Inverse Problems

PAPER

Well posedness and convergence analysis of the ensemble Kalman inversion

To cite this article: Dirk Blömker *et al* 2019 *Inverse Problems* **35** 085007

IOP Publishing Inverse Problems

Inverse Problems **35** (2019) 085007 (32pp) https://doi.org/10.1088/1361-6420/ab149c

Well posedness and convergence analysis of the ensemble Kalman inversion

Dirk Blömker[1], Claudia Schillings[2], Philipp Wacker[3] and Simon Weissmann[2]

[1] Universität Augsburg, Augsburg, Germany
[2] Universität Mannheim, Mannheim, Germany
[3] Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

E-mail: dirk.bloemker@math.uni-augsburg.de, c.schillings@uni-mannheim.de, phkwacker@gmail.com and sweissma@mail.uni-mannheim.de

Received 19 October 2018, revised 26 February 2019
Accepted for publication 29 March 2019
Published 30 July 2019

Abstract

The ensemble Kalman inversion is widely used in practice to estimate unknown parameters from noisy measurement data. Its low computational costs, straightforward implementation, and non-intrusive nature makes the method appealing in various areas of application. We present a complete analysis of the ensemble Kalman inversion with perturbed observations for a fixed ensemble size when applied to linear inverse problems. The well-posedness and convergence results are based on the continuous time scaling limits of the method. The resulting coupled system of stochastic differential equations allows one to derive estimates on the long-time behaviour and provides insights into the convergence properties of the ensemble Kalman inversion. We view the method as a derivative free optimization method for the least-squares misfit functional, which opens up the perspective to use the method in various areas of applications such as imaging, groundwater flow problems, biological problems as well as in the context of the training of neural networks.

Keywords: Bayesian inverse problems, ensemble Kalman inversion, optimization, well-posedness and accuracy

(Some figures may appear in colour only in the online journal)

1. Introduction

Inverse problems arise in various fields of sciences and engineering. Methods to efficiently incorporate data into models are needed to reduce the overall uncertainty and to ensure the reliability of the simulations under real world conditions. The Bayesian approach to inverse problems provides a rigorous framework for the incorporation and quantification of uncertainties in measurements, parameters and models. However, in computationally intense applications, the approximation of the solution of the Bayesian inverse problem, the posterior, might be prohibitively expensive; see [36]. In such settings, the ensemble Kalman filter (EnKF), originally introduced by Evensen [1] for data assimilation, has been reported to produce reliable estimates of the unknown parameters with low computational cost, making the method very appealing for large scale problems. Areas of applications include, among others, groundwater flow problems [2], climate models [3], biological problems [4], image reconstruction [5] and building [6] and material sciences [7]. Most recent directions involve the use of the ensemble Kalman inversion as a derivative free optimization method, in particular in the context of the training of neural networks [8]. Despite its documented success, the ensemble Kalman inversion is underpinned by limited theoretical understanding. The goal of our work is to give useful insights into properties of the method and provide tools for a systematic development and improvement.

For linear dynamical systems and Gaussian initial conditions, analysis of the large ensemble size limit has been done, for example in [9, 10]. Convergence to the mean-field Kalman filter for nonlinear systems can be found in [11]. Multilevel extensions are proposed, e.g. in [12, 13]. In [14–16], the authors present an analysis of the long-time behaviour and ergodicity of the ensemble Kalman filter with arbitrary ensemble size establishing time uniform bounds to control the filter divergence with variance inflation techniques and ensuring in addition the existence of an invariant measure. Accuracy results have been recently established for a fixed ensemble size in the linear Gaussian setting, see [17, 18] and for ensemble Kalman–Bucy filters applied to continuous-time filtering problems, see [19, 20].

For inverse problems, the large ensemble size limit has been investigated in [21]. It has been shown that the ensemble Kalman inