$ with mean zero and precision (inverse covariance) operator given by

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

equipped with Neumann boundary conditions on the space of mean-zero functions. Here r and τ are parameters controlling the smoothness and characteristic inverse length scale of samples drawn from the prior, respectively. The eigenfunctions and eigenvalues of the covariance operator are

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

By the Karhunen–Loève (KL) expansion,[69] it holds for any $a \sim \mathcal{N}(0, \mathcal{C})$ that

$$a(x) = \sum_{\ell \in \mathbf{N}^2} (a, \psi_\ell)\psi_\ell(x) =: \sum_{\ell \in \mathbf{N}^2} \sqrt{\lambda_\ell}\, \theta_\ell\, \psi_\ell(x), \tag{42}$$

for independent coefficients $\theta_\ell \sim \mathcal{N}(0,1)$, and where $(\bullet, \bullet)$ denotes the L^2-inner product.

To approach the problem numerically, we take as object of inference a finite number of terms $\{\theta_\ell\}_{|\ell|_\infty \leqslant N}$ in the KL expansion of the log-permeability, which may be ordered as a linear vector given an ordering of $\{0, \dots, N\}^2$. The associated prior distribution is given by the finite-dimensional Gaussian $\mathcal{N}(0, I_d)$, where $d = (N+1)^2$. At the numerical level, the forward model is evaluated as follows: for a given vector of coefficients $\theta \in \mathbf{R}^d$, a log-permeability field is calculated by summation as $a(\bullet; \theta) := \sum_{|\ell|_\infty \leqslant N} \sqrt{\lambda_\ell}\, \theta_\ell\, \psi_\ell(\bullet)$, and the corresponding solution to (41) is approximated with a finite element method (FEM). Linear shape functions over a regular mesh with 100 subdivisions per direction are employed for the finite element solution.

For the numerical experiments presented below, a true value $\theta^\dagger \in \mathbf{R}^d$ for the vector of coefficients is drawn from $\mathcal{N}(0, I_d)$ and employed to construct the true permeability field which, in turn, is used with the FEM described above to generate the data. In particular, we employ only $(N+1)^2$ terms in the KL expansion of the true permeability. We note that, with this approach, the resulting random field should be viewed only as an approximate sample from $\mathcal{N}(0, \mathcal{C})$. Our aim is to study the performance of CBS, not the effect of FEM discretization and truncation of the KL series on the solution of the inverse problem.

The ensemble obtained after 100 iterations of CBS with adaptive β, with $\alpha = 0$ and with $J = 512$ is depicted in Figure 4, along with the marginals of the Gaussian distribution with the same first and second moments as the empirical measure associated with the ensemble. The particles forming the initial ensembles were drawn independently from $\mathsf{N}(0, 9I_d)$. To validate our results, we use as point of reference the solution provided by the ensemble Kalman sampling method,[34] combined with the adaptive time-stepping scheme from Ref. 57. It appears from the simulations that the agreement between the posterior distribution obtained by CBS and that obtained by ensemble Kalman sampling is very good, and both approximate posteriors are in good agreement with the true solution.

Using the final ensemble as initial condition for (30) in optimization mode, and running 50 more iterations of the algorithm, one obtains an approximation of the MAP estimator, whose associated permeability field is illustrated in Figure 5. Here we use as point of comparison the solution provided by the EKI approach.[31] We present below the values of the first nine

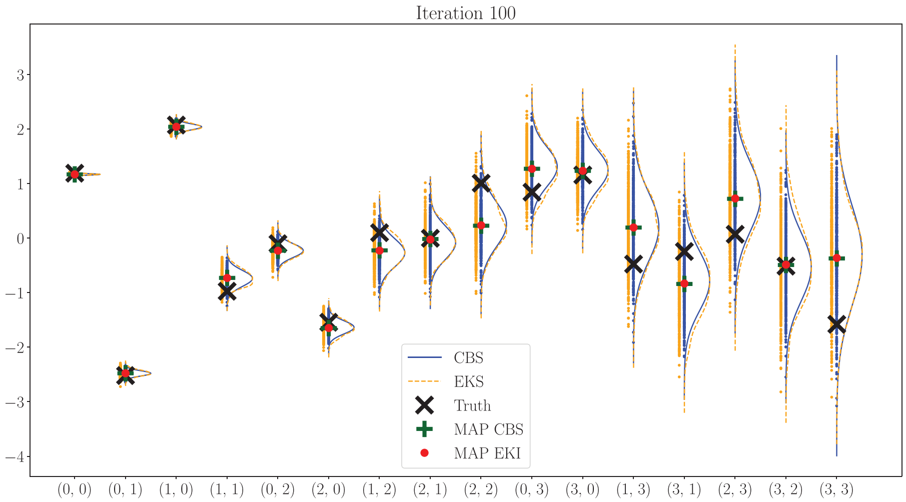

FIGURE 4 Approximate posterior samples produced by (30) with $\alpha = 0$ and adaptive β. Here, the labels on the x-axis denote the multi-indices associated with the KL coefficients of the permeability. The (nonnormalized) solid curves represent the marginals of the Gaussian distribution whose mean and covariance are calculated from the samples produced by CBS. The (nonnormalized) dashed curves are the marginal distributions obtained by kernel density estimation using Gaussian kernels from the samples produced by ensemble Kalman sampling.[34] The black crosses denote the true values of the KL coefficients, i.e., the values employed to generate the data

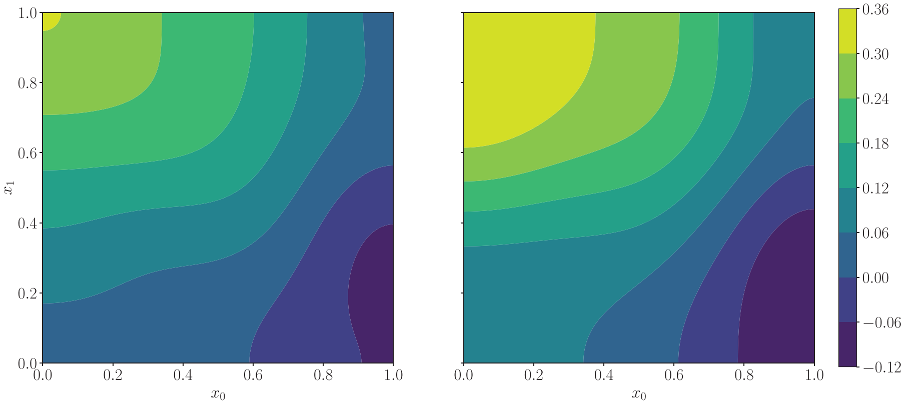

FIGURE 5 Logarithms of true (left) and approximate permeability profiles (right). The approximate permeability profile was constructed from the approximation of the MAP estimator provided by (30) with $\alpha = 0$, adaptive β and $\lambda = 1$ (optimization mode), with $J = 512$ particles

Karhunen–Loève coefficients of (i) the true permeability, (ii) the MAP estimator obtained by CBS, and (iii) the MAP estimator obtained by EKI:

$$(u^{\dagger})^{\mathsf{T}} = \begin{pmatrix} 1.19 & -2.52 & 2.07 & -0.97 & -0.10 & -1.54 & 0.10 & -0.00 & 1.01 & \dots \end{pmatrix},$$

$$(u_{\mathrm{MAP}}^{\mathrm{CBS}})^{\mathsf{T}} = \begin{pmatrix} 1.17 & -2.48 & 2.04 & -0.73 & -0.23 & -1.65 & -0.22 & -0.02 & 0.23 & \dots \end{pmatrix},$$

$$(u_{\mathrm{MAP}}^{\mathrm{EKI}})^{\mathsf{T}} = \begin{pmatrix} 1.17 & -2.48 & 2.04 & -0.73 & -0.23 & -1.65 & -0.23 & -0.02 & 0.24 & \dots \end{pmatrix}.$$

(All the numbers displayed here were rounded to two decimals.) The agreement between the MAP estimators as approximated by EKI and by our method is very good, and both vectors are close to the KL series of the logarithm of the true permeability.

4.4 | Discussion

We draw the following conclusions from the numerical experiments presented in this section.

- It is crucial to dynamically adapt the parameter β during a simulation for our method to be competitive, both for optimization and sampling tasks. We obtained very good numerical results with the adaptation scheme based on the effective sample size in (39).
- For optimization tasks, our method generally requires more particles than CBO[9] to consistently find the global minimizer when the number of local minima is large. Relatedly, for a given number of particles, the probability of converging to (a small neighborhood) of the correct minimizer appears to be better for CBO.
- For sampling tasks, our numerical experiments suggest that the CBS method is competitive with the ensemble Kalman sampling scheme.[34] The number of iterations required by both methods to reach equilibrium is of the same order of magnitude, and the quality of the posterior approximation appears similar in the test cases we considered.

In future work, we will aim to give our proposed β-adaptation scheme a theoretical footing, and to investigate other adaptation strategies. It will also be worthwhile to more precisely compare our method with discretizations of CBO and EKS in terms of computational cost, especially for PDE-based inverse problems, where evaluations of the forward model are typically the predominant computational cost. Finally, it would be interesting, both for optimization and sampling tasks, to investigate whether ideas from Refs. 36,46,70 could be leveraged to improve the performance of our method when the number of particles is of the same order of magnitude as the dimension of the parameter space.

5 | PROOF OF THE MAIN RESULTS

Throughout this section, for a given $\mathbf{m} \in \mathbf{R}^d$ and $C \in \mathbf{R}^{d\times d}$, we will use the notation

$$\rho_\beta(\theta; \mathbf{m}, C) = \frac{1}{Z_\beta} e^{-V_\beta(\theta)}, \qquad V_\beta(\theta; \mathbf{m}, C) := \frac{1}{2}|\theta - \mathbf{m}|_C^2 + \beta f(\theta), \tag{43}$$

where $Z_\beta = Z_\beta(\mathbf{m}, C)$ is the normalization constant. When the parameters $\mathbf{m}$, C are clear from the context, we will often write just $\rho_\beta(\theta)$ and $V_\beta(\theta)$ for conciseness.

5.1 | Proof of the convergence estimates in the Gaussian setting

Proof of Proposition 1. Consider first the sampling case $\lambda = (1+\beta)^{-1}$. Using the same notation as in the proof of Lemma 3, we have

$$\widetilde{C}_n^{-1} - \beta^{-1} I_d = \lambda^n \big(\widetilde{C}_0^{-1} - \beta^{-1} I_d\big). \tag{44}$$

Rearranging the equation, we obtain

$$\widetilde{C}_n - \beta I_d = (\widetilde{C}_n \widetilde{C}_0^{-1})\lambda^n \big(\widetilde{C}_0 - \beta I_d\big).$$

Because $\widetilde{C}_n$ commutes with $\widetilde{C}_0^{-1}$ from (44), the matrix $\widetilde{C}_n\widetilde{C}_0^{-1}$ is symmetric and positive definite. By (44), the eigenvalues $\{\ell_i\}$ of $\widetilde{C}_n\widetilde{C}_0^{-1}$ are of the form

$$\ell_i = \frac{1}{\beta^{-1}m_i + \lambda^n(1-\beta^{-1}m_i)} \leqslant \max\left\{\frac{\beta}{m_i}, 1\right\} \leqslant \max\{1, k_0\},$$

where $\{m_i\}$ denote the eigenvalues of $\widetilde{C}_0$. Hence,

$$\begin{aligned}\|C_n - A\|_A = \beta^{-1}\|\tilde{C}_n - \beta I_d\| &= \lambda^n\beta^{-1}\left\|\left(\tilde{C}_n\tilde{C}_0^{-1}\right)\left(\tilde{C}_0 - \beta I_d\right)\right\| \\ &\leqslant \lambda^n\|\tilde{C}_n\tilde{C}_0^{-1}\|\beta^{-1}\|\tilde{C}_0 - \beta I_d\| \leqslant \lambda^n \max\left(1, k_0\right)\|C_0 - A\|_A\,.\end{aligned}$$

This shows the convergence result of the covariance, and the convergence result for the mean follows similarly using Lemma 3:

$$\begin{aligned}|\mathbf{m}_n - \mathbf{a}|_A = |\tilde{\mathbf{m}}_n| &= \lambda^n|\tilde{C}_n\tilde{C}_0^{-1}\tilde{\mathbf{m}}_0| \leqslant \lambda^n\|\tilde{C}_n\tilde{C}_0^{-1}\||\tilde{\mathbf{m}}_0| \\ &\leqslant \lambda^n \max(1, k_0)\,|\tilde{\mathbf{m}}_0| = \lambda^n \max(1, k_0)\,|\mathbf{m}_0 - \mathbf{a}|_A\,.\end{aligned}$$

In the optimization case $\lambda = 1$, we have using the definition of k_0 that

$$\widetilde{C}_n^{-1} = \widetilde{C}_0^{-1} + nI_d \succcurlyeq \left(1 + \frac{\beta n}{k_0}\right)\widetilde{C}_0^{-1} \qquad \Rightarrow \qquad \widetilde{C}_n \preccurlyeq \left(\frac{k_0}{k_0 + \beta n}\right)\widetilde{C}_0.$$

This shows the convergence result for the covariance, which directly implies the convergence estimate for the mean. ■

Proof of Proposition 2. Notice that the right-hand side of (25b) commutes with $\widetilde{C}_n$, so there exists an orthogonal matrix Q such that $\widehat{C}_n := Q^\top\widetilde{C}_n Q$ is diagonal for all $n \in \mathbf{N}$. Introducing $\widehat{m}_n = Q^\top\widetilde{m}_n$, we can check that $\widehat{m}_n$ and $\widehat{C}_n$ solve again (25). Therefore, for all $i \in \{1, \dots, d\}$, it holds that $(u_{i,n}, v_{i,n}) := ((\widehat{m}_n)_i, (\widehat{C}_n)_{ii})$ solves the discrete-time equation (26) with initial conditions which depend on i. The convergence of the solution for the two-dimensional difference equation (26) is then given by Lemma A.1. Note that $v_{i,0} \geqslant \beta/k_0$ for all $i \in \{1, \dots, d\}$, because by definition $k_0 = \beta\|\widetilde{C}_0^{-1}\| = \beta\|\widehat{C}_0^{-1}\|$. In the sampling case, we have

$$\begin{aligned}|\mathbf{m}_n - \mathbf{a}|_A = |\widehat{m}_n| &\leqslant \max(1, k_0)^{\frac{1}{1+\alpha}}((1-\alpha)\lambda + \alpha)^n|\widehat{m}_0| \\ &= \max(1, k_0)^{\frac{1}{1+\alpha}}((1-\alpha)\lambda + \alpha)^n|\mathbf{m}_0 - \mathbf{a}|_A\,.\end{aligned}$$

On the other hand, it holds for any $1 \leqslant i \leqslant d$ that

$$|(\widehat{C}_n)_{ii} - \beta| \leqslant \max(1, k_0)\big((1-\alpha^2)\lambda + \alpha^2\big)^n|(\widehat{C}_0)_{ii} - \beta|.$$

From this, we deduce

$$\|\widehat{C}_n - \beta I_d\| \leqslant \max(1, k_0)\big((1-\alpha^2)\lambda + \alpha^2\big)^n \|\widehat{C}_0 - \beta I_d\|.$$

Because $\|QMQ^{\mathsf{T}}\| = \|M\|$ for any symmetric matrix M and orthogonal matrix Q, we deduce

$$\|\tilde{C}_n - \beta I_d\| \leqslant \max(1, k_0)\big((1-\alpha^2)\lambda + \alpha^2\big)^n \|\tilde{C}_0 - \beta I_d\|.$$

The statement then follows because $\|C_n - C_\infty\|_A = \beta^{-1}\|\tilde{C}_n - \beta I_d\|$ by definition of $\|\bullet\|_A$. An analogous argument, using the estimates (A.3a) and (A.3b) in Lemma A.1 and noting that the function $s \mapsto (s+1)/(s+1+(1-\alpha^2)n)$ is strictly decreasing for $s \geqslant 0$, yields the bounds for the optimization case $\lambda = 1$. ■

Proof of Proposition 3. Letting $\widetilde{\mathbf{m}}(t) = A^{-1/2}(\mathbf{m}(t) - \mathbf{a})$ and $\widetilde{C}(t) = \beta A^{-1/2}C(t)A^{-1/2}$, we can verify that $\widetilde{\mathbf{m}}$ and $\widetilde{C}$ solve

$$\begin{aligned}
\dot{\widetilde{\mathbf{m}}} &= -\widetilde{C}\big(I_d + \widetilde{C}\big)^{-1}\widetilde{\mathbf{m}}, \\
\dot{\widetilde{C}} &= -2\widetilde{C}\left(I_d + \widetilde{C}\right)^{-1}\left(\widetilde{C} - \left(\frac{1-\lambda}{\lambda}\right)I_d\right).
\end{aligned}$$

It is then straightforward to show the result by employing the same reasoning as in the discrete-time case and using Lemma A.2, which characterizes the convergence to equilibrium for the following ordinary differential equation (ODE) system with u, v scalar functions :

$$\dot{u} = -\Big(\frac{v}{1+v}\Big)u, \qquad \dot{v} = -2\Big(\frac{v}{1+v}\Big)(v - v_\infty), \qquad v_\infty = \frac{1-\lambda}{\lambda}. \tag{46}$$

We leave the details to the reader. ■

5.2 | Proof of the preliminary bounds

Proof of Lemma 5. Recall notation (43), and let $\widetilde{\theta}$ denote the unique global minimizer of $V_\beta(\theta)$. The function g defined by

$$\mathrm{g}(\theta) = f(\theta) - \left(f(\widetilde{\theta}) + \nabla f(\widetilde{\theta})^{\mathsf{T}}(\theta - \widetilde{\theta}) + \frac{1}{2}|\theta - \widetilde{\theta}|^2_{L^{-1}}\right)$$

is such that $\mathrm{g}(\widetilde{\theta}) = \nabla \mathrm{g}(\widetilde{\theta}) = 0$ and $\mathrm{D}^2\,\mathrm{g}(\theta) \succcurlyeq 0$ for all $\theta \in \mathbf{R}^d$, by the convexity assumption on the function f. We denote

$$\widetilde{V}_\beta(\theta) := \frac{1}{2}|\theta|^2 + \beta\widetilde{\mathrm{g}}(\theta), \qquad \widetilde{\mathrm{g}}(\theta) := \mathrm{g}\Big(\widetilde{\theta} + \big(C^{-1} + \beta L\big)^{-1/2}\theta\Big),$$

and define $\widetilde{\rho}_\beta(\theta) = \frac{1}{\widetilde{Z}_\beta}\,\mathrm{e}^{-\widetilde{V}_\beta(\theta)}$ where $\widetilde{Z}_\beta$ is the normalization constant. By a change of variables, it holds

$$\mathcal{C}_\beta(\mathbf{m}, C) = \mathcal{C}(\rho_\beta) = \left(C^{-1} + \beta L\right)^{-1/2} \mathcal{C}(\widetilde{\rho}_\beta) \left(C^{-1} + \beta L\right)^{-1/2},$$

It remains to show $\mathcal{C}(\widetilde{\rho}_\beta) \preccurlyeq I_d$ or, equivalently, that for every unit vector $\mathbf{a} \in \mathbf{R}^d$ it holds

$$\mathbf{a}^\mathsf{T} \mathcal{C}(\tilde{\rho}_\beta)\mathbf{a} = \int_{\mathbf{R}^d} \left| \mathbf{a}^\mathsf{T}\theta - \int_{\mathbf{R}^d} (\mathbf{a}^\mathsf{T}\theta)\,\tilde{\rho}_\beta(\theta)\mathrm{d}\theta \right|^2 \tilde{\rho}_\beta(\theta)\,\mathrm{d}\theta \leqslant 1, \tag{47}$$

Clearly $\widetilde{g}(0) = \nabla\widetilde{g}(0) = 0$ and $\mathrm{D}^2\,\widetilde{g} \succcurlyeq 0$, so $\mathrm{D}^2\,\widetilde{V}_\beta \succcurlyeq I_d$. Therefore, by the Bakry–Emery criterion Ref. [65, Theorem 2.10], the probability distribution $\mathrm{d}\mu(\theta) := \widetilde{\rho}_\beta(\theta)\mathrm{d}\theta$ satisfies a logarithmic Sobolev inequality, and thus also a Poincaré inequality by Ref. [65, Proposition 2.12], with the factor on the right equal to 1. That is, it holds

$$\forall u \in H^1(\mu), \qquad \int_{\mathbf{R}^d} \left| u - \int_{\mathbf{R}^d} u\,\mathrm{d}\mu \right|^2 \mathrm{d}\mu \leqslant \int_{\mathbf{R}^d} |\nabla u|^2 \mathrm{d}\mu.$$

Applying this inequality with $u(\theta) = \mathbf{a}^\mathsf{T}\theta$ gives (47). ■

Proof of Lemma 6. Let $\widetilde{\theta}$ denote again the unique global minimizer of $V_\beta(\theta)$, where V_β is given in (43). The function g defined by

$$g(\theta) = f(\theta) - \left(f(\widetilde{\theta}) + \nabla f(\widetilde{\theta})^\mathsf{T}(\theta - \widetilde{\theta}) + \frac{1}{2}|\theta - \widetilde{\theta}|^2_{U^{-1}} \right)$$

is such that $g(\widetilde{\theta}) = \nabla g(\widetilde{\theta}) = 0$ and $\mathrm{D}^2\,g(\theta) \preccurlyeq 0$ for all $\theta \in \mathbf{R}^d$, by assumption 2. By a change of variables, it holds

$$\mathcal{C}_\beta(\mathbf{m}, C) = \mathcal{C}(\rho_\beta) = \left(C^{-1} + \beta U\right)^{-1/2} \mathcal{C}(\widetilde{\rho}_\beta) \left(C^{-1} + \beta U\right)^{-1/2},$$

where $\widetilde{\rho}_\beta(\theta) = \frac{1}{\widetilde{Z}_\beta}\,\mathrm{e}^{-\widetilde{V}_\beta(\theta)}$, with $\widetilde{Z}_\beta$ the normalization constant and

$$\widetilde{V}_\beta(\theta) := \frac{1}{2}|\theta|^2 + \beta\widetilde{g}(\theta), \qquad \widetilde{g}(\theta) := g\left(\widetilde{\theta} + \left(C^{-1} + \beta U\right)^{-1/2}\theta\right).$$

It remains to show that $\mathcal{C}(\widetilde{\rho}_\beta) \succcurlyeq I_d$. To this end, let $\bar{\theta} = \mathcal{M}(\widetilde{\rho}_\beta)$ for brevity and, for a given unit vector $\mathbf{a} \in \mathbf{R}^d$, let $\partial_{\mathbf{a}} = \mathbf{a}^\mathsf{T}\nabla$ and so $\partial_{\mathbf{a}}\widetilde{\rho}_\beta = -(\partial_{\mathbf{a}}\widetilde{V}_\beta)\widetilde{\rho}_\beta$. By the Cauchy–Schwarz inequality,

$$\int_{\mathbf{R}^d} \partial_{\mathbf{a}}\widetilde{V}_\beta(\theta)\,\mathbf{a}^\mathsf{T}(\theta - \bar{\theta})\,\widetilde{\rho}_\beta(\theta)\,\mathrm{d}\theta \leqslant \sqrt{\int_{\mathbf{R}^d} |\partial_{\mathbf{a}}\widetilde{V}_\beta(\theta)|^2 \widetilde{\rho}_\beta(\theta)\,\mathrm{d}\theta}\,\sqrt{\int_{\mathbf{R}^d} |\mathbf{a}^\mathsf{T}(\theta - \bar{\theta})|^2 \widetilde{\rho}_\beta(\theta)\,\mathrm{d}\theta}.$$

After rearranging and using integration by parts, this gives

$$\mathbf{a}^\top C(\widetilde{\rho}_\beta)\mathbf{a} \geqslant \frac{\left(\int_{\mathbf{R}^d} \partial_{\mathbf{a}} \widetilde{V}_\beta(\theta)\, \mathbf{a}^\top(\theta - \bar{\theta})\, \widetilde{\rho}_\beta(\theta)\, \mathrm{d}\theta\right)^2}{\int_{\mathbf{R}^d} |\partial_{\mathbf{a}} \widetilde{V}_\beta(\theta)|^2 \widetilde{\rho}_\beta(\theta)\, \mathrm{d}\theta}$$

$$= \frac{\left(\int_{\mathbf{R}^d} \mathbf{a}^\top(\theta - \bar{\theta})\, \partial_{\mathbf{a}} \widetilde{\rho}_\beta(\theta)\, \mathrm{d}\theta\right)^2}{-\int_{\mathbf{R}^d} \partial_{\mathbf{a}} \widetilde{V}_\beta(\theta) \partial_{\mathbf{a}} \widetilde{\rho}_\beta(\theta)\, \mathrm{d}\theta} = \frac{1}{\int_{\mathbf{R}^d} \partial_{\mathbf{a}}^2 \widetilde{V}_\beta(\theta) \widetilde{\rho}_\beta(\theta)\, \mathrm{d}\theta},$$

where we denote $\partial_{\mathbf{a}}^2 h(\theta) = \mathbf{a}^T D^2 h(\theta) \mathbf{a}$. Because $\mathrm{D}^2\, \widetilde{V}_\beta \preccurlyeq I_d$ because $\mathrm{D}^2\, \widetilde{g} \preccurlyeq 0$, it follows immediately that $C(\widetilde{\rho}_\beta) \succcurlyeq I_d$. ■

Proof of Lemma 7. Let $\widetilde{\theta}$ denote again the unique global minimizer of $V_\beta(\theta)$ given by (43). We first show a bound on $\widetilde{\theta} - \theta_*$. By the assumptions on f, it holds

$$V_\beta(\theta) \geqslant \frac{1}{2}|\theta - m|_C^2 + \frac{\ell\beta}{2}|\theta - \theta_*|^2 + \beta f(\theta_*) \geqslant \frac{\ell\beta}{2}|\theta - \theta_*|^2 + \beta f(\theta_*).$$

Likewise, it holds $V_\beta(\theta_*) \leqslant \frac{1}{2}\|C^{-1}\||\theta_* - m|^2 + \beta f(\theta_*)$, so we obtain

$$V_\beta(\theta) - V_\beta(\theta_*) \geq \frac{\ell\beta}{2}|\theta - \theta_*|^2 - \frac{1}{2}\|C^{-1}\||\theta_* - m|^2.$$

In particular, for any θ such that

$$|\theta - \theta_*| > \left(\frac{\|C^{-1}\|}{\ell\beta}\right)^{1/2} |\theta_* - m| =: R,$$

it holds $V_\beta(\theta) - V_\beta(\theta_*) > 0$, implying that $|\widetilde{\theta} - \theta_*| \leqslant R$. Now,

$$|\mathbf{m}_\beta(\mathbf{m}, C) - \tilde{\theta}| = |\mathcal{M}(\rho_\beta) - \tilde{\theta}| = \left|\int_{\mathbf{R}} (\theta - \tilde{\theta}) \rho_\beta(\theta)\, \mathrm{d}\theta\right|$$

$$\leqslant \sqrt{\int_{\mathbf{R}^d} |\theta - \tilde{\theta}|^2\, \rho_\beta(\theta)\, \mathrm{d}\theta} = \sqrt{\frac{\int_{\mathbf{R}^d} |\theta - \tilde{\theta}|^2\, \mathrm{e}^{-V_\beta(\theta)}\, \mathrm{d}\theta}{\int_{\mathbf{R}^d} \mathrm{e}^{-V_\beta(\theta)}\, \mathrm{d}\theta}}. \qquad (48)$$

Because $V_\beta(\theta)$ is minimized at $\theta = \widetilde{\theta}$, it holds

$$V_\beta(\tilde{\theta}) + \frac{1}{2}|\theta - \tilde{\theta}|^2_{(C^{-1} + \beta L)^{-1}} \leqslant V_\beta(\theta) \leqslant V_\beta(\tilde{\theta}) + \frac{1}{2}|\theta - \tilde{\theta}|^2_{(C^{-1} + \beta U)^{-1}}.$$

Using these inequalities, we can obtain an upper bound for the numerator in (48) and a lower bound for the denominator in (48), respectively:

$$\int_{\mathbf{R}^d} |\theta - \tilde{\theta}|^2\, \mathrm{e}^{-V_\beta(\theta)}\, \mathrm{d}\theta \leqslant \mathrm{e}^{-V_\beta(\tilde{\theta})}\, \mathrm{tr}\left(\left(C^{-1} + \beta L\right)^{-1}\right) \det\left(C^{-1} + \beta L\right)^{-1/2} (2\pi)^{d/2}$$

$$\int_{\mathbf{R}^d} \mathrm{e}^{-V_\beta(\theta)}\,\mathrm{d}\theta \geq \mathrm{e}^{-V_\beta(\tilde{\theta})} \det\left(C^{-1}+\beta U\right)^{-1/2}(2\pi)^{d/2}.$$

Combining these inequalities, writing the determinant as a product of eigenvalues, and using the inequality $\frac{1+x}{1+y} \leqslant \frac{x}{y}$ for all $0 < y \leqslant x$, we deduce

$$\begin{aligned}\left|\mathbf{m}_\beta(\mathbf{m}, C) - \tilde{\theta}\right| &\leqslant \sqrt{\operatorname{tr}\left(\left(C^{-1}+\beta L\right)^{-1}\right)} \frac{\det\left(C^{-1}+\beta U\right)^{1/4}}{\det\left(C^{-1}+\beta L\right)^{1/4}} \\ &\leqslant \sqrt{d\|(C^{-1}+\beta L)^{-1}\|} \frac{\det\left(C^{-1}+\beta u I_d\right)^{1/4}}{\det\left(C^{-1}+\beta \ell I_d\right)^{1/4}} \leqslant \sqrt{d\|(C^{-1}+\beta L)^{-1}\|}\left(\frac{u}{\ell}\right)^{d/4}.\end{aligned}$$

The statement then follows from the triangle inequality,

$$|\mathbf{m}_\beta(\mathbf{m}, C) - \theta_*| \leqslant |\theta_* - \tilde{\theta}| + |\mathbf{m}_\beta(\mathbf{m}, C) - \tilde{\theta}|,$$

and from the fact that $\|(C^{-1}+\beta L)^{-1}\| \leqslant \|(C^{-1}+\beta\ell I_d)^{-1}\| \leqslant (\|C\|^{-1}+\beta\ell)^{-1}$. ■

5.3 | Proof of Proposition 4 and Theorem 1

Proof of Proposition 4. Let $x = \alpha^2$ for simplicity. It holds by (10b) and Lemma 5 that

$$C_{n+1} \preccurlyeq x C_n + (1-x)(C_n^{-1}+\beta L)^{-1}.$$

Therefore, introducing $\bar{C}_n = \beta L^{1/2} C_n L^{1/2}$, it holds

$$\bar{C}_{n+1} \preccurlyeq x \bar{C}_n + (1-x)(\bar{C}_n^{-1}+I_d)^{-1}.$$

Let $\bar{D}_n$ denote the solution to the discrete-time equation

$$\bar{D}_{n+1} = x\bar{D}_n + (1-x)(\bar{D}_n^{-1}+I_d)^{-1}, \qquad \bar{D}_0 = \bar{C}_0.$$

It is clear that $\bar{C}_n \preccurlyeq \bar{D}_n$ for all $n \geqslant 0$. Indeed, this is true for $n = 0$, and if $\bar{C}_n \preccurlyeq \bar{D}_n$ then

$$\begin{aligned}\bar{D}_{n+1} - \bar{C}_{n+1} &\succcurlyeq x(\bar{D}_n - \bar{C}_n) + (1-x)\left((\bar{D}_n^{-1}+I_d)^{-1} - (\bar{C}_n^{-1}+I_d)^{-1}\right) \\ &\succcurlyeq (1-x)\left((\bar{D}_n^{-1}+I_d)^{-1} - (\bar{C}_n^{-1}+I_d)^{-1}\right).\end{aligned}$$

By Ref. [71, Proposition V.1.6], the function $\mathbf{R} \ni s \mapsto -1/s$ is operator monotone on $(0, \infty)$, meaning that if two symmetric positive definite matrices M_1 and M_2 are such that $M_1 \succcurlyeq M_2$, then it holds that $M_1^{-1} \preccurlyeq M_2^{-1}$. Therefore

$$\bar{C}_n \preccurlyeq \bar{D}_n \;\Rightarrow\; \bar{C}_n^{-1} \succcurlyeq \bar{D}_n^{-1} \;\Rightarrow\; \bar{C}_n^{-1} + I_d \succcurlyeq \bar{D}_n^{-1} + I_d \;\Rightarrow\; (\bar{C}_n^{-1}+I_d)^{-1} \preccurlyeq (\bar{D}_n^{-1}+I_d)^{-1},$$

which shows that $\bar{D}_{n+1} - \bar{C}_{n+1} \succcurlyeq 0$. Now note that $\bar{D}_n$ satisfies the same equation as $\widetilde{C}_n$ in (25b), so we deduce by a reasoning similar to the proof of Proposition 2 that $\bar{D}_n$ satisfies

$$\bar{D}_n \preccurlyeq \left(\frac{\|\bar{C}_0^{-1}\| + 1}{\|\bar{C}_0^{-1}\| + 1 + (1-x)n} \right) \bar{C}_0,$$

which implies the statement for the discrete-time case $\alpha \in (0,1)$. If $\alpha = 0$, then it follows from Proposition 1 that

$$\bar{D}_n \preccurlyeq \left(\frac{\|\bar{C}_0^{-1}\|}{\|\bar{C}_0^{-1}\| + n} \right) \bar{C}_0.$$

Similarly in the continuous-time case, let $\bar{C}(t) = \beta L^{1/2} C(t) L^{1/2}$ and let $\bar{D}(t)$ denote the solution to the equation

$$\frac{\mathrm{d}}{\mathrm{d}t} \bar{D}(t) = -2\bar{D}(t) + 2\big(\bar{D}(t)^{-1} + I_d\big)^{-1}, \qquad \bar{D}(0) = \bar{C}(0).$$

We have by (22b) and Lemma 5 that

$$\frac{\mathrm{d}}{\mathrm{d}t} \bar{C}(t) \preccurlyeq -2\bar{C}(t) + 2\big(\bar{C}(t)^{-1} + I_d\big)^{-1}.$$

Using the same reasoning as in the discrete-time case, we derive that

$$\frac{\mathrm{d}}{\mathrm{d}t} \big(\bar{D}(t) - \bar{C}(t)\big) \succcurlyeq -2\big(\bar{D}(t) - \bar{C}(t)\big) \qquad \Leftrightarrow \qquad \frac{\mathrm{d}}{\mathrm{d}t} \left(\mathrm{e}^{2t} \left(\bar{D}(t) - \bar{C}(t)\right)\right) \succcurlyeq 0,$$

and so $\bar{C}(t) \preccurlyeq \bar{D}(t)$ for all $t \geqslant 0$. Employing a reasoning similar to that in Proposition 3, we obtain the statement. ■

We show a similar result establishing a lower bound on C_n.

Lemma 8 (Lower bound on the covariance in optimization mode). *Let $\lambda = 1$, $\beta > 0$ and $\alpha \in [0,1)$, and assume that Assumption 2 holds. Then, for any solution $\{(m_n, C_n)\}_{n \in \mathbf{N}}$ to Equations (23a) and (23b) with $C_0 \in \mathcal{S}^d_{++}$, it holds that*

$$C_n \succcurlyeq \left(C_0^{-1} + n(1-\alpha^2)\beta U\right)^{-1}. \tag{49}$$

Likewise, for any solution $\big\{(m(t), C(t))\big\}_{t \in \mathbf{R}_{\geqslant 0}}$ to Equations (24a) and (24b) with $C(0) \in \mathcal{S}^d_{++}$, the following inequality holds:

$$C(t) \succcurlyeq \left(C(0)^{-1} + 2t\beta U\right)^{-1}. \tag{50}$$

Proof. Let us now use the notation $\widehat{C}_n = \beta U^{1/2} C_n U^{1/2}$. It holds by Lemma 6 that

$$\widehat{C}_{n+1} \succcurlyeq x\widehat{C}_n + (1-x)(\widehat{C}_n^{-1} + I_d)^{-1}.$$

Defining $\widehat{P}_n = \widehat{C}_n^{-1}$ for $n \in \{0, 1, \dots\}$ we have

$$\widehat{P}_{n+1} \preccurlyeq (xI_d + \widehat{P}_n)^{-1}(I_d + \widehat{P}_n)\widehat{P}_n = \widehat{P}_n + (1-x)(I_d + x\widehat{P}_n^{-1})^{-1} \preccurlyeq \widehat{P}_n + (1-x)I_d,$$

so we deduce (49). For the continuous-time case, we employ the notation $\widehat{C}(t) = \beta U^{1/2}C(t)U^{1/2}$ and $\widehat{P}(t) = \widehat{C}(t)^{-1}$. By (24b) and Lemma 6, we have that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\widehat{C}(t) &\succcurlyeq -2\widehat{C}(t) + 2(\widehat{C}(t)^{-1} + I_d)^{-1} \\ &= -2(\widehat{C}(t)^{-1} + I_d)^{-1}\left[(\widehat{C}(t)^{-1} + I_d)\widehat{C}(t) - I_d\right] = -2\widehat{C}(t)\left(\widehat{C}(t) + I_d\right)^{-1}\widehat{C}(t).\end{aligned}$$

Hence,

$$\frac{\mathrm{d}}{\mathrm{d}t}\widehat{P}(t) = -\widehat{C}(t)^{-1}\frac{\mathrm{d}}{\mathrm{d}t}\widehat{C}(t)\widehat{C}(t)^{-1} \preccurlyeq 2\left(I_d + \widehat{C}(t)\right)^{-1} \preccurlyeq 2I_d,$$

leading to the statement. ■

Remark 8. A simple corollary of Proposition 4 and Lemma 8 is that the condition number

$$\operatorname{cond}(C_n) = \|C_n\|\|C_n^{-1}\|$$

of C_n remains bounded as $n \to \infty$, and similarly in continuous time.

To prove Theorem 1, we first show the following auxiliary result.

Lemma 9. *Let $\beta > 0$ and suppose f satisfies Assumptions 1 and 2. Then there exists a constant $K = K(\beta, d, \ell, u) > 0$ such that the following inequality holds*

$$|\mathbf{m}_\beta(\mathbf{m}, C) - \mathbf{m} + \beta C\nabla f(\mathbf{m})| \leqslant \mathrm{e}^{\beta f(\mathbf{m})}\frac{K\beta|C\nabla f(\mathbf{m})|\,\|C\| + K\|C\|^{3/2}}{1 - K\mathrm{e}^{\beta f(\mathbf{m})}\|C\|},$$

for all $(\mathbf{m}, C) \in \mathbf{R}^d \times \mathcal{S}^d_{++}$ such that the denominator is positive.

Proof. By Taylor's theorem, there exists for all $(\theta, \mathbf{m}) \in \mathbf{R}^d \times \mathbf{R}^d$ a point $\xi = \xi(\theta, \mathbf{m}) \in \mathbf{R}^d$ on the straight segment between θ and $\mathbf{m}$ such that

$$\begin{aligned}\mathrm{e}^{-\beta f(\theta)} &= \mathrm{e}^{-\beta f(\mathbf{m})} - \mathrm{e}^{-\beta f(\mathbf{m})}\,\beta\nabla f(\mathbf{m})\cdot(\theta - \mathbf{m}) \\ &\quad + \frac{1}{2}\,\mathrm{e}^{-\beta f(\xi)}\left(\beta^2(\nabla f(\xi)\otimes\nabla f(\xi)) - \beta\,\mathrm{D}^2 f(\xi)\right) : \left((\theta - \mathbf{m})\otimes(\theta - \mathbf{m})\right) \\ &=: \mathrm{e}^{-\beta f(\mathbf{m})} - \mathrm{e}^{-\beta f(\mathbf{m})}\,\beta\nabla f(\mathbf{m})\cdot(\theta - \mathbf{m}) + R(\theta; \mathbf{m}).\end{aligned}$$

By Assumptions 1 and 2, it is clear that

$$\frac{1}{2}\sup_{\xi\in\mathbf{R}^d}\left(\mathrm{e}^{-\beta f(\xi)}\left(\beta^2|\nabla f(\xi)|^2+\beta\|\mathrm{D}^2 f(\xi)\|_{\mathrm{F}}\right)\right)<\infty,$$

where $\|\bullet\|_{\mathrm{F}}$ denotes the Frobenius norm. Consequently, there exists a constant M such that

$$\forall(\theta,\mathbf{m})\in\mathbf{R}^d\times\mathbf{R}^d,\qquad |R(\theta;\mathbf{m})|\leqslant M|\theta-\mathbf{m}|^2. \tag{51}$$

We therefore deduce

$$\int_{\mathbf{R}^d} g(\theta;\mathbf{m},C)\ \mathrm{e}^{-\beta f(\theta)}\ \mathrm{d}\theta=\mathrm{e}^{-\beta f(\mathbf{m})}+R_0(\mathbf{m},C), \tag{52a}$$

$$\int_{\mathbf{R}^d}(\theta-\mathbf{m})\,g(\theta;\mathbf{m},C)\ \mathrm{e}^{-\beta f(\theta)}\ \mathrm{d}\theta=-\,\mathrm{e}^{-\beta f(\mathbf{m})}\,\beta C\nabla f(\mathbf{m})+R_1(\mathbf{m},C), \tag{52b}$$

with remainder terms satisfying the bounds

$$\forall(\mathbf{m},C)\in\mathbf{R}^d\times\mathcal{S}^d_{++},\qquad \begin{cases}|R_0(\mathbf{m},C)|\leqslant K\|C\|,\\ |R_1(\mathbf{m},C)|\leqslant K\|C\|^{3/2}.\end{cases} \tag{53}$$

The second bound holds because, by (51) and a change of variable, we have

$$\begin{aligned}|R_1(\mathbf{m},C)|&\leqslant M\int_{\mathbf{R}^d}|\theta-\mathbf{m}|^3\,g(\theta;\mathbf{m},C)\,\mathrm{d}\theta\\ &=M\int_{\mathbf{R}^d}|C^{1/2}u|^3\,g(u;\mathbf{0},I_d)\,\mathrm{d}u\leqslant M\|C^{3/2}\|\int_{\mathbf{R}^d}|u|^3 g(u;\mathbf{0},I_d)\,\mathrm{d}u.\end{aligned}$$

Using eqns. (52a, 52b), we obtain

$$\mathbf{m}_\beta(\mathbf{m},C)-\mathbf{m}=\frac{-\,\mathrm{e}^{-\beta f(\mathbf{m})}\,\beta C\nabla f(\mathbf{m})+R_1(\mathbf{m},C)}{\mathrm{e}^{-\beta f(\mathbf{m})}+R_0(\mathbf{m},C)}.$$

In view of (53), it therefore holds

$$\begin{aligned}|\mathbf{m}_\beta(\mathbf{m},C)-\mathbf{m}+\beta C\nabla f(\mathbf{m})|&=\left|\frac{\beta C\nabla f(\mathbf{m})R_0(\mathbf{m},C)+R_1(\mathbf{m},C)}{\mathrm{e}^{-\beta f(\mathbf{m})}+R_0(\mathbf{m},C)}\right|\\ &\leqslant \mathrm{e}^{\beta f(\mathbf{m})}\frac{K\beta|C\nabla f(\mathbf{m})|\,\|C\|+K\|C\|^{3/2}}{|1+\mathrm{e}^{\beta f(\mathbf{m})}R_0(\mathbf{m},C)|}.\end{aligned}$$

Using the bound on $R_0(\mathbf{m},C)$ given in (53), we obtain the statement. ■

Proof of Theorem 1. For a contradiction, assume $\mathbf{m}_n\to\hat{\theta}$ and $\hat{\theta}\neq\theta_*$, where θ_* denotes the global minimizer of f. Then, by the convexity assumption on f, it holds that $|\nabla f(\widehat{\theta})|>0$. By Proposition

4, it holds $C_n \to 0$, and by Remark 8, the condition number of C_n satisfies $\operatorname{cond}(C_n) \leqslant \varkappa$ for some $\varkappa > 0$ and all $n \in \{0, 1, \dots\}$. By continuity of ∇f at $\hat{\theta}$, we have that for any $\varepsilon > 0$, there is $\delta = \delta(\varepsilon) > 0$ such that

$$\forall \mathbf{m} \in B_\delta(\hat{\theta}), \qquad |\nabla f(\hat{\theta}) - \nabla f(\mathbf{m})| \leqslant \frac{\varepsilon}{\varkappa} |\nabla f(\hat{\theta})|. \tag{54}$$

Fix $0 < \varepsilon \ll 1$ and let $\delta = \delta(\varepsilon)$. From Lemma 9, there exists $K > 0$ such that the inequality

$$|\mathbf{m}_\beta(\mathbf{m}, C) - \mathbf{m} + \beta C \nabla f(\mathbf{m})| \leqslant \frac{K\beta \mathrm{e}^{\beta f(\mathbf{m})} |C \nabla f(\mathbf{m})| \, \|C\| + K\mathrm{e}^{\beta f(\mathbf{m})} \|C\|^{3/2}}{1 - K\mathrm{e}^{\beta f(\mathbf{m})} \|C\|} \tag{55}$$

is satisfied for all $(\mathbf{m}, C) \in (\mathbf{R}^d \times \mathcal{S}^d_{++})$ such that the denominator is positive. We claim that there exists $\widetilde{c} > 0$ such that the following inequalities are satisfied for all $\mathbf{m} \in B_\delta(\hat{\theta})$ and all matrices $0 < C \leqslant \widetilde{c} I_d$ such that $\operatorname{cond}(C) \leqslant \varkappa$:

$$\begin{cases} \left|1 - K\mathrm{e}^{\beta f(\mathbf{m})} \|C\|\right| \geqslant \dfrac{1}{2}, \\ K\beta \, \mathrm{e}^{\beta f(\mathbf{m})} \, \|C\| \leqslant \dfrac{\varepsilon}{4}, \\ K \, \mathrm{e}^{\beta f(\mathbf{m})} \, \|C\|^{3/2} \leqslant \dfrac{\varepsilon}{4} |C \nabla f(\mathbf{m})|. \end{cases} \tag{56}$$

Indeed, it suffices to choose

$$\widetilde{c} = \min\left(\frac{I}{2K}, \frac{\varepsilon I}{4K\beta}, \left(\frac{\varepsilon I}{4K\varkappa} \inf_{\mathbf{m} \in B_\delta(\hat{\theta})} |\nabla f(\mathbf{m})|\right)^2\right), \qquad \text{where } I = \inf_{\mathbf{m} \in B_\delta(\hat{\theta})} \mathrm{e}^{-\beta f(\mathbf{m})}.$$

Here the arguments of the minimum guarantee that each of the three inequalities in (56) are satisfied, respectively. We note that $\inf_{\mathbf{m} \in B_\delta(\hat{\theta})} |\nabla f(\mathbf{m})| > 0$ by (54) and the fact that $\varepsilon/\varkappa < 1$. To justify that the third inequality in (56) is indeed satisfied for this choice of $\widetilde{c}$, notice that

$$|C \nabla f(\mathbf{m})| \geq \lambda_{\min}(C) |\nabla f(\mathbf{m})| \geq \operatorname{cond}(C)^{-1} |\nabla f(\mathbf{m})| \, \|C\|.$$

Substituting the three inequalities in (56) into the estimate (55) from Lemma 9, we obtain that, for all $\mathbf{m} \in B_\delta(\hat{\theta})$ and all $0 < C \leqslant \widetilde{c} I_d$ such that $\operatorname{cond}(C) \leqslant \varkappa$, it holds

$$|\mathbf{m}_\beta(\mathbf{m}, C) - \mathbf{m} + \beta C \nabla f(\mathbf{m})| \leqslant \varepsilon |C \nabla f(\mathbf{m})|. \tag{57}$$

Now because $(\mathbf{m}_n, C_n) \to (\hat{\theta}, 0)$ as $n \to \infty$ by assumption, there exists N sufficiently large such that $\mathbf{m}_n \in B_\delta(\hat{\theta})$ and $0 < C_n \leqslant \widetilde{c} I_d$ and $\operatorname{cond}(C_n) \leqslant \varkappa$ for all $n \geqslant N$. By (10), we have that for any $n \geqslant N$ it holds

$$\mathbf{m}_{n+1} - \mathbf{m}_n = (1 - \alpha)\big(\mathbf{m}_\beta(\mathbf{m}_n, C_n) - \mathbf{m}_n\big) = -(1 - \alpha)\big(\beta C_n \nabla f(\mathbf{m}_n) + \mathbf{r}(\mathbf{m}_n, C_n)\big),$$

where $\mathbf{r}(\mathbf{m}_n, C_n)$ is the remainder term, bounded by (57). Taking the inner product of both sides with $\nabla f(\hat{\theta})$ and using (57), we deduce

$$-(\mathbf{m}_{n+1} - \mathbf{m}_n)^{\mathrm{T}} \nabla f(\widehat{\theta}) \geq \beta(1-\alpha)\left(\nabla f(\widehat{\theta})^{\mathrm{T}} C_n \nabla f(\mathbf{m}_n)\right) - \varepsilon(1-\alpha)\|C_n\| |\nabla f(\mathbf{m}_n)| |\nabla f(\widehat{\theta})|.$$

For any $(\mathbf{x}, \mathbf{y}) \in \mathbf{R}^d \times \mathbf{R}^d$ with $|\mathbf{x} - \mathbf{y}| \leqslant \zeta |\mathbf{x}|$, it holds

$$\begin{aligned}\mathbf{x}^{\mathrm{T}} C_n \mathbf{y} &= \mathbf{x}^{\mathrm{T}} C_n \mathbf{x} - \mathbf{x}^{\mathrm{T}} C_n (\mathbf{x} - \mathbf{y}) \\ &\geqslant \mathbf{x}^{\mathrm{T}} C_n \mathbf{x} - \sqrt{\mathbf{x}^{\mathrm{T}} C_n \mathbf{x}^{\mathrm{T}}} \sqrt{(\mathbf{x} - \mathbf{y})^{\mathrm{T}} C_n (\mathbf{x} - \mathbf{y})} \geqslant \lambda_{\min}(C_n) |\mathbf{x}|^2 (1 - \operatorname{cond}(C_n)\zeta).\end{aligned}$$

Together with (54), this implies

$$\begin{aligned}\forall n \geq N, \qquad -(\mathbf{m}_{n+1} - \mathbf{m}_n)^{\mathrm{T}} \nabla f(\widehat{\theta}) \geq& \beta(1-\alpha)(1-\varepsilon)\lambda_{\min}(C_n)|\nabla f(\widehat{\theta})|^2 \\ &- \varepsilon\left(1 + \frac{\varepsilon}{\kappa}\right)(1-\alpha)\lambda_{\max}(C_n)|\nabla f(\widehat{\theta})|^2.\end{aligned}$$

By repeating this reasoning with a smaller ε if necessary, we can ensure

$$\forall n \geq N, \qquad -(\mathbf{m}_{n+1} - \mathbf{m}_n)^{\mathrm{T}} \nabla f(\widehat{\theta}) \geq K \lambda_{\min}(C_n) |\nabla f(\widehat{\theta})|^2, \tag{58}$$

with a constant K independent of n. Because $\lambda_{\min}(C_n) \geqslant \frac{\lambda}{n}$ by (49), for some other constant λ independent of n, we conclude that for any $n \geqslant N$, it holds

$$-(\mathbf{m}_{n+1} - \mathbf{m}_N)^{\mathrm{T}} \nabla f(\hat{\theta}) \geqslant \left(\sum_{s=N}^{n} \frac{1}{s}\right) K\lambda |\nabla f(\hat{\theta})|^2 \xrightarrow[n\to\infty]{} \infty,$$

which is a contradiction because we assumed $\mathbf{m}_n \to \hat{\theta}$. A similar reasoning applies in the continuous-time setting. ■

5.4 | Proof of Propositions 5 and 6

For simplicity, we introduce the “dimensionless” notation $\widetilde{m} = \sqrt{\ell\beta}(m - \theta_*)$ and $\widetilde{C} = \ell\beta C$. We also introduce

$$\begin{aligned}\widetilde{m}_\beta(\widetilde{m}, \widetilde{C}) &= \sqrt{\ell\beta}\left(m_\beta\left(\theta_* + \frac{\widetilde{m}}{\sqrt{\ell\beta}}, \frac{\widetilde{C}}{\ell\beta}\right) - \theta_*\right) = \sqrt{\ell\beta}\big(m_\beta(m, C) - \theta_*\big), \\ \widetilde{C}_\beta(\widetilde{m}, \widetilde{C}) &= \ell\beta\, C_\beta\left(\theta_* + \frac{\widetilde{m}}{\sqrt{\ell\beta}}, \frac{\widetilde{C}}{\ell\beta}\right) = \ell\beta\, C_\beta(m, C).\end{aligned}$$

We begin by obtaining auxiliary results.

Lemma 10 (Bound on the weighted mean). *Let $d = 1$ and $\beta > 0$. If Assumption 1 is satisfied, then it holds*

$$\forall(\widetilde{m}, \widetilde{C}) \in \mathbf{R} \times \mathbf{R}_{>0}, \qquad |\widetilde{m}_\beta(\widetilde{m}, \widetilde{C})| \leqslant \frac{|\widetilde{m}|}{1+\widetilde{C}} \left(1 + 2 \frac{\phi\left(\frac{|\widetilde{m}|}{\sqrt{\widetilde{C}(1+\widetilde{C})}} \right)}{\frac{|\widetilde{m}|}{\sqrt{\widetilde{C}(1+\widetilde{C})}}} \right), \tag{59}$$

with ϕ the probability density function of the standard normal distribution, i.e., $\phi = \mathrm{g}(\bullet; 0, 1)$.

Proof. Let $\rho_+(\theta) := \frac{1}{Z_+} 1_{[\theta_*, \infty)}(\theta) \rho_\beta(\theta)$ and $\rho_-(\theta) := \frac{1}{Z_-} 1_{(-\infty, \theta_*)}(\theta) \rho_\beta(\theta)$, where ρ_β is defined as in (43) and Z_+, Z_- are the normalization constants. It is clear that

$$\mathcal{M}(\rho_-) \leqslant \mathcal{M}(\rho_\beta) \leqslant \mathcal{M}(\rho_+) \quad \text{and} \quad \mathcal{M}(\rho_-) \leqslant \theta_* \leqslant \mathcal{M}(\rho_+).$$

For example, we have

$$\mathcal{M}(\rho_+) - \theta_* = \frac{\int_{\theta_*}^{\infty} (\theta - \theta_*) \rho_\beta(\theta)}{\int_{\theta_*}^{\infty} \rho_\beta(\theta)} \geqslant \frac{\int_{\theta_*}^{\infty} (\theta - \theta_*) \rho_\beta(\theta)}{\int_{-\infty}^{\infty} \rho_\beta(\theta)} \geqslant \frac{\int_{-\infty}^{\infty} (\theta - \theta_*) \rho_\beta(\theta)}{\int_{-\infty}^{\infty} \rho_\beta(\theta)} = \mathcal{M}(\rho_\beta) - \theta_*.$$

Now notice that, because $f(\theta) = f(\theta_*) + \frac{\ell}{2}|\theta - \theta_*|^2 + \mathrm{g}(\theta)$ for a function g that is nondecreasing on $[\theta_*, \infty)$ and such that $\mathrm{g}(\theta_*) = \mathrm{g}'(\theta_*) = 0$, it holds by Lemma A.3 that

$$\begin{aligned} \mathcal{M}(\rho_+) - \theta_* &= \frac{\int_{\theta_*}^{\infty} (\theta - \theta_*) \exp\left(-\frac{(\theta - m)^2}{2C} - \beta f(\theta) \right) \mathrm{d}\theta}{\int_{\theta_*}^{\infty} \exp\left(-\frac{(\theta - m)^2}{2C} - \beta f(\theta) \right) \mathrm{d}\theta} \\ &\leqslant \frac{\int_{\theta_*}^{\infty} (\theta - \theta_*) \exp\left(-\frac{(\theta - m)^2}{2C} - \frac{\beta\ell}{2} |\theta - \theta_*|^2 \right) \mathrm{d}\theta}{\int_{\theta_*}^{\infty} \exp\left(-\frac{(\theta - m)^2}{2C} - \frac{\beta\ell}{2} |\theta - \theta_*|^2 \right) \mathrm{d}\theta}. \end{aligned}$$

Completing the square in the last expression, we obtain

$$\mathcal{M}(\rho_+) - \theta_* \leqslant \frac{\displaystyle\int_{\theta_*}^{\infty} (\theta - \theta_*) \exp\left(-\frac{1}{2} \left(\frac{1}{C} + \beta\ell \right) \left(\theta - \frac{\frac{m}{C} + \ell\beta\theta_*}{\frac{1}{C} + \ell\beta} \right)^2 \right) \mathrm{d}\theta}{\displaystyle\int_{\theta_*}^{\infty} \exp\left(-\frac{1}{2} \left(\frac{1}{C} + \beta\ell \right) \left(\theta - \frac{\frac{m}{C} + \ell\beta\theta_*}{\frac{1}{C} + \ell\beta} \right)^2 \right) \mathrm{d}\theta} =: D(m, C).$$

We claim that $D(\bullet, C)$, is a nondecreasing function for fixed C. Indeed, let us introduce the function $\mu : (m, C) \mapsto \frac{m/C + \ell\beta\theta_*}{1/C + \ell\beta}$. Because $\mu(m, C)$ is an increasing function of m for fixed C, it is sufficient to show that the function

$$\mu \mapsto \frac{\displaystyle\int_{\theta_*}^{\infty} (\theta - \theta_*) \exp\left(-\frac{1}{2}\left(\frac{1}{C} + \beta\ell\right)(\theta - \mu)^2\right) \mathrm{d}\theta}{\displaystyle\int_{\theta_*}^{\infty} \exp\left(-\frac{1}{2}\left(\frac{1}{C} + \beta\ell\right)(\theta - \mu)^2\right) \mathrm{d}\theta} \tag{60}$$

is nondecreasing for fixed C. To this end, assume that $\mu_1 \leqslant \mu_2$ and note that

$$\begin{aligned}\exp\left(-\frac{1}{2}\left(\frac{1}{C} + \beta\ell\right)|\theta - \mu_1|^2\right)& \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{1}{C} + \beta\ell\right)|\theta - \mu_2|^2\right) \exp\left(-\left(\frac{1}{C} + \beta\ell\right)(\mu_2 - \mu_1)\theta\right).\end{aligned}$$

Because the second factor is decreasing for $\theta \in [\theta_*, \infty)$, we deduce by Lemma A.3 that the function defined in (60) is nondecreasing, and therefore $D(\bullet, C)$ is also nondecreasing.

Using the standard formula for the mean of a truncated normal distribution, we deduce

$$D(m, C) = \mu(m, C) - \theta_* + \frac{\phi\left(\sqrt{\frac{1}{C} + \ell\beta}(\theta_* - \mu(m, C))\right)}{1 - \Phi\left(\sqrt{\frac{1}{C} + \ell\beta}(\theta_* - \mu(m, C))\right)} \frac{1}{\sqrt{\frac{1}{C} + \ell\beta}},$$

where Φ denotes the CDF of the standard normal distribution. Using the notation introduced at the beginning of this section and the fact that $\Phi(x) + \Phi(-x) = 1$, this rewrites

$$\sqrt{\ell\beta} D(m, C) = \frac{1}{1 + \widetilde{C}} \left(\widetilde{m} + \frac{\phi\left(\frac{\widetilde{m}}{\sqrt{\widetilde{C}(1+\widetilde{C})}}\right)}{\Phi\left(\frac{\widetilde{m}}{\sqrt{\widetilde{C}(1+\widetilde{C})}}\right)} \sqrt{\widetilde{C}(1 + \widetilde{C})} \right) =: \widetilde{D}(\widetilde{m}, \widetilde{C}).$$

Because $D(\bullet, C)$ is nondecreasing, we deduce that

$$\sqrt{\ell\beta}(\mathcal{M}(\rho_+) - \theta_*) \leqslant \widetilde{D}(|\widetilde{m}|, \widetilde{C}).$$

Employing the same reasoning for $\mathcal{M}(\rho_-)$, we obtain similarly

$$\sqrt{\ell\beta}(\mathcal{M}(\rho_-) - \theta_*) \geqslant -\widetilde{D}(|\widetilde{m}|, \widetilde{C}).$$

Using the fact that $\Phi(x) \geqslant \Phi(0) = 1/2$ for all $x \geqslant 0$, and

$$\sqrt{\ell\beta}(\mathcal{M}(\rho_-) - \theta_*) \leqslant \widetilde{m}_\beta(\widetilde{m}, \widetilde{C}) \leqslant \sqrt{\ell\beta}(\mathcal{M}(\rho_+) - \theta_*),$$

we obtain the statement. ■

To establish Proposition 6, we prove the following technical result.

Lemma 11 (Bound on the ratio of weighted moments). *Let $d = 1$ and $\beta > 0$. If Assumptions 1 and 2 are satisfied, then there exists for all $\varepsilon \in (0,1)$ a constant $\gamma = \gamma(\ell, u, \varepsilon) > 0$ such that*

$$\forall (\widetilde{m}, \widetilde{C}) \in \mathbf{R} \times \mathbf{R}_{>0}, \qquad \frac{|\widetilde{m}_\beta(\widetilde{m}, \widetilde{C})|}{\widetilde{C}_\beta(\widetilde{m}, \widetilde{C})^{\frac{1}{r}}} \leqslant \max\left(\gamma, \frac{|\widetilde{m}|}{\widetilde{C}^{\frac{1}{r}}}\right),$$

where $r = \max(\frac{u}{\ell}, (2 + \varepsilon))$.

Proof. Using Lemma 6 and Lemma 10, we deduce

$$\left|\frac{\widetilde{m}_\beta(\widetilde{m}, \widetilde{C})}{\widetilde{C}_\beta(\widetilde{m}, \widetilde{C})^{\frac{1}{r}}}\right| \leqslant \left|\frac{\widetilde{m}_\beta(\widetilde{m}, \widetilde{C})}{\left(\frac{\widetilde{C}}{1 + \frac{u}{\ell}\widetilde{C}}\right)^{\frac{1}{r}}}\right| \leqslant \frac{|\widetilde{m}|}{\widetilde{C}^{\frac{1}{r}}} \frac{\left(1 + r\widetilde{C}\right)^{\frac{1}{r}}}{1 + \widetilde{C}} \left(1 + 2 \frac{\phi\left(\frac{|\widetilde{m}|}{\sqrt{\widetilde{C}(1+\widetilde{C})}}\right)}{\frac{|\widetilde{m}|}{\sqrt{\widetilde{C}(1+\widetilde{C})}}}\right) =: B(\widetilde{m}, \widetilde{C}). \tag{61}$$

If $|\widetilde{m}| \geqslant \gamma \widetilde{C}^{1/r}$ for some $\gamma > 0$, then it holds that

$$\frac{\left(1 + r\widetilde{C}\right)^{\frac{1}{r}}}{1 + \widetilde{C}} \left(1 + 2 \frac{\phi\left(\frac{|\widetilde{m}|}{\sqrt{\widetilde{C}(1+\widetilde{C})}}\right)}{\frac{|\widetilde{m}|}{\sqrt{\widetilde{C}(1+\widetilde{C})}}}\right) \leqslant \frac{\left(1 + r\widetilde{C}\right)^{\frac{1}{r}}}{1 + \widetilde{C}} \left(1 + 2 \frac{\phi\left(\frac{\gamma \widetilde{C}^{\frac{1}{r}}}{\sqrt{\widetilde{C}(1+\widetilde{C})}}\right)}{\frac{\gamma \widetilde{C}^{\frac{1}{r}}}{\sqrt{\widetilde{C}(1+\widetilde{C})}}}\right) \tag{62}$$

because $\phi(z)/z$ is nonincreasing. We claim that, for γ sufficiently large, the right-hand side of this inequality is bounded from above by 1 for all $\widetilde{C} > 0$. Checking this claim is technical but not difficult, so we postpone the proof to Lemma A.4 in the Appendix. For such a value of γ, it holds by (61) that if $|\widetilde{m}| \geqslant \gamma \widetilde{C}^{1/r}$, then

$$\frac{|\widetilde{m}_\beta(\widetilde{m}, \widetilde{C})|}{\widetilde{C}_\beta(\widetilde{m}, \widetilde{C})^{\frac{1}{r}}} \leqslant \frac{|\widetilde{m}|}{\widetilde{C}^{\frac{1}{r}}}.$$

On the other hand, because $B(\bullet, \widetilde{C})$ (because the function $x \mapsto x + 2\phi(x)$ is increasing), it holds that, if $|\widetilde{m}| \leqslant \gamma \widetilde{C}^{1/r}$, then $B(|\widetilde{m}|, \widetilde{C}) \leqslant B(\gamma \widetilde{C}^{1/r}, \widetilde{C}) \leqslant \gamma$ by Lemma A.4 again, which proves the result. ■

Proof of Proposition 5. Let us first assume that $\alpha = 0$. Then, by (23), because the moments of successive iterates are related by

$$\widetilde{m}_{n+1} = \widetilde{m}_\beta(\widetilde{m}_n, \widetilde{C}_n) \qquad \text{and} \qquad \widetilde{C}_{n+1} = \widetilde{C}_\beta(\widetilde{m}_n, \widetilde{C}_n)$$

for this value of α, it holds by Lemma 11 that

$$\frac{|\widetilde{m}_{n+1}|}{\widetilde{C}_{n+1}^{1/r}} \leqslant \max\left(\gamma, \frac{|\widetilde{m}_n|}{\widetilde{C}_n^{1/r}}\right) \leqslant \dots \leqslant \max\left(\gamma, \frac{|\widetilde{m}_0|}{\widetilde{C}_0^{1/r}}\right), \tag{63}$$

which gives directly the convergence of $\widetilde{m}_n$ to 0, in view of the fact that $\widetilde{C}_n \to 0$ by Proposition 4. In the case where $\alpha \in (0,1)$, the moments of successive iterates are related by the equations

$$\begin{aligned}\widetilde{m}_{n+1} &= (1-\alpha)\widetilde{m}_\beta(\widetilde{m}_n, \widetilde{C}_n) + \alpha\widetilde{m}_n,\\ \widetilde{C}_{n+1} &= (1-\alpha^2)\widetilde{C}_\beta(\widetilde{m}_n, \widetilde{C}_n) + \alpha^2\widetilde{C}_n,\end{aligned}$$

so clearly

$$\begin{aligned}|\widetilde{m}_{n+1}| &\leqslant (1-\alpha)|\widetilde{m}_\beta(\widetilde{m}_n, \widetilde{C}_n)| + \alpha|\widetilde{m}_n|\\ &\leqslant (1-\alpha)|\widetilde{m}_\beta(\widetilde{m}_n, \widetilde{C}_n)| + \alpha \max\left(|\widetilde{m}_n|, \gamma\widetilde{C}_n^{1/r}\right) =: \widehat{m}_{n+1}.\end{aligned}$$

We will now use the technical Lemma A.5 in the Appendix with parameters

$$(\widehat{C}_\beta, \widehat{C}_n, \widehat{m}_\beta, \widehat{u}) = \left(\widetilde{C}_\beta(\widetilde{m}_n, \widetilde{C}_n), \widetilde{C}_n, |\widetilde{m}_\beta(\widetilde{m}_n, \widetilde{C}_n)|, \max\left(|\widetilde{m}_n|, \gamma\widetilde{C}_n^{1/r}\right)\right).$$

Using Lemma 11, we check that the assumptions of Lemma A.5 are satisfied:

$$\frac{\widehat{m}_\beta}{\widehat{C}_\beta^{1/r}} = \frac{|\widetilde{m}_\beta(\widetilde{m}_n, \widetilde{C}_n)|}{\widetilde{C}_\beta^{1/r}} \leqslant \max\left(\gamma, \frac{|\widetilde{m}_n|}{\widetilde{C}_n^{1/r}}\right) = \frac{\widehat{u}}{\widehat{C}_n^{1/r}},$$

so we deduce that, for $q = 2r$,

$$\frac{|\widetilde{m}_{n+1}|}{\widetilde{C}_{n+1}^{1/q}} \leqslant \frac{\widehat{m}_{n+1}}{\widetilde{C}_{n+1}^{1/q}} \leqslant \frac{\widehat{u}}{\widehat{C}_n^{1/q}} = \frac{\max\left(|\widetilde{m}_n|, \gamma\widetilde{C}_n^{1/r}\right)}{\widetilde{C}_n^{1/q}} = \max\left(\frac{|\widetilde{m}_n|}{\widetilde{C}_n^{1/q}}, \gamma\widetilde{C}_n^{1/r-1/q}\right).$$

Because $\widetilde{C}_n \leqslant \widetilde{C}_0$ by Proposition 4, this implies

$$\frac{|\widetilde{m}_{n+1}|}{\widetilde{C}_{n+1}^{1/q}} \leqslant \max\left(\frac{|\widetilde{m}_n|}{\widetilde{C}_n^{1/q}}, \gamma\widetilde{C}_0^{1/r-1/q}\right) \leqslant \dots \leqslant \max\left(\frac{|\widetilde{m}_0|}{\widetilde{C}_0^{1/q}}, \gamma\widetilde{C}_0^{1/r-1/q}\right),$$

implying the convergence of $\widetilde{m}_n \to 0$ with rate $n^{-1/q}$.

A similar reasoning can be employed to show the convergence in continuous time ; the details are omitted for conciseness. ∎

Proof of Proposition 6. Let us now obtain a convergence rate in the case where $\alpha = 0$. To this end, the main idea is to express that, close to equilibrium, i.e., when $C_n \ll 1$ and $|m_n - \theta_*| \ll 1$, the algorithm behaves similarly to how it would in a quadratic potential. Employing the same reasoning as in the derivation of (52a) and (52b), now using Taylor's theorem up to higher orders, we deduce

$$\int_{\mathbf{R}} g(\theta; m, C)\, \mathrm{e}^{-\beta f(\theta)}\, \mathrm{d}\theta = \mathrm{e}^{-\beta f(m)} \left(1 + \left(\beta^2 |f'(m)|^2 - \beta f''(m) \right) \frac{C}{2} \right) \\ + R_0(m, C), \tag{64a}$$

$$\int_{\mathbf{R}} (\theta - m)\, g(\theta; m, C)\, \mathrm{e}^{-\beta f(\theta)}\, \mathrm{d}\theta = -\,\mathrm{e}^{-\beta f(m)} \beta C f'(m) + R_1(m, C),$$

$$\int_{\mathbf{R}} (\theta - m)^2\, g(\theta; m, C)\, \mathrm{e}^{-\beta f(\theta)}\, \mathrm{d}\theta = \mathrm{e}^{-\beta f(m)}\, C \left(1 + \left(\beta^2 |f'(m)|^2 - \beta f''(m) \right) \frac{3C}{2} \right) \\ + R_2(m, C), \tag{64b}$$

with remainder terms (different from the ones in the proof of Theorem 1) satisfying

$$\forall m \in (\theta_* - 1, \theta_* + 1), \quad \forall 0 < C \leqslant 1, \qquad \begin{cases} |R_0(m, C)| \leqslant K |C|^2, \\ |R_1(m, C)| \leqslant K |C|^2, \\ |R_2(m, C)| \leqslant K |C|^3, \end{cases}$$

for an appropriate constant K. We claim that

$$m_\beta(m, C) - \theta_* = \left(C^{-1} + \beta f''(m) \right)^{-1} C^{-1} (m - \theta_*) + R_m(m, C) \tag{65a}$$

$$C_\beta(m, C) = \left(C^{-1} + \beta f''(m) \right)^{-1} + R_C(m, C), \tag{65b}$$

with R_m and R_C satisfying

$$\forall m \in (\theta_* - \overline{m}, \theta_* + \overline{m}), \quad \forall 0 < C < \overline{C} \qquad \begin{cases} |R_m(m, C)| \leqslant K \left(C^2 + C^2 |m - \theta_*| + C |m - \theta_*|^2 \right), \\ |R_C(m, C)| \leqslant K |C|^3, \end{cases}$$

for a possibly different constant K independent of m and C and appropriate positive constants $\overline{m}$ and $\overline{C}$. For completeness, let us present the details of the proof of (65a). To simplify the notation, we will write $u(m, C) = \mathcal{O}(v(m, C))$ to mean that there exist constants K, $\widetilde{m}$ and $\widetilde{C}$ such that $|u(m, C)| \leqslant K\, v(m, C)$ for all $m \in (\theta_* - \widetilde{m}, \theta_* + \widetilde{m})$ and for all $0 < C < \widetilde{C}$. It holds, by a Taylor expansion of the function $x \mapsto (1 + x)^{-1}$ around $x = 0$,

$$\begin{aligned} \left(C^{-1} + \beta f''(m) \right)^{-1} C^{-1} (m - \theta_*) &= m - \theta_* - C \beta f''(m)(m - \theta_*) + \mathcal{O}(C^2 |m - \theta_*|) \\ &= m - \theta_* - C \beta f'(m) + \mathcal{O}(C^2 |m - \theta_*| + C |m - \theta_*|^2) \\ &= m_\beta(m, C) - \theta_* + \mathcal{O}(C^2 + C^2 |m - \theta_*| + C |m - \theta_*|^2). \end{aligned}$$

In the second line, we used that $f''(m)(\theta_* - m) = f'(\theta_*) - f'(m) - \frac{1}{2} f'''(\xi)|\theta_* - m|^2$ by Taylor's theorem, for some appropriate ξ. Moreover, the third line is a consequence of the estimate

$$|m_\beta(m, C) - m + C\beta f'(m)| = \mathcal{O}(C^2)$$

due to (64a) and (64b). Equation (65b) can be shown using a similar approach, so we will omit its derivation. Combining (65a) and (65b), we deduce

$$\begin{aligned} C_\beta(m,C)^{-1}\big(m_\beta(m,C) - \theta_*\big) &= \frac{\big(C^{-1} + \beta f''(m)\big)^{-1} C^{-1}(m - \theta_*) + R_m(m,C)}{\frac{1}{C^{-1} + \beta f''(m)} + R_C(m,C)} \\ &= \frac{C^{-1}(m - \theta_*) + \big(C^{-1} + \beta f''(m)\big) R_m(m,C)}{1 + (C^{-1} + \beta f''(m)) R_C(m,C)} \\ &= C^{-1}(m - \theta_*) + \mathcal{O}(C + C|m - \theta_*| + |m - \theta_*|^2). \end{aligned}$$

Now let (m_n, C_n) denote the iterates of the optimization scheme. In view of the definition of the $\mathcal{O}$ notation, and because we already showed that $C_n \leqslant Kn^{-1}$ and $|m_n - \theta_*| \leqslant Kn^{-\frac{1}{r}}$ for some positive constant K and some $r > 2$ due to (63), the previous equation implies that there exists another constant K and an index k sufficiently large such that, for all $n \geqslant k$,

$$|C_n^{-1}(m_n - \theta_*) - C_k^{-1}(m_k - \theta_*)| \leqslant K \sum_{i=k}^{n-1} \left(C_i + \underbrace{C_i |m_i - \theta_*|}_{\text{summable}} + |m_i - \theta_*|^2 \right). \tag{66}$$

All the summands are bounded from above by the worst decay given by the last summand $i^{-\frac{2}{r}}$, up to a constant factor. Because

$$\sum_{i=k}^{n-1} i^{-\frac{2}{r}} \leqslant \int_{k-1}^{n-1} x^{-\frac{2}{r}}\, \mathrm{d}x \leqslant \widetilde{K} n^{1-\frac{2}{r}}, \tag{67}$$

with $\widetilde{K}$ a constant independent of n changing from occurrence to occurrence, we deduce that the right-hand side of (66) is controlled by $\widetilde{K}(1 + n^{1-\frac{2}{r}})$. Therefore, using the fact that $C_n \to 0$ with rate $1/n$, we obtain

$$\forall n \geqslant k, \qquad |m_n - \theta_*| \leqslant \left(\frac{\widetilde{K}}{n}\right) C_k^{-1}(m_k - \theta_*) + \widetilde{K}\left(n^{-1} + n^{-\frac{2}{r}}\right).$$

We have thus upgraded the convergence rate to $n^{-\frac{2}{r}}$. This procedure can be repeated until only the first term in the sum on the right-hand side of (66) is nonsummable, leading finally to the estimate

$$|m_n - \theta_*| \leqslant \widetilde{K}\left(\frac{\log n}{n}\right),$$

by a similar argument as in (67) applied to the decay $1/i$. ∎

5.5 | Proof of Theorem 3

In this section, we analyze the mean-field dynamics eqns. (18, 16). We show, in the convex one-dimensional case, the existence and uniqueness of a steady state close to the Laplace approximation of the Bayesian posterior at the MAP estimator. We begin by showing a version of Laplace's method, which is based on reducing all information about the objective function f into the unique smooth and increasing function $\tau : \mathbf{R} \to \mathbf{R}$ satisfying

$$\forall \theta \in \mathbf{R}, \qquad f(\theta_* + \tau(\theta)) = f(\theta_*) + \theta^2. \tag{68}$$

with $\tau(0) = 0$. For details, see Lemma A.7.

Proposition 7 (Laplace's method). *Let $d = 1$. Suppose Assumptions 1 and 4 hold, and assume additionally that φ is a smooth function such that*

$$\forall i \in \{0, \dots, 2N+2\}, \qquad \|\varphi^{(i)}\|_\infty \leqslant M_\varphi < \infty, \tag{69}$$

for some $N \in \mathbf{N}$ and some $M_\varphi \geqslant 0$. Then, introducing the function $\psi(\theta) = \varphi(\theta_* + \tau(\theta))\,\tau'(\theta)$, where τ is the map provided by Lemma A.7, it holds

$$I_\beta := \int_{\mathbf{R}} \mathrm{e}^{-\beta f(\theta)}\, \varphi(\theta)\, \mathrm{d}\theta = \mathrm{e}^{-\beta f(\theta_*)} \left(\sum_{n=0}^{N} \psi_{2n} \frac{\Gamma(n+1/2)}{\beta^{n+1/2}} + R_\beta \right), \qquad \psi_{2n} := \frac{\psi^{(2n)}(0)}{(2n)!},$$

and the remainder R_β satisfies the bound

$$|R_\beta| \leqslant \frac{K M_\varphi}{(\beta - \beta_0)^{N+3/2}},$$

for some constants $K = K(f, N) > 0$ and $\beta_0 = \beta_0(f, N) \geqslant 0$.

Proof. Applying Lemma A.7, we can use the change of variable $\theta \mapsto \theta_* + \tau(\theta)$ to obtain

$$\forall \varphi \in C^\infty(\mathbf{R}), \qquad \int_{\mathbf{R}} \mathrm{e}^{-\beta f(\theta)}\, \varphi(\theta)\, \mathrm{d}\theta = \mathrm{e}^{-\beta f(\theta_*)} \int_{\mathbf{R}} \mathrm{e}^{-\beta \theta^2}\, \varphi(\theta_* + \tau(\theta))\, \tau'(\theta)\, \mathrm{d}\theta =: \mathrm{e}^{-\beta f(\theta_*)}\, \widetilde{I}_\beta.$$

By Faà di Bruno's formula (generalized chain rule), we have

$$\forall n \in \mathbf{N}, \qquad \frac{\mathrm{d}^n}{\mathrm{d}\theta^n} \big(\varphi(\theta_* + \tau(\theta))\, \tau'(\theta)\big) = \sum_{i=0}^{n} \varphi^{(i)}(\theta_* + \tau(\theta))\, B_{n+1,i+1}\big(\tau'(\theta), \dots, \tau^{(n-i+1)}(\theta)\big),$$

where, for $n \in \mathbf{N}$, the functions $\{B_{n,i}\}_{i\in\{0,\dots,n\}}$ are polynomials (more precisely, Bell polynomials) of degree 0 to n. By Lemma A.7, there exists a constant $\lambda = \lambda(f, N) \geqslant 0$ such that

$$\forall i \in \{0, \dots, 2N+3\}, \qquad \| \mathrm{e}^{-\lambda\theta^2}\, \tau^{(i)}(\theta)\|_\infty < \infty.$$

It is clear, therefore, that

$$\forall n \in \{0, \dots, 2N+2\}, \quad \forall i \in \{0, \dots, n\},$$
$$\|\mathrm{e}^{-(i+1)\lambda\theta^2} B_{n+1,i+1}\left(\tau'(\theta), \dots, \tau^{(n-i+1)}(\theta)\right)\|_\infty < \infty.$$

Combining this inequality with (69), we deduce that there exists $K = K(f, N)$ such that

$$\forall n \in \{0, \dots, 2N+2\}, \qquad \|\mathrm{e}^{-((2N+3)\lambda)\theta^2} \frac{\mathrm{d}^n}{\mathrm{d}\theta^n} \left(\varphi\left(\theta_* + \tau(\theta)\right) \tau'(\theta)\right)\|_\infty \leqslant K\, M_\varphi < \infty.$$

It follows that, in particular, the assumptions of Lemma A.6 are satisfied for the function $\psi(\theta)$, with the parameters $M = K\,M_\varphi$ and $\beta_0 = (2N+3)\lambda$. By Lemma A.6, it holds that

$$\widetilde{I}_\beta = \sum_{n=0}^{N} \psi_{2n} \frac{\Gamma(n+1/2)}{\beta^{n+1/2}} + R_\beta, \qquad \psi_{2n} := \frac{\psi^{(2n)}(0)}{(2n)!},$$

where the remainder R_β satisfies the bound

$$|R_\beta| \leqslant \frac{M}{(2N+2)!} \frac{\Gamma(N+3/2)}{(\beta-\beta_0)^{N+3/2}},$$

which concludes the proof. ■

To prove Theorem 3, let us now introduce the following map on $\mathbf{R}\times\mathbf{R}_{>0}$:

$$\Phi_\beta : \begin{pmatrix} m \\ C \end{pmatrix} \mapsto \begin{pmatrix} m_\beta(m, C) \\ \lambda^{-1}\, C_\beta(m, C) \end{pmatrix}, \qquad \lambda = (1+\beta)^{-1}. \tag{70}$$

In view of Lemma 1, existence of a fixed point of Φ_β implies the existence of a steady-state solution both for the iterative scheme (16) with any $\alpha \in [0, 1)$ and for the nonlinear Fokker–Planck equation eq. 18. To prove the existence of a fixed point of Φ_β we will apply Laplace's method Proposition 7, and therefore need to calculate the coefficients ψ_{2n}, which requires the calculation of the derivatives of the smooth function τ at 0. This can be achieved by implicit differentiation of the Equation (68). For example, differentiating twice, we obtain

$$\tau'(0) = \pm\sqrt{\frac{2}{f''(\theta_*)}}.$$

Because, τ refers here to the unique increasing function such that (68) holds, only the positive solution is retained. Differentiating (68) again we obtain

$$\tau''(0) = -\frac{f'''(\theta_*)}{3f''(\theta_*)}\,|\tau'(0)|^2.$$

The following result therefore implies the existence of steady state close to the Laplace approximation of the target distribution both for the iterative scheme (16) with any $\alpha \in [0,1)$ and for the nonlinear Fokker–Planck equation eq. 18.

Proposition 8 (Existence of a fixed point of Φ_β)**.** *Let $d = 1$ and assume that Assumptions 1 and 4 hold. Then there exist $\widetilde{k} = \widetilde{k}(f)$ and $\widetilde{\beta} = \widetilde{\beta}(f)$ such that, for all $\beta \geqslant \widetilde{\beta}$, there exists a fixed point $(m_\infty(\beta), C_\infty(\beta))$ of Φ_β satisfying*

$$|m_\infty(\beta) - \theta_*|^2 + |C_\infty(\beta) - C_*|^2 \leqslant \left|\frac{\widetilde{k}}{\beta}\right|^2.$$

Proof. It is clear from the definitions of m_β and C_β that the map Φ_β is continuous. Our approach to show the existence of a fixed point is to use Brouwer's fixed point theorem. To this end, let us define

$$\varphi_j(\theta) = (\theta - \theta_*)^j\, g(\theta; m, C), \qquad j = 0, 1, \dots, J.$$

Introducing the function $\hat{\theta} : \mathbf{R} \ni u \mapsto m + \sqrt{C}u$, and using the notation $g(u) := g(u; 0, 1)$ for conciseness, we calculate

$$\begin{aligned}\hat{\varphi}_j(u) := \varphi_j\left(\hat{\theta}(u)\right) &= \frac{1}{\sqrt{C}}\left(m - \theta_* + \sqrt{C}u\right)^j g(u; 0, 1)\\ &= C^{\frac{j-1}{2}}\left(\frac{m-\theta_*}{\sqrt{C}} + u\right)^j g(u; 0, 1) = C^{\frac{j-1}{2}} \sum_{k=0}^{j} \binom{j}{k}\left(\frac{m-\theta_*}{\sqrt{C}}\right)^k u^{j-k} g(u; 0, 1),\end{aligned}$$

so we deduce

$$\left\|\hat{\varphi}_j^{(n)}\right\|_\infty \leqslant K_{j,n}\, C^{\frac{j-1}{2}}\left(1 + \left|\frac{m-\theta_*}{\sqrt{C}}\right|^j\right) = K_{j,n}\left(C^{\frac{j-1}{2}} + C^{-\frac{1}{2}}|m - \theta_*|^j\right),$$

for some constant $K_{j,n}$ independent of m and C. Because $\varphi_j^{(n)}(\theta) = C^{-n/2}\hat{\varphi}_j^{(n)}\left(C^{-1/2}(\theta - m)\right)$, this directly implies

$$\left\|\varphi_j^{(n)}\right\|_\infty \leqslant K_{j,n}\left(C^{\frac{j-n-1}{2}} + C^{-\frac{n+1}{2}}|m - \theta_*|^j\right). \tag{71}$$

Let us take any $R \in (0, C_*)$ and introduce the notation $u(\beta, m, C) = \mathcal{O}_R(v(\beta))$ for any functions $u(\beta, m, C)$ and $v(\beta)$ to mean that there exist constants c and $\widetilde{\beta}$ such that

$$\forall (m, C) \in B_R(\theta_*, C_*), \quad \forall \beta > \widetilde{\beta}, \qquad |u(\beta, m, C)| \leqslant c v(\beta),$$

where $B_R(\theta_*, C_*)$ denotes the closed ball of radius R centered at (θ_*, C_*). Because $R < C_*$, it is clear that, for all $j \in \mathbf{N}$ and $N \in \mathbf{N}$, the right-hand side of (71) is bounded from above by a constant over $B_R(\theta_*, C_*)$, uniformly in m, C and $n \in \{0, \dots, 2N+2\}$. Thus, we can apply Laplace's method, Proposition 7. Letting $\psi_j(\theta) = \varphi_j(\theta_* + \tau(\theta))\tau'(\theta)$, we calculate

$$\psi_j(0) = \varphi_j(\theta_*)\tau'(0) = \varphi_j(\theta_*)\sqrt{\frac{2}{f''(\theta_*)}},$$

$$\psi_j''(0) = \varphi_j''(\theta_*)\tau'(0)^3 + 3\varphi_j'(\theta_*)\,\tau''(0)\,\tau'(0) + \varphi_j(\theta_*)\,\tau'''(0).$$

Note that only the first term in the expression of $\psi_2''(0)$ is nonzero. Therefore, Laplace's method applied with $N = 0$ or $N = 1$ gives

$$\mathrm{e}^{\beta f(\theta_*)} \int_{\mathbf{R}} \mathrm{e}^{-\beta f}\, \mathrm{g}(\theta; m, C)\, \mathrm{d}\theta = \mathrm{g}(\theta_*; m, C)\, \Gamma(1/2)\, \frac{\tau'(0)}{\beta^{1/2}} + \mathcal{O}_R\left(\frac{1}{\beta^{3/2}}\right), \tag{73a}$$

$$\mathrm{e}^{\beta f(\theta_*)} \int_{\mathbf{R}} (\theta - \theta_*)\, \mathrm{e}^{-\beta f}\, \mathrm{g}(\theta; m, C)\, \mathrm{d}\theta = \mathcal{O}_R\left(\frac{1}{\beta^{3/2}}\right), \tag{73b}$$

$$\mathrm{e}^{\beta f(\theta_*)} \int_{\mathbf{R}} (\theta - \theta_*)^2\, \mathrm{e}^{-\beta f}\, \mathrm{g}(\theta; m, C)\, \mathrm{d}\theta = \mathrm{g}(\theta_*; m, C)\, \Gamma(3/2)\, \frac{\tau'(0)^3}{\beta^{3/2}} + \mathcal{O}_R\left(\frac{1}{\beta^{5/2}}\right). \tag{73c}$$

Further, $\mathrm{g}(\theta_*; m, C)$ is bounded above and below on $B_R(\theta_*, C_*)$ by positive constants. Hence, Equation (73b) leads to

$$m_\beta(m, C) = \frac{\int_{\mathbf{R}} \theta\, \mathrm{e}^{-\beta f(\theta)}\, \mathrm{g}(\theta; m, C)\, \mathrm{d}\theta}{\int_{\mathbf{R}} \mathrm{e}^{-\beta f(\theta)}\, \mathrm{g}(\theta; m, C)\, \mathrm{d}\theta} = \theta_* + \mathcal{O}_R\big(\beta^{-1}\big). \tag{74}$$

For the covariance term, note that

$$C_\beta(m, C) = \frac{\int_{\mathbf{R}} (\theta - \theta_*)^2\, \mathrm{e}^{-\beta f(\theta)}\, \mathrm{g}(\theta; m, C)\, \mathrm{d}\theta}{\int_{\mathbf{R}} \mathrm{e}^{-\beta f(\theta)}\, \mathrm{g}(\theta; m, C)\, \mathrm{d}\theta} - \big(m_\beta(m, C) - \theta_*\big)^2,$$

which by (73c) and the equality $\Gamma(1/2) = 2\Gamma(3/2)$, leads to

$$\lambda^{-1} C_\beta(m, C) = \frac{1 + \beta}{\beta} \left(\frac{\Gamma(3/2)}{\Gamma(1/2)} |\tau'(0)|^2 + \mathcal{O}_R(\beta^{-1}) \right) + \mathcal{O}_R(\beta^{-2}) = \frac{1}{f''(\theta_*)} + \mathcal{O}_R(\beta^{-1}).$$

Consequently, we deduce by definition of $\mathcal{O}_R$ that there exist constants $\beta^\dagger$ and $\widetilde{k}$ such that

$$\forall \beta > \beta^\dagger, \qquad \sup_{(m,C)\in B_R(\theta_*,C_*)} |\Phi_\beta(m,C) - (\theta_*,C_*)| \leqslant \frac{\widetilde{k}}{\beta^\dagger},$$

where $|\bullet|$ for any $\beta \geqslant \beta^\dagger$. If additionally $\beta \geqslant \widetilde{k}/R$, we have $B_{\widetilde{k}/\beta}(\theta_*,C_*) \subset B_R(\theta_*,C_*)$ and so

$$\Phi_\beta\left(B_{\widetilde{k}/\beta}(\theta_*,C_*)\right) \subset \Phi_\beta(B_R(\theta_*,C_*)) \subset B_{\widetilde{k}/\beta}(\theta_*,C_*).$$

Consequently, in this case Brouwer's theorem implies the existence of a fixed point of Φ_β in $B_{\widetilde{k}/\beta}(\theta_*,C_*)$. This proves the statement with $\widetilde{\beta} = \max(\beta^\dagger, \widetilde{k}/R)$. ■

Next, we show that the map Φ_β given in (70) is a contraction for sufficiently large β.

Proposition 9 (Φ_β is a contraction). *Under the same assumptions as in Proposition 8 and for any $R \in (0, C_*)$, there exists a constant $\widehat{\beta} = \widehat{\beta}(f,R)$ and $\widehat{k} = \widehat{k}(f,R)$ such that, for all $\beta \geqslant \widehat{\beta}$, the map Φ_β is a contraction with constant $\widehat{k}/\beta$ for the Euclidean norm over the closed ball of radius R centered at (θ_*,C_*): for all (m_1,C_1) and (m_2,C_2) in $B_R(\theta_*,C_*)$, it holds that*

$$\left|\Phi_\beta(m_1,C_1) - \Phi_\beta(m_2,C_2)\right| \leqslant \frac{\widehat{k}}{\beta}\left|\begin{pmatrix} m_2 \\ C_2 \end{pmatrix} - \begin{pmatrix} m_1 \\ C_1 \end{pmatrix}\right|.$$

Proof. We assume without loss of generality that $\theta_* = 0$, which is justified because the method is affine invariant, discussed in Section 2.3, and we recall that Φ_β relates the moments of successive iterates from (5) with $\alpha = 0$ when this scheme is initialized at a Gaussian density. Let us introduce the notation

$$J_\beta(\varphi) = \int_{\mathbf{R}} \varphi(\theta) \exp\left(-\frac{|\theta - m|^2}{2C}\right) \mathrm{e}^{-\beta f(\theta)}\, \mathrm{d}\theta.$$

Using the fact that

$$m_\beta(m,C) = \frac{J_\beta(\theta)}{J_\beta(1)}, \qquad C_\beta(m,C) = \frac{J_\beta(\theta^2)}{|J_\beta(1)|} - |m_\beta(m,C)|^2 = \frac{J_\beta(\theta^2)J_\beta(1) - |J_\beta(\theta)|^2}{|J_\beta(1)|^2},$$

and noting that

$$\partial_m J_\beta(\varphi) = \frac{J_\beta(\varphi(\theta)(\theta - m))}{C}, \qquad \partial_C J_\beta(\varphi) = \frac{J_\beta\left(\varphi(\theta)|\theta - m|^2\right)}{2C^2},$$

we calculate

$$\partial_m m_\beta = \frac{1}{C|J_\beta(1)|^2}\left(J_\beta(\theta^2)J_\beta(1) - |J_\beta(\theta)|^2\right),$$

$$\partial_C m_\beta = \frac{1}{2C^2|J_\beta(1)|^2}\left(J_\beta\left(\theta|\theta - m|^2\right)J_\beta(1) - J_\beta\left(|\theta - m|^2\right)J_\beta(\theta)\right)$$

$$
\begin{aligned}
&= \frac{1}{2C^2|J_\beta(1)|^2}\Big[J_\beta\big(\theta^3\big)J_\beta(1) - J_\beta\big(\theta^2\big)J_\beta(\theta) - 2m\big(J_\beta\big(\theta^2\big)J_\beta(1) - J_\beta(\theta)^2\big)\Big]\,, \\
\partial_m C_\beta &= \frac{1}{C|J_\beta(1)|^2}\big(J_\beta\big(\theta^3\big)J_\beta(1) - J_\beta(\theta)J_\beta(\theta^2)\big) - 2m_\beta\,\partial_m m_\beta\,, \\
\partial_C C_\beta &= \frac{1}{2C^2|J_\beta(1)|^2}\big(J_\beta\big(\theta^2|\theta-m|^2\big)J_\beta(1) - J_\beta\big(|\theta-m|^2\big)J_\beta(\theta^2)\big) - 2m_\beta\,\partial_C m_\beta \\
&= \frac{1}{2C^2|J_\beta(1)|^2}\Big[J_\beta\big(\theta^4\big)J_\beta(1) - J_\beta(\theta^2)^2 - 2m\big(J_\beta\big(\theta^3\big)J_\beta(1) - J_\beta(\theta^2)J_\beta(\theta)\big)\Big] - 2m_\beta\,\partial_C m_\beta\,.
\end{aligned}
$$

Applying Laplace's method, and noting that $\frac{\mathrm{d}^n}{\mathrm{d}\theta^n}(\theta^j g(\theta;m,C))$ vanishes at $\theta = \theta_* = 0$ for all $n < j$, we obtain that

$$
\begin{aligned}
\mathrm{e}^{\beta f(\theta_*)}\int_{\mathbf{R}} \theta^3\,\mathrm{e}^{-\beta f}\,g(\theta;m,C)\,\mathrm{d}\theta &= \mathcal{O}_R\left(\frac{1}{\beta^{5/2}}\right), \\
\mathrm{e}^{\beta f(\theta_*)}\int_{\mathbf{R}} \theta^4\,\mathrm{e}^{-\beta f}\,g(\theta;m,C)\,\mathrm{d}\theta &= \mathcal{O}_R\left(\frac{1}{\beta^{5/2}}\right).
\end{aligned}
$$

Combining these estimates with eqns. (73a), (73b), (73c) and (74), and using the same notation as in the proof of Proposition 8, we deduce

$$
\begin{aligned}
\partial_m m_\beta(m,C) &= \mathcal{O}_R(\beta^{-1}), & \partial_C m_\beta(m,C) &= \mathcal{O}_R(\beta^{-1}), \\
\partial_m C_\beta(m,C) &= \mathcal{O}_R(\beta^{-2}), & \partial_C C_\beta(m,C) &= \mathcal{O}_R(\beta^{-2}).
\end{aligned}
$$

It easily follows that

$$
D\Phi_\beta := \begin{pmatrix} \partial_m \Phi^m_\beta & \partial_C \Phi^m_\beta \\ \partial_m \Phi^C_\beta & \partial_C \Phi^C_\beta \end{pmatrix} = \begin{pmatrix} \mathcal{O}_R(\beta^{-1}) & \mathcal{O}_R(\beta^{-1}) \\ \mathcal{O}_R(\beta^{-1}) & O_R(\beta^{-1}) \end{pmatrix}. \tag{75}
$$

Therefore, for all $(m_1, C_1) \in B_R(\theta_*, C_*)$ and $(m_2, C_2) \in B_R(\theta_*, C_*)$, it holds

$$
\begin{aligned}
|\Phi_\beta(m_1, C_1) - \Phi_\beta(m_2, C_2)| &= \left|\int_0^1 D\Phi_\beta\,(m_t, C_t)\begin{pmatrix} m_2 - m_1 \\ C_2 - C_1 \end{pmatrix}\,\mathrm{d}t\right| \\
&\leqslant \int_0^1 \|D\Phi_\beta\,(m_t, C_t)\|\,\mathrm{d}t\left|\begin{pmatrix} m_2 \\ C_2 \end{pmatrix} - \begin{pmatrix} m_1 \\ C_1 \end{pmatrix}\right|,
\end{aligned}
$$

where $(m_t, C_t)^\top = \big(m_1 + t(m_2 - m_1), C_1 + t(C_2 - C_1)\big)^\top$. Since $\|D\Phi_\beta\| = \mathcal{O}_R(\beta^{-1})$, by (75), this concludes the proof of the statement. ■

Theorem 3 is now a simple consequence of Propositions 8 and 9.

Proof of Theorem 3. Let $\widetilde{\beta}$ and $\widehat{\beta}$, as well as $\widetilde{k}$ and $\widehat{k}$, be as given in the statements of Propositions 8 and 9, respectively. Let $\underline{\beta}(f,R)$ and $k(f,R)$ be defined by

$$\underline{\beta} = \max\left(\widetilde{\beta}(f), \widehat{\beta}(f,R), \frac{\widetilde{k}}{R}\right), \qquad k(f,R) = \max\left(\widetilde{k}(f), \widehat{k}(f,R)\right).$$

By Proposition 8, there exists for all $\beta \geqslant \underline{\beta}$ a fixed point of Φ_β in $B_{\widetilde{k}/\beta}(\theta_*, C_*) \subset B_R(\theta_*, C_*)$. Because Φ_β is a contraction over $B_R(\theta_*, C_*)$ for such value of β by Proposition 9, this fixed point is unique in $B_R(\theta_*, C_*)$. Let us now show the convergence to the fixed point in the discrete and continuous-time cases.

(i) Case $\alpha \in [0,1)$. We consider the iteration (10),

$$\begin{aligned} m_{n+1} &= \alpha m_n + (1-\alpha) m_\beta(m_n, C_n), \\ C_{n+1} &= \alpha^2 C_n + (1-\alpha^2)\lambda^{-1} C_\beta(m_n, C_n). \end{aligned}$$

Denoting the fixed point by $(m_\infty, C_\infty)^\mathsf{T}$, we rewrite this system as

$$\begin{aligned} m_{n+1} - m_\infty &= \alpha(m_n - m_\infty) + (1-\alpha)\big(m_\beta(m_n, C_n) - m_\beta(m_\infty, C_\infty)\big), \\ C_{n+1} - C_\infty &= \alpha^2(C_n - C_\infty) + (1-\alpha^2)\lambda^{-1}\big(C_\beta(m_n, C_n) - C_\beta(m_\infty, C_\infty)\big), \end{aligned}$$

and so, by the triangle inequality,

$$\begin{aligned} \left|\begin{pmatrix} m_{n+1} \\ C_{n+1} \end{pmatrix} - \begin{pmatrix} m_\infty \\ C_\infty \end{pmatrix}\right| &\leqslant \alpha \left|\begin{pmatrix} m_n \\ C_n \end{pmatrix} - \begin{pmatrix} m_\infty \\ C_\infty \end{pmatrix}\right| + (1-\alpha^2)|\Phi_\beta(m_n, C_n) - \Phi_\beta(m_\infty, C_\infty)| \\ &\leqslant \left(\alpha + (1-\alpha^2)\frac{k}{\beta}\right)\left|\begin{pmatrix} m_n \\ C_n \end{pmatrix} - \begin{pmatrix} m_\infty \\ C_\infty \end{pmatrix}\right|, \end{aligned}$$

from where the statement follows easily.

(ii) Case $\alpha = 1$. Similarly, in the continuous-time setting, we can rewrite Equations (22) for the moments as

$$\begin{aligned} \dot{m}(t) &= -(m(t) - m_\infty) + \big(m_\beta(m(t), C(t)) - m_\beta(m_\infty, C_\infty)\big), \\ \dot{C}(t) &= -2(C(t) - C_\infty) + 2\lambda^{-1}\big(C_\beta(m(t), C(t)) - C_\beta(m_\infty, C_\infty)\big). \end{aligned}$$

Therefore

$$\begin{aligned} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left|\begin{pmatrix} m(t) \\ C(t) \end{pmatrix} - \begin{pmatrix} m_\infty \\ C_\infty \end{pmatrix}\right|^2 &\leqslant -\left|\begin{pmatrix} m(t) \\ C(t) \end{pmatrix} - \begin{pmatrix} m_\infty \\ C_\infty \end{pmatrix}\right|^2 \\ &\quad + 2|\Phi_\beta(m_n, C_n) - \Phi_\beta(m_\infty, C_\infty)|\left|\begin{pmatrix} m(t) \\ C(t) \end{pmatrix} - \begin{pmatrix} m_\infty \\ C_\infty \end{pmatrix}\right| \end{aligned}$$

$$\leqslant -\left(1-\frac{2k}{\beta}\right)\left|\begin{pmatrix} m_n \\ C_n \end{pmatrix} - \begin{pmatrix} m_\infty \\ C_\infty \end{pmatrix}\right|^2,$$

which leads to the statement by Grönwall's inequality. ■

ACKNOWLEDGMENTS

The authors are grateful to Zehua Lai for pointing out that the Poincaré inequality could be employed for proving Lemma 5. JAC was supported by the Advanced Grant Nonlocal-CPD (Nonlocal PDEs for Complex Particle Dynamics: Phase Transitions, Patterns and Synchronization) of the European Research Council Executive Agency (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 883363) and by EPSRC grant number EP/T022132/1. JAC and UV were also supported by EPSRC grant number EP/P031587/1. FH was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - GZ 2047/1, Projekt-ID 390685813. AMS is supported by NSF (award AGS-1835860), by NSF (award DMS-1818977) and by the Office of Naval Research (award N00014-17-1-2079). UV was also supported by the Fondation Sciences Mathématiques de Paris (FSMP), through a postdoctoral fellowship in the "mathematical interactions" program.

ORCID

J. A. Carrillo https://orcid.org/0000-0001-8819-4660
F. Hoffmann https://orcid.org/0000-0002-1182-5521
A. M. Stuart https://orcid.org/0000-0001-9091-7266
U. Vaes https://orcid.org/0000-0002-7629-7184

REFERENCES

1. Kaipio J, Somersalo E. *Statistical and Computational Inverse Problems. Vol. 160 of Applied Mathematical Sciences*. New York: Springer-Verlag; 2005.
2. Dashti M, Law KJH, Stuart AM, Voss J. MAP estimators and their consistency in Bayesian nonparametric inverse problems. *Inverse Probl*. 2013;29(9):095017, 27.
3. Engl HW, Hanke M, Neubauer A. *Regularization of Inverse Problems. Vol. 375 of Mathematics and Its Applications*. Dordrecht: Kluwer Academic Publishers Group; 1996.
4. Lu Y, Stuart A, Weber H. Gaussian approximations for probability measures on $\mathbb{R}^d$. *SIAM/ASA J Uncertain Quantif*. 2017;5(1):1136–1165.
5. van der Vaart AW. *Asymptotic Statistics. Vol. 3 of Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge: Cambridge University Press; 1998.
6. Shun Z, McCullagh P. Laplace approximation of high-dimensional integrals. *J Roy Statist Soc Ser B*. 1995;57(4):749–760.
7. Carrillo JA, Choi YP, Totzeck C, Tse O. An analytical framework for consensus-based global optimization method. *Math Models Methods Appl Sci*. 2018;28(6):1037–1066.
8. Carrillo JA, Jin S, Li L, Zhu Y. A consensus-based global optimization method for high dimensional machine learning problems. *ESAIM Control Optim Calc Var*. 2021;27(suppl.):Paper No. S5, 22.
9. Pinnau R, Totzeck C, Tse O, Martin S. A consensus-based model for global optimization and its mean-field limit. *Math Models Methods Appl Sci*. 2017;27(1):183–204.
10. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. *J Chem Phys*. 1953;21(6):1087–1092.
11. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*. 1970;57(1):97–109.
12. Brooks S, Gelman A, Jones GL, Meng XL, editors. *Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. Boca Raton, FL: CRC Press; 2011.

13. Del Moral P, Doucet A, Jasra A. Sequential Monte Carlo samplers. *J R Stat Soc Ser B Stat Methodol*. 2006;68(3):411–436.
14. Liggett TM. *Interacting Particle Systems. Classics in Mathematics*. Berlin: Springer-Verlag; 2005. Reprint of the 1985 original.
15. Swart JM. A course in interacting particle systems. arXiv e-prints. 2017 Mar;1703.10007.
16. Bolley F, Cañizo JA, Carrillo JA. Stochastic mean-field limit: non-Lipschitz forces and swarming. *Math Models Methods Appl Sci*. 2011;21(11):2179–2210.
17. Bolley F, Carrillo JA. Nonlinear diffusion: geodesic convexity is equivalent to Wasserstein contraction. *Comm Partial Diff Equat*. 2014;39(10):1860–1869.
18. Carrillo JA, Fornasier M, Toscani G, Vecil F. Particle, kinetic, and hydrodynamic models of swarming. In: *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences*. Naldi Giovanni, Pareschi Lorenzo, Toscani Giuseppe, (eds). Boston, MA: Birkhäuser 2010;297–336.
19. Jabin PE, Wang Z. Mean field limit for stochastic particle systems. In: *Active Particles. Vol. 1. Advances in Theory, Models, and Applications*. Bellomo Nicola, Degond Pierre, Tadmor Eitan, (eds). Cham: Birkhäuser/Springer. 2017;379–402.
20. Sznitman AS. Topics in propagation of chaos. In: *École d'Été de Probabilités de Saint-Flour XIX—1989. vol. 1464 of Lecture Notes in Math*. Berlin: Springer; 1991:165–251.
21. Bunch P, Godsill S. Approximations of the optimal importance density using Gaussian particle flow importance sampling. *J Amer Statist Assoc*. 2016;111(514):748–762.
22. Reich S. A dynamical systems framework for intermittent data assimilation. *BIT*. 2011;51(1):235–249.
23. Van Leeuwen PJ, Künsch HR, Nerger L, Potthast R, Reich S. Particle filters for high-dimensional geoscience applications: a review. *Q J R Meteorol Soc*. 2019;145(723):2335–2365.
24. Yang T, Mehta PG, Meyn SP. Feedback particle filter. *IEEE Trans Automat Control*. 2013;58(10):2465–2480.
25. Reich S, Cotter C. *Probabilistic Forecasting and Bayesian Data Assimilation*. New York: Cambridge University; 2015.
26. Chen Y, Oliver DS. Ensemble randomized maximum likelihood method as an iterative ensemble smoother. *Math Geosci*. 2012 Jan;44(1):1–26.
27. Emerick AA, Reynolds AC. Investigation of the sampling performance of ensemble-based methods with a simple reservoir model. *Comput Geosci*. 2013;17(2):325–350.
28. Bergemann K, Reich S. A localization technique for ensemble Kalman filters. *Q J R Meteorol Soc*. 2010;136(648):701–707.
29. Bergemann K, Reich S. An ensemble Kalman–Bucy filter for continuous data assimilation. *Meteorol Z*. 2012;21(3):213.
30. Iglesias MA. A regularizing iterative ensemble Kalman method for PDE-constrained inverse problems. *Inverse Probl*. 2016;32(2):025002, 45.
31. Iglesias MA, Law KJH, Stuart AM. Ensemble Kalman methods for inverse problems. *Inverse Probl*. 2013;29(4):045001, 20.
32. Schillings C, Stuart AM. Analysis of the ensemble Kalman filter for inverse problems. *SIAM J Numer Anal*. 2017;55(3):1264–1290.
33. Schillings C, Stuart AM. Convergence analysis of ensemble Kalman inversion: the linear, noisy case. *Appl Anal*. 2018;97(1):107–123.
34. Garbuno-Inigo A, Hoffmann F, Li W, Stuart AM. Interacting Langevin diffusions: gradient structure and ensemble Kalman sampler. *SIAM J Appl Dyn Syst*. 2020;19(1):412–441.
35. Carrillo JA, Vaes U. Wasserstein stability estimates for covariance-preconditioned Fokker-Planck equations. *Nonlinearity*. 2021;34:2275–2295.
36. Garbuno-Inigo A, Nüsken N, Reich S. Affine invariant interacting Langevin dynamics for Bayesian inference. *SIAM J Appl Dyn Syst*. 2020;19(3):1633–1658.
37. Nüsken N, Reich S. Note on interacting Langevin diffusions: gradient structure and ensemble Kalman sampler by Garbuno-Inigo, Hoffmann, Li and Stuart. arXiv e-prints. 2019;1908.10890.
38. Dorigo M., Blum C. Ant colony optimization theory: A survey. *Theoretical Computer Science*. 2005;344:(2-3):243–278. https://doi.org/10.1016/j.tcs.2005.05.020
39. Kennedy J. Particle swarm optimization. In: *Encyclopedia of Machine Learning*. Springer; 2010:760–766.
40. Cucker F, Smale S. On the mathematics of emergence. *Jpn J Math*. 2007;2(1):197–227.

41. Ha SY, Liu JG. A simple proof of the Cucker-Smale flocking dynamics and mean-field limit. *Commun Math Sci*. 2009;7(2):297–325.
42. Motsch S, Tadmor E. Heterophilious dynamics enhances consensus. *SIAM Rev*. 2014;56(4):577–621.
43. Toscani G. Kinetic models of opinion formation. *Commun Math Sci*. 2006;4(3):481–496.
44. Carrillo JA, Fornasier M, Rosado J, Toscani G. Asymptotic flocking dynamics for the kinetic Cucker-Smale model. *SIAM J Math Anal*. 2010;42(1):218–236.
45. Ha SY, Jin S, Kim D. Convergence of a first-order consensus-based global optimization algorithm. *Math Models Methods Appl Sci*. 2020;30(12):2417–2444.
46. Jin S, Li L, Liu JG. Random batch methods (RBM) for interacting particle systems. *J Comput Phys*. 2020;400:108877, 30.
47. Fornasier M, Huang H, Pareschi L, Sünnen P. Consensus-based optimization on hypersurfaces: well-posedness and mean-field limit. arXiv e-prints. 2020;2001.11994.
48. Fornasier M, Huang H, Pareschi L, Sünnen P. Consensus-based optimization on hypersurfaces: well-posedness and mean-field limit. *Math Models Methods Appl Sci*. 2020;30(14):2725–2751.
49. Fornasier M, Klock T, Riedl K. Consensus-based optimization methods converge globally in mean-field law. arXiv e-prints. 2021;2103.15130.
50. Borovykh A, Kantas N, Parpas P, Pavliotis G. Stochastic mirror descent for fast distributed optimization and federated learning. In: *OPT2020: 12th Annual Workshop on Optimization for Machine Learning*; 2020.
51. Borovykh A, Kantas N, Parpas P, Pavliotis GA. To interact or not? The convergence properties of interacting stochastic mirror descent. In: *International Conference on Machine Learning (ICML) Workshop on 'Beyond First order methods in ML Systems*; 2020.
52. Borovykh A, Kantas N, Parpas P, Pavliotis GA. On stochastic mirror descent with interacting particles: convergence properties and variance reduction. *Physica D: Nonlin Phenomena*. 2021;418:132844.
53. Kantas N, Parpas P, Pavliotis GA. The sharp, the flat and the shallow: Can weakly interacting agents learn to escape bad minima? arXiv preprint arXiv:190504121. 2019;.
54. Goodman J, Weare J. Ensemble samplers with affine invariance. *Commun Appl Math Comput Sci*. 2010;5(1):65–80.
55. Lu Y, Lu J, Nolen J. Accelerating Langevin sampling with birth-death. arXiv e-prints. 2019;1905.09863.
56. Reich S, Weissmann S. Fokker–Planck particle systems for Bayesian inference: computational approaches. *SIAM/ASA J Uncertain Quantif*. 2021;9(2):446–482.
57. Kovachki NB, Stuart AM. Ensemble Kalman inversion: a derivative-free technique for machine learning tasks. *Inverse Probl*. 2019;35(9):095005, 35.
58. Pidstrigach J, Reich S. Affine-invariant ensemble transform methods for logistic regression. arXiv preprint arXiv:210408061. 2021.
59. Duncan AB, Stuart AM, Wolfram MT. Ensemble inference methods for models with noisy and expensive likelihoods. arXiv preprint arXiv:210403384. 2021;.
60. Pavliotis GA, Stuart AM, Vaes U. Derivative-free Bayesian inversion using multiscale dynamics. arXiv e-prints. 2021 Feb;2102.00540.
61. Cleary E, Garbuno-Inigo A, Lan S, Schneider T, Stuart AM. Calibrate, emulate, sample. *J Comput Phys*. 2021;424:109716, 20.
62. Leimkuhler B, Matthews C, Weare J. Ensemble preconditioning for Markov chain Monte Carlo simulation. *Statistics Comput*. 2018;28(2):277–290.
63. Petersen KB, Pedersen MS. *The Matrix Cookbook*. Technical University of Denmark; 2008. Version 20081110.
64. Iglesias MA, Law KJH, Stuart AM. Ensemble Kalman methods for inverse problems. *Inverse Probl*. 2013;29(4):045001, 20.
65. Lelièvre T, Stoltz G. Partial differential equations and stochastic methods in molecular dynamics. *Acta Numer*. 2016;25:681–880.
66. Carrillo JA, Hoffmann F, Stuart AM, Vaes U. Consensus based sampling: figshare media; 2021.
67. Ernst OG, Sprungk B, Starkloff H. Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems. *SIAM/ASA J Uncertain Quantif*. 2015;3(1):823–851.
68. Herty M, Visconti G. Kinetic methods for inverse problems. *Kinet Relat Models*. 2019;12(5):1109–1130.
69. Pavliotis GA. *Stochastic Processes and Applications. Vol. 60 of Texts in Applied Mathematics*. New York: Springer; 2014. [Diffusion processes, the Fokker-Planck and Langevin equations.]

70. Nüsken N, Reich S. Note on interacting Langevin diffusions: gradient structure and ensemble Kalman sampler by Garbuno-Inigo, Hoffmann, Li and Stuart. arXiv e-prints. 2019;1908.10890.
71. Bhatia R. *Matrix Analysis. Vol. 169 of Graduate Texts in Mathematics*. New York: Springer-Verlag; 1997.
72. Miller PD. *Applied Asymptotic Analysis. Vol. 75 of Graduate Studies in Mathematics*. Providence, RI: American Mathematical Society; 2006.

How to cite this article: Carrillo JA, Hoffmann F, Stuart AM, Vaes U. Consensus-based sampling. *Stud Appl Math*. 2022;148:1069–1140. https://doi.org/10.1111/sapm.12470

APPENDIX A: AUXILIARY TECHNICAL RESULTS

Lemma A.1. *Let (u_n, v_n) denote the solution to the recurrence relation (26)*

$$u_{n+1} = \left[\alpha + (1-\alpha)(1+v_n)^{-1}\right] u_n, \tag{A.1a}$$

$$v_{n+1} = \left[\alpha^2 + (1-\alpha^2)\lambda^{-1}(1+v_n)^{-1}\right] v_n, \tag{A.1b}$$

with initial condition (u_0, v_0) and $v_0 > 0$. Denote $v_\infty = (1-\lambda)/\lambda$. We separate the sampling and optimization cases.

(i) Case $\lambda \in (0,1)$. It holds, for all $n \in \mathbf{N}$, that

$$\min\left(1, \frac{v_\infty}{v_0}\right)^{\frac{1}{1+\alpha}} \left((1-\alpha)\lambda + \alpha\right)^n \leqslant \left|\frac{u_n}{u_0}\right| \leqslant \max\left(1, \frac{v_\infty}{v_0}\right)^{\frac{1}{1+\alpha}} \left((1-\alpha)\lambda + \alpha\right)^n, \tag{A.2a}$$

$$\min\left(1, \frac{v_\infty}{v_0}\right)\left((1-\alpha^2)\lambda + \alpha^2\right)^n \leqslant \left|\frac{v_n - v_\infty}{v_0 - v_\infty}\right| \leqslant \max\left(1, \frac{v_\infty}{v_0}\right)\left((1-\alpha^2)\lambda + \alpha^2\right)^n; \tag{A.2b}$$

(ii) Case $\lambda = 1$. For all $n \in \mathbf{N}$, it holds that

$$\left(\frac{1}{1 + v_0(1-\alpha^2)n}\right)^{\frac{1}{1+\alpha}} \leqslant \left|\frac{u_n}{u_0}\right| \leqslant \left(\frac{1+v_0}{1 + v_0 + v_0(1-\alpha^2)n}\right)^{\frac{1}{1+\alpha}} \tag{A.3a}$$

and

$$\left(\frac{1}{1 + v_0(1-\alpha^2)n}\right) \leqslant \frac{v_n}{v_0} \leqslant \left(\frac{1+v_0}{1 + v_0 + v_0(1-\alpha^2)n}\right). \tag{A.3b}$$

Proof. **Case $\lambda \in (0,1)$.** Rearranging the equation for $\{v_n\}_{n=0,\dots}$, we obtain

$$v_{n+1} - v_\infty = \gamma(v_n)(v_n - v_\infty), \qquad \gamma(s) := \frac{1+\alpha^2 s}{1+s}. \tag{A.4}$$

If $v_0 \geqslant v_\infty$, then clearly $v_0 \geqslant v_n \geqslant v_\infty$ for all $n \in \mathbf{N}$. Therefore, because $0 < \gamma(\cdot) < 1$, it holds that $0 \leqslant v_{n+1} - v_\infty \leqslant \gamma(v_\infty)(v_n - v_\infty)$, which leads directly to the convergence estimate

$$|v_n - v_\infty| \leqslant \gamma(v_\infty)^n |v_0 - v_\infty|.$$

Similarly, for $v_0 < v_\infty$, we obtain $v_0 < v_n < v_\infty$ for all $n \in \mathbf{N}$, which leads to the lower bound $|v_n - v_\infty| \geqslant \gamma(v_\infty)^n |v_0 - v_\infty|$. For the opposite bounds, we calculate using (A.4) that

$$\begin{aligned}\frac{v_{n+1}^{-1}(v_{n+1} - v_\infty)}{v_n^{-1}(v_n - v_\infty)} &= \left(\frac{1+\alpha^2 v_n}{1+v_n}\right)\frac{v_n}{v_\infty + \frac{1+\alpha^2 v_n}{1+v_n}(v_n - v_\infty)} \\ &= \frac{1+\alpha^2 v_n}{(1-\alpha^2)v_\infty + 1 + \alpha^2 v_n} = \frac{1+\alpha^2 v_\infty + \alpha^2(v_n - v_\infty)}{1 + v_\infty + \alpha^2(v_n - v_\infty)}\,.\end{aligned}$$

Hence, if $v_0 > v_\infty$, then

$$\frac{v_{n+1}^{-1}(v_{n+1} - v_\infty)}{v_n^{-1}(v_n - v_\infty)} \geqslant \gamma(v_\infty) \quad \Longrightarrow \quad \frac{|v_{n+1} - v_\infty|}{v_{n+1}} \geqslant \gamma(v_\infty)\frac{|v_n - v_\infty|}{v_n}\,,$$

with the inequalities reversed in the case $v_0 < v_\infty$. Iterating the last inequality, combining the above estimates and noting that $\gamma(v_\infty) = (1-\alpha^2)\lambda + \alpha^2$ gives (A.2b).

Next, notice that the equation for u_n can be rewritten as

$$u_{n+1} = \tilde{\gamma}(v_n)u_n\,, \qquad \tilde{\gamma}(s) := \frac{1+\alpha s}{1+s}, \tag{A.5}$$

where $\tilde{\gamma}$ is strictly decreasing on $[0, \infty)$. Clearly, if $v_0 \geqslant v_\infty$, then we have

$$|u_n| \leqslant \tilde{\gamma}(v_\infty)^n |u_0|\,, \tag{A.6}$$

with the reversed inequality holding for $v_0 < v_\infty$. Noting that u_n and $v_n - v_\infty$ do not change sign with n, we calculate by analogy with the continuous-time case Lemma A.2 that

$$\begin{aligned}\frac{|u_{n+1}|}{|u_n|}\left(\frac{|v_n - v_\infty|}{|v_{n+1} - v_\infty|}\right)^{\frac{1}{1+\alpha}} &= \frac{u_{n+1}}{u_n}\left(\frac{v_n - v_\infty}{v_{n+1} - v_\infty}\right)^{\frac{1}{1+\alpha}} \\ &= \left(\frac{1+\alpha v_n}{1+v_n}\right)\left(\frac{1+v_n}{1+\alpha^2 v_n}\right)^{\frac{1}{1+\alpha}} =: h_\alpha(v_n).\end{aligned} \tag{A.7}$$

Because $h_\alpha'(s) \geqslant 0$ for all $\alpha \in (0,1)$ and all $s > 0$, we deduce for $v_0 < v_\infty$,

$$\frac{|u_{n+1}|}{|u_n|}\left(\frac{|v_n - v_\infty|}{|v_{n+1} - v_\infty|}\right)^{\frac{1}{1+\alpha}} \leqslant h_\alpha(v_\infty),$$

and iterating this inequality, then using (A.2b), we have

$$|u_n| \leqslant |u_0| h_\alpha(v_\infty)^n \left(\frac{|v_n - v_\infty|}{|v_0 - v_\infty|} \right)^{\frac{1}{1+\alpha}} \leqslant |u_0| \left(\frac{v_\infty}{v_0} \right)^{\frac{1}{1+\alpha}} \left(h_\alpha(v_\infty) \gamma(v_\infty)^{\frac{1}{1+\alpha}} \right)^n$$
$$= |u_0| \left(\frac{v_\infty}{v_0} \right)^{\frac{1}{1+\alpha}} \widetilde{\gamma}(v_\infty)^n \,.$$

with reversed inequality of $v_0 > v_\infty$. Because $\widetilde{\gamma}(v_\infty) = (1-\alpha)\lambda + \alpha$, this concludes the proof of (A.2a).

Case $\lambda = 1$. Rearranging the equation for v_n, we have

$$v_{n+1}^{-1} = \left(\frac{1 + v_n}{1 + \alpha^2 v_n} \right) v_n^{-1} = \gamma(v_n)^{-1} v_n^{-1} \,. \tag{A.8}$$

Because clearly

$$\forall (x, y) \in \mathbf{R}_+^2, \qquad \frac{1+x}{1+y} \leqslant 1 + |x - y|, \tag{A.9}$$

we have

$$v_{n+1}^{-1} \leqslant \big(1 + (1-\alpha^2) v_n\big) v_n^{-1} \leqslant v_n^{-1} + (1 - \alpha^2),$$

so we obtain a lower bound on v_n:

$$\forall n \in \mathbf{N}, \qquad v_n^{-1} \leqslant v_0^{-1} + (1-\alpha^2) n =: \underline{v}_n^{-1}. \tag{A.10}$$

To obtain an upper bound for v_n, we note that

$$v_{n+1} = \left(\frac{1 + \alpha^2 v_n}{1 + v_n} \right) v_n \leqslant \left(\frac{1 + \alpha^2 \underline{v}_n}{1 + \underline{v}_n} \right) v_n = \left(\frac{v_0^{-1} + n(1-\alpha^2) + \alpha^2}{v_0^{-1} + n(1-\alpha^2) + 1} \right) v_n.$$

Therefore, we deduce

$$v_n \leqslant \prod_{k=0}^{n-1} \left(1 - \frac{1 - \alpha^2}{v_0^{-1} + k(1-\alpha^2) + 1} \right) v_0 =: \Pi_{n-1} v_0.$$

Using $\log(1-\epsilon) \leqslant -\epsilon$ for all $\epsilon \in (0, 1)$, we have

$$\log \Pi_{n-1} \leqslant -\sum_{k=0}^{n-1} \frac{1-\alpha^2}{v_0^{-1} + k(1-\alpha^2) + 1} \leqslant -\int_0^n \frac{1-\alpha^2}{v_0^{-1} + x(1-\alpha^2) + 1} \,\mathrm{d}x$$
$$= -\log \left(\frac{v_0^{-1} + n(1-\alpha^2) + 1}{v_0^{-1} + 1} \right),$$

so we conclude that the upper bound in (A.3b) holds. A similar reasoning with the inequality

$$|u_{n+1}| \leqslant \left(\frac{1+\alpha \underline{v}_n}{1+\underline{v}_n} \right) |u_n|$$

can be employed to show the upper bound on u_n in (A.3a). To obtain the lower bound on u_n, we use the fact that h_α is increasing to estimate from (A.7) that

$$\left| \frac{u_{n+1}}{u_n} \right| \left| \frac{v_n}{v_{n+1}} \right|^{\frac{1}{1+\alpha}} = h_\alpha(v_n) \geq h_\alpha(0) = 1 \quad \Leftrightarrow \quad \frac{|u_{n+1}|}{v_{n+1}^{\frac{1}{1+\alpha}}} \geq \frac{|u_n|}{v_n^{\frac{1}{1+\alpha}}}.$$

By iterating this inequality and using (A.10), we conclude

$$\left| \frac{u_n}{u_0} \right| \geq \left| \frac{v_n}{v_0} \right|^{\frac{1}{1+\alpha}} \geq \left(\frac{1}{1+v_0(1-\alpha^2)n} \right)^{\frac{1}{1+\alpha}},$$

which is the result. ■

Lemma A.2. *Let $\lambda \in (0,1]$, and let $(u(t), v(t))$ denote the unique global solution to the ODE system (46) with initial condition (u_0, v_0) and $v_0 > 0$.*

(i) Case $\lambda \in (0,1)$. It holds that

$$\min\left(1, \left(\frac{v_\infty}{v_0}\right)^{\lambda/2}\right) \mathrm{e}^{-(1-\lambda)t} \leqslant \left| \frac{u(t)}{u_0} \right| \leqslant \max\left(1, \left(\frac{v_\infty}{v_0}\right)^{\lambda/2}\right) \mathrm{e}^{-(1-\lambda)t}, \tag{A.11a}$$

$$\min\left(1, \left(\frac{v_\infty}{v_0}\right)^{\lambda}\right) \mathrm{e}^{-2(1-\lambda)t} \leqslant \left| \frac{v(t)-v_\infty}{v_0-v_\infty} \right| \leqslant \max\left(1, \left(\frac{v_\infty}{v_0}\right)^{\lambda}\right) \mathrm{e}^{-2(1-\lambda)t}. \tag{A.11b}$$

(ii) Case $\lambda = 1$. For all $t \geqslant 0$, it holds that

$$\left(\frac{1}{1+2v_0 t} \right)^{\frac{1}{2}} \leqslant \left| \frac{u(t)}{u_0} \right| \leqslant \left(\frac{1+v_0}{1+v_0+2v_0 t} \right)^{\frac{1}{2}} \tag{A.12a}$$

and

$$\frac{1}{1+2v_0 t} \leqslant \frac{v(t)}{v_0} \leqslant \frac{1+v_0}{1+v_0+2v_0 t}. \tag{A.12b}$$

Proof. Note that solutions to (46) are unique, and exist globally in time.

Case $\lambda \in (0,1)$. We begin with the sampling case (i) when $\lambda \neq 1$, The second equation in (46) can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t}(v - v_\infty) = -2\Big(\frac{v}{v+1}\Big)(v - v_\infty).$$

For $x = v/v_\infty$, we obtain

$$\dot{x} = -2\left(\frac{(x-1)x}{v_\infty^{-1} + x}\right) = -2\left(\frac{1 + v_\infty^{-1}}{x - 1} - \frac{v_\infty^{-1}}{x}\right)^{-1} = -2(1-\lambda)\left(\frac{-1}{1-x} - \frac{\lambda}{x}\right)^{-1}.$$

We can rewrite this equation as

$$\frac{\mathrm{d}}{\mathrm{d}t}(\log(1 - x(t)) - \lambda \log(x(t))) = -2(1-\lambda), \tag{A.13}$$

leading to

$$|1 - x(t)| = \left(\frac{x(t)}{x(0)}\right)^{\lambda} \mathrm{e}^{-2(1-\lambda)t} \, |1 - x(0)|.$$

Because $v(t)$ is decreasing if $v_0 > v_\infty$ and increasing if $v_0 < v_\infty$, estimate (A.11b) directly follows.

Next, we consider the first equation in (46), and note that it can be rewritten as

$$\frac{\dot{u}}{u} = \frac{1}{2}\left(\frac{1}{v - v_\infty}\right)\frac{\mathrm{d}}{\mathrm{d}t}(v - v_\infty).$$

This implies

$$\log\left(\frac{u(t)}{u_0}\right) = \frac{1}{2}\log\left(\frac{v(t) - v_\infty}{v_0 - v_\infty}\right), \tag{A.14}$$

where it is not difficult to verify that the arguments of the logarithms are positive for all times. Applying (A.11b), we conclude that (A.11a) holds.

Case $\lambda = 1$. The argument follows analogously; the second equation in (46) reads

$$\dot{v} = -2\left(\frac{v}{v+1}\right)v, \tag{A.15}$$

Because the right hand is bounded from below by $-2v^2$, we directly deduce that

$$\forall t > 0, \qquad v(t) \geqslant \frac{1}{v_0^{-1} + 2t} := \underline{v}(t). \tag{A.16}$$

Now, because the function $s \mapsto \frac{s}{1+s}$ is increasing, it is clear that $v(t)$ satisfies

$$\dot{v}(t) \leqslant -2\left(\frac{\underline{v}(t)}{\underline{v}(t) + 1}\right)v(t).$$

Using Grönwall's inequality, we obtain the upper bound in (A.12b).

The bounds on $u(t)$ are then obtained from (A.14) and the bounds on $v(t)$. ■

Remark A.1. Notice that, by letting $\alpha = \mathrm{e}^{-t/n}$ in the bounds obtained in Lemma A.1 and taking the limit $n \to \infty$, we recover the bounds in Lemma A.2.

Remark A.2. It is possible to slightly improve the upper bounds in (A.3b) and (A.12b).

- In the discrete-time case, rearranging the equation for v_{n+1} and using that $\log(1+\varepsilon) \geqslant \frac{\varepsilon}{1+\varepsilon}$ for all $\varepsilon > 0$, we have

$$\begin{aligned} &v_{n+1}^{-1} - \log(v_{n+1}) - v_n^{-1} + \log(v_n) \\ &\quad = \frac{1-\alpha^2}{1+\alpha^2 v_n} + \alpha^2 \log\left(\frac{1+v_n}{1+\alpha^2 v_n}\right) = \frac{1-\alpha^2}{1+\alpha^2 v_n} + \log\left(1 + \frac{(1-\alpha^2)v_n}{1+\alpha^2 v_n}\right) \\ &\quad \geqslant \frac{1-\alpha^2}{1+\alpha^2 v_n} + \frac{(1-\alpha^2)v_n}{1+v_n} \geqslant (1-\alpha^2)\left(\frac{1}{1+\alpha^2 v_n} + \frac{v_n}{1+v_n}\right) \geqslant 1-\alpha^2. \end{aligned} \tag{A.17}$$

  Because v_n is decreasing with n, this directly implies, using the lower bound (A.10),

$$v_n^{-1} \geqslant v_0^{-1} + n(1-\alpha^2) + \log\left(\frac{v_n}{v_0}\right) \geqslant v_0^{-1} + (1-\alpha^2)n - \log\left(1 + v_0(1-\alpha^2)n\right).$$

  so we deduce the following inequality:

$$\frac{v_n}{v_0} \leqslant \frac{1}{1 + v_0(1-\alpha^2)n - v_0 \log\left(1 + v_0(1-\alpha^2)n\right)}$$

  which holds for $n \in \mathbf{N}$ large enough to ensure that the right-hand side is strictly positive.
- In the continuous-time case, one may rewrite (A.15) as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\log v(t) - \frac{1}{v(t)}\right) = -2.$$

  Integrating, rearranging, and taking reciprocals, we obtain

$$v(t) = \frac{1}{v_0^{-1} + 2t + \log\left(\frac{v}{v_0}\right)}.$$

  Using the lower bound (A.16) to bound the argument of the logarithm, we obtain

$$v(t) \leqslant \frac{v_0}{1 + 2v_0 t - v_0 \log(1 + 2v_0 t)}.$$

Though slightly better in the long time limit, these bounds are more cumbersome to manipulate than the ones presented in Lemmas A.1 and A.2.

Lemma A.3. *Assume that μ is a probability measure on $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$, with $\mathcal{B}(\mathbf{R})$ the Borel σ-algebra on $\mathbf{R}$, and that $f : \mathbf{R} \to \mathbf{R}$ is a positive and nondecreasing (resp. nonincreasing) function. Let $\widetilde{\mu}$ be the probability measure defined by*

$$\widetilde{\mu} : \mathcal{B}(\mathbf{R}) \ni A \mapsto \frac{\int_A f(x) \mathrm{d}\mu(x)}{\int_{\mathbf{R}} f(x)\, \mathrm{d}\mu(x)}.$$

Then it holds $\mathbf{E}_{X\sim\widetilde{\mu}}(X) \geqslant \mathbf{E}_{X\sim\mu}(X)$ *(resp.* $\mathbf{E}_{X\sim\widetilde{\mu}}(X) \leqslant \mathbf{E}_{X\sim\mu}(X)$*).*

Proof. Let us assume that f is nondecreasing, and let us denote the cumulative distribution functions (CDFs) by $F(x) := \mathbf{P}_{X\sim\mu}(X \leqslant x)$ and $\widetilde{F}(x) := \mathbf{P}_{X\sim\widetilde{\mu}}(X \leqslant x)$. For any probability measure ν with CDF F_ν, it holds

$$\mathbf{E}_{X\sim\nu}(X) = \int_0^\infty 1 - F_\nu(x) - F_\nu(-x)\,\mathrm{d}x,$$

so it is sufficient to show $\widetilde{F}(x) \leqslant F(x)$ for all $x \in \mathbf{R}$. If $\widetilde{F}(x) = 0$, this inequality is clearly satisfied, so let us verify the inequality for any x such that $\widetilde{F}(x) > 0$. For such a value of x, employing the fact that f is nondecreasing, we obtain

$$\frac{1-\widetilde{F}(x)}{\widetilde{F}(x)} = \frac{\int_{(x,\infty)} f(y)\,\mathrm{d}\mu(y)}{\int_{(-\infty,x]} f(y)\,\mathrm{d}\mu(y)} \geqslant \frac{\int_{(x,\infty)} f(x)\,\mathrm{d}\mu(y)}{\int_{(-\infty,x]} f(x)\,\mathrm{d}\mu(y)} = \frac{\mu((x,\infty))}{\mu((-\infty,x])} = \frac{1-F(x)}{F(x)}.$$

Applying the function $y \mapsto \frac{1}{1+y}$ to both sides of this inequality, and flipping the direction of the inequality accordingly (because this function is decreasing over $[0,\infty)$), we obtain the desired inequality $\widetilde{F}(x) \leqslant F(x)$. ■

Lemma A.4. *Let* $r > 2$ *be given. There exists* $\gamma > 0$ *sufficiently large such that*

$$\forall \widetilde{C} > 0, \qquad h(\widetilde{C};\gamma) := \frac{(1+r\widetilde{C})^{\frac{1}{r}}}{1+\widetilde{C}}\left(1 + 2\frac{\phi\left(\frac{\gamma\widetilde{C}^{\frac{1}{r}}}{\sqrt{\widetilde{C}(1+\widetilde{C})}}\right)}{\frac{\gamma\widetilde{C}^{\frac{1}{r}}}{\sqrt{\widetilde{C}(1+\widetilde{C})}}}\right) \leqslant 1,$$

where ϕ *denotes the density of the standard normal distribution, i.e.,* $\phi = g(\cdot\,;0,1)$.

Proof. If $\widetilde{C} \geqslant 1$, then

$$h(\widetilde{C},\gamma) \leqslant \frac{(1+r\widetilde{C})^{\frac{1}{r}}}{1+\widetilde{C}}\left(1 + 2\frac{\phi(0)}{\frac{\gamma\widetilde{C}^{\frac{1}{r}}}{\sqrt{\widetilde{C}(1+\widetilde{C})}}}\right) \leqslant \frac{(1+r\widetilde{C})^{\frac{1}{r}}}{1+\widetilde{C}} + \frac{2\phi(0)}{\gamma}\frac{(1+r\widetilde{C})^{\frac{1}{r}}}{\widetilde{C}^{\frac{1}{r}}}\sqrt{\frac{\widetilde{C}}{1+\widetilde{C}}}.$$

By concavity of $\widetilde{C} \mapsto (1+r\widetilde{C})^{\frac{1}{r}}$, and the fact that the first term is strictly decreasing, we have

$$h(\widetilde{C},\gamma) \leqslant \frac{(1+r)^{\frac{1}{r}}}{2} + \frac{2\phi(0)}{\gamma}\left(\frac{1+(r\widetilde{C})^{\frac{1}{r}}}{\widetilde{C}^{\frac{1}{r}}}\right) \leqslant \frac{(1+r)^{\frac{1}{r}}}{2} + \frac{2\phi(0)}{\gamma}\left(1 + r^{\frac{1}{r}}\right).$$

Because the first term is strictly less than 1, there exists γ sufficiently large such that the right-hand side is bounded from above by 1.

If $0 < \widetilde{C} < 1$, on the other hand, we have

$$h(\widetilde{C},\gamma) \leqslant \frac{(1+r\widetilde{C})^{\frac{1}{r}}}{1+\widetilde{C}}\left(1+\frac{4}{\gamma}\phi\left(\frac{\gamma\widetilde{C}^{\frac{1}{r}}}{\sqrt{2\widetilde{C}}}\right)\right).$$

Therefore,

$$\log\left(h(\widetilde{C},\gamma)\right) \leqslant \frac{1}{r}\log(1+r\widetilde{C}) - \log(1+\widetilde{C}) + \log\left(1+\frac{4}{\gamma}\phi\left(\frac{\gamma\widetilde{C}^{\frac{1}{r}}}{\sqrt{2\widetilde{C}}}\right)\right).$$

The sum of the first two terms is bounded as follows (where we employ that $\widetilde{C} \leqslant 1$):

$$\begin{aligned}\frac{1}{r}\log(1+r\widetilde{C}) - \log(1+\widetilde{C}) &= \int_0^{\widetilde{C}} \left(\frac{1}{1+rx} - \frac{1}{1+x}\right) \mathrm{d}x \\ &\leqslant -(r-1)\int_0^{\widetilde{C}} \frac{x}{2(1+r)}\,\mathrm{d}x = -\frac{1}{4}\left(\frac{r-1}{r+1}\right)\widetilde{C}^2.\end{aligned}$$

Employing this estimate together with the elementary bound $\log(1+\varepsilon) \leqslant \varepsilon$, we have

$$\log\left(h(\widetilde{C},\gamma)\right) \leqslant -\frac{1}{4}\left(\frac{r-1}{r+1}\right)\widetilde{C}^2 + \frac{4}{\gamma}\phi\left(\frac{\gamma}{\sqrt{2}}\widetilde{C}^{-\frac{r-2}{2r}}\right).$$

Clearly, there exists K such that $\phi(x) \leqslant K(1+x)^{-\frac{4r}{r-2}}$ uniformly, so we deduce

$$\log\left(h(\widetilde{C},\gamma)\right) \leqslant -\frac{1}{4}\left(\frac{r-1}{r+1}\right)\widetilde{C}^2 + \frac{4K}{\gamma}\left(\frac{\sqrt{2}}{\gamma}\right)^{\frac{4r}{r-2}}\widetilde{C}^2.$$

It is possible to choose γ sufficiently large such that the right-hand side of this equation is bounded from above by 0 for $\widetilde{C} \in (0,1]$, and the statement then follows easily. ■

Lemma A.5. *Assume that* $\alpha \in [0,1]$ *and that* $\widehat{C}_\beta$, $\widehat{C}_n$, $\widehat{m}_\beta$, *and* $\widehat{u}$ *are nonnegative real numbers satisfying* $0 < \widehat{C}_\beta \leqslant \widehat{C}_n$ *and*

$$\frac{\widehat{m}_\beta}{\widehat{C}_\beta^{1/r}} \leqslant \frac{\widehat{u}}{\widehat{C}_n^{1/r}}$$

for some $r \geqslant 2$. *Then* $(\widehat{m}_{n+1}, \widehat{C}_{n+1})$ *defined by*

$$\begin{aligned}\widehat{m}_{n+1} &= (1-\alpha)\widehat{m}_\beta + \alpha\widehat{u}, \\ \widehat{C}_{n+1} &= (1-\alpha^2)\widehat{C}_\beta + \alpha^2\widehat{C}_n\end{aligned}$$

satisfy

$$\frac{\widehat{m}_{n+1}}{\widehat{C}_{n+1}^{1/2r}} \leqslant \frac{\widehat{u}}{\widehat{C}_{n}^{1/2r}}.$$

Proof. Letting $m_{n+1} = \widehat{m}_{n+1}/\widehat{u}$, $C_{n+1} = \widehat{C}_{n+1}/\widehat{C}_n$, $m_\beta = \widehat{m}_\beta/\widehat{u}$, and $C_\beta = \widehat{C}_\beta/\widehat{C}_n$, we can rewrite the equations for $\widehat{m}_{n+1}$ and $\widehat{C}_{n+1}$ as

$$m_{n+1} = (1-\alpha)m_\beta + \alpha,$$
$$C_{n+1} = (1-\alpha^2)C_\beta + \alpha^2.$$

By the assumptions, it holds that $C_\beta \leqslant 1$ and $m_\beta \leqslant C_\beta^{1/r}$, and so

$$\frac{m_{n+1}^{2r}}{C_{n+1}} = \frac{\big((1-\alpha)m_\beta + \alpha\big)^{2r}}{(1-\alpha^2)C_\beta + \alpha^2} \leqslant \frac{((1-\alpha)x + \alpha)^{2r}}{(1-\alpha^2)x^r + \alpha^2} =: h(x,\alpha), \qquad x := C_\beta^{1/r} \in (0,1].$$

We claim that

$$\forall (y,\alpha) \in (0,1] \times [0,1), \qquad \partial_x h(y,\alpha) \geqslant 0. \tag{A.18}$$

This will imply that $h(x,\alpha) = h(1,\alpha) - \int_x^1 \partial_x h(y,\alpha)\,\mathrm{d}y \leqslant h(1,\alpha) = 1$ and thus $m_{n+1}^{2r} \leqslant C_{n+1}$, giving the statement. Let us now prove (A.18). A simple calculation gives

$$\begin{aligned}\operatorname{sign}(\partial_x h(y,\alpha)) &= \operatorname{sign}\big(2r(1-\alpha)\big((1-\alpha^2)y^r + \alpha^2\big) - r(1-\alpha^2)y^{r-1}((1-\alpha)y+\alpha)\big)\\ &= \operatorname{sign}\big(2\big((1-\alpha^2)y^r + \alpha^2\big) - (1+\alpha)y^{r-1}((1-\alpha)y+\alpha)\big)\\ &= \operatorname{sign}\big(\alpha^2\big(2 - y^r - y^{r-1}\big) - \alpha y^{r-1} + y^r\big) =: \operatorname{sign}(g(y,\alpha)).\end{aligned}$$

The argument of the sign function in the last line, i.e., $g(y,\alpha)$, is a quadratic function of α with a minimizer at $\alpha_*(y) = \frac{1}{2}y^{r-1}(2-y^r-y^{r-1})^{-1}$. If $\alpha_*(y) \geqslant 1$, then $g(y,\alpha) \geqslant g(y,1) \geqslant 0$. On the other hand, for any y such that $\alpha_*(y) \leqslant 1$, it holds

$$\forall \alpha \in [0,1], \qquad g(y,\alpha) \geqslant g(y,\alpha_*) = y^r\left(1 - \frac{1}{2y}\left(\frac{\frac{1}{2}y^{r-1}}{2-y^r-y^{r-1}}\right)\right).$$

If $y \in (0,\frac{1}{2}]$, a direct bound of the right-hand side of the previous equation shows that $g(y,\alpha_*) \geqslant 0$, and if $y \geqslant 1/2$ we have by the constraint $\alpha_*(y) \leqslant 1$ that

$$g(y,\alpha) \geqslant g(y,\alpha_*) \geqslant y^r\left(1 - \frac{1}{2y}\right) \geqslant 0,$$

which concludes the proof of (A.18). ■

Lemma A.6 (Generalization of Watson's lemma with bound on remainder)**.** *Assume that ϕ is a smooth function satisfying*

$$M := \| \mathrm{e}^{-\beta_0 \theta^2} \, \phi^{(2N+2)}(\theta) \|_\infty < \infty. \tag{A.19}$$

for some constant $\beta_0 \in \mathbf{R}$ and $N \in \mathbf{N}$. Then for $\beta > \beta_0$ it holds

$$I_\beta := \int_{-\infty}^{\infty} \mathrm{e}^{-\beta\theta^2} \, \phi(\theta) \, \mathrm{d}\theta = \sum_{n=0}^{N} \phi_{2n} \, \frac{\Gamma(n+1/2)}{\beta^{n+1/2}} + R_\beta, \qquad \phi_{2n} := \frac{\phi^{(2n)}(0)}{(2n)!},$$

where the remainder R_β satisfies the bound

$$|R_\beta| \leqslant \frac{M}{(2N+2)!} \, \frac{\Gamma(N+3/2)}{(\beta-\beta_0)^{N+3/2}}.$$

Proof. We follow here the approach of Ref. [72, Chapter 2]. We first notice that

$$I_\beta = 2 \int_0^\infty \mathrm{e}^{-\beta\theta^2} \left(\frac{\phi(\theta) + \phi(-\theta)}{2} \right) \mathrm{d}\theta =: 2 \int_0^\infty \mathrm{e}^{-\beta\theta^2} \, \psi(\theta) \, \mathrm{d}\theta \, .$$

The function ψ is even and smooth, all its odd derivatives vanish at $\theta = 0$. Therefore, by Taylor's theorem, for any $\theta \geqslant 0$ there exists $\xi(\theta) \in [0, \theta]$ such that

$$\psi(\theta) = \sum_{n=0}^{N} \phi_{2n} \, \theta^{2n} + \frac{\psi^{(2N+2)}(\xi(\theta))}{(2N+2)!} \, \theta^{2N+2}.$$

With a change of variables $\sigma = \theta^2$, this leads to

$$I_\beta = \sum_{n=0}^{N} \phi_{2n} \int_0^\infty \mathrm{e}^{-\beta\sigma} \, \sigma^{n-1/2} \, \mathrm{d}\sigma + R_\beta = \sum_{n=0}^{N} \phi_{2n} \, \frac{\Gamma(n+1/2)}{\beta^{n+1/2}} + R_\beta,$$

where, by (A.19) and for $\beta > \lambda_0$, the remainder term is bounded from above as follows:

$$|R_\beta| \leqslant \frac{M}{(2N+2)!} \int_0^\infty \mathrm{e}^{-(\beta-\beta_0)\sigma} \, \sigma^{N+1/2} \, \mathrm{d}\sigma = \frac{M}{(2N+2)!} \, \frac{\Gamma(N+3/2)}{(\beta-\beta_0)^{N+3/2}},$$

which concludes the proof. ■

Lemma A.7. *Suppose that Assumptions 1 and 4 are satisfied. Then there exists a unique smooth and increasing function $\tau(\theta)$ such that*

$$\forall \theta \in \mathbf{R}, \qquad f(\theta_* + \tau(\theta)) = f(\theta_*) + \theta^2.$$

In addition, the function τ and all its derivatives are bounded from above by the reciprocal of a Gaussian, in the sense that for all $i \in \{0, 1, 2, \dots\}$ there exists $\mu_i \in \mathbf{R}$ such that

$$\|e^{-\mu_i \theta^2} \tau^{(i)}(\theta)\|_\infty < \infty.$$

Proof. Introducing $g(\theta) := f(\theta + \theta_*) - f(\theta_*)$, we must prove the existence of a function τ satisfying

$$\forall \theta \in \mathbf{R}, \qquad g(\tau(\theta)) = \theta^2. \tag{A.20}$$

By assumption $g''(\theta) \geqslant \ell$, so $g(\theta) \geqslant \ell\, \theta^2/2$ and $|g'(\theta)| \geqslant \ell |\theta|$ for all $\theta \in \mathbf{R}$. This implies that the preimage set $g^{-1}(\theta^2)$ contains exactly two elements for any value of $\theta \neq 0$, a positive one $g_+^{-1}(\theta^2)$ and a negative one $g_-^{-1}(\theta^2)$. Further, the preimage $g^{-1}(0)$ is simply $\{0\}$. If τ satisfies (A.20) and is increasing, then it holds necessarily that

$$\tau(\theta) = \begin{cases} g_-^{-1}(\theta^2) & \text{if } \theta < 0, \\ 0 & \text{if } \theta = 0, \\ g_+^{-1}(\theta^2) & \text{if } \theta > 0. \end{cases}$$

By the inverse function theorem, we observe that g_+^{-1} and g_-^{-1} are smooth on $(0, +\infty)$, because g is smooth and strictly monotonic over $(-\infty, 0)$ and $(0, \infty)$, and consequently τ is smooth on $(-\infty, 0)$ and $(0, \infty)$. Therefore, to show that τ is a smooth function over $\mathbf{R}$, it is sufficient to verify that τ is also infinitely differentiable in a neighborhood of $\theta = 0$. To this end, we define, analogously to Ref. [72, Chapter 3],

$$G(u, \theta) = \begin{cases} \frac{g(u\theta)}{\theta^2} - 1 & \text{if } \theta \neq 0, \\ \frac{u^2}{2} g''(0) - 1 & \text{if } \theta = 0. \end{cases}$$

The function G is smooth over $\mathbf{R}^2$ and it is simple to verify that $G(u^*, 0) = 0$ for $u^* = \sqrt{2/g''(0)}$ and $\partial_u G(u^*, 0) = u^* g''(0) > 0$. Therefore, the implicit function theorem implies the existence of a unique smooth function $\hat{u}(\theta)$, defined on an interval $(-\varepsilon, \varepsilon)$, such that $\hat{u}(0) = u^*$ and $G(\hat{u}(\theta), \theta) = 0$ for any $\theta \in (-\varepsilon, \varepsilon)$. Because the function $\hat{\tau} : (-\varepsilon, \varepsilon) \ni \theta \mapsto \hat{u}(\theta)\theta$ satisfies $g(\hat{\tau}(\theta)) = \theta^2$ by construction, and because it is increasing for ε sufficiently small because $\hat{u}(0) > 0$, this function must necessarily coincide with τ on the interval $(-\varepsilon, \varepsilon)$, implying that τ is indeed smooth over $\mathbf{R}$.

Now note that, because the function f and its derivatives are bounded by the reciprocal of a Gaussian by assumption, then clearly so are the function g and its derivatives; for any $i \in \{0, 1, 2, \dots\}$, there exists r_i such that

$$\|e^{-r_i \theta^2} g^{(i)}(\theta)\|_\infty < \infty.$$

Differentiating (A.20) repeatedly, we obtain

$$g'(\tau(\theta))\, \tau'(\theta) = 2\theta \tag{A.21a}$$

$$g''(\tau(\theta))\,|\tau'(\theta)|^2 + g'(\tau(\theta))\,\tau''(\theta) = 2, \tag{A.21b}$$

$$p_i\big(g'(\tau(\theta)),\dots,g^{(i)(\tau(\theta))},\tau'(\theta),\dots,\tau^{(i-1)}(\theta)\big) + g'(\tau(\theta))\,\tau^{(i)}(\theta) = 0, \qquad i = 3,\dots \tag{A.21c}$$

where p_i are polynomials. Recalling that $|g'(\theta)| \geqslant \ell\,|\theta|$ for all $\theta \in \mathbf{R}$, we can therefore divide the equations in appendix A.21 by $g'(\tau(\theta))$ to obtain expressions for the derivatives $\tau^{(i)}(\theta)$, which are valid when $\theta \neq 0$. From these expressions, it is then easy to obtain the desired bounds. For example, if we have already shown that $\|\,\mathrm{e}^{-\mu_1\theta^2}\,\tau'\|_\infty < \infty$, which follows from (A.21a), then from (A.21b) we obtain, using the fact that $\theta^2 = g(\tau(\theta)) \geqslant \frac{\ell}{2}|\tau(\theta)|^2$,

$$\begin{aligned}|\tau''(\theta)| &\leqslant \frac{2 + |g''(\tau(\theta))\,|(\tau'(\theta))^2}{|g'(\tau(\theta))|} \leqslant \frac{2 + C\,\mathrm{e}^{r_2|\tau(\theta)|^2}\ \mathrm{e}^{2\mu_1\theta^2}}{\ell\,|\tau(\theta)|}\\ &\leqslant \frac{2 + C\,\mathrm{e}^{\frac{2r_2}{\ell}\theta^2}\ \mathrm{e}^{2\mu_1\theta^2}}{\ell\,|\tau(\theta)|} \leqslant C\,\mathrm{e}^{\left(\frac{2r_2}{\ell}+2\mu_1\right)\theta^2} \qquad \text{if } |\theta| \geqslant 1,\end{aligned}$$

where C is a constant changing from occurrence to occurrence. The last inequality is justified because $\max_{|\theta|\geqslant 1}|\tau(\theta)| > 0$. Because τ'' is continuous and the set $\{\theta\,:\,|\theta| \leqslant 1\}$ is compact, this shows the existence of $\mu_2 \in \mathbf{R}$ that $\|\tau''(\theta)\,\mathrm{e}^{-\mu_2\theta^2}\|_\infty < \infty$. ■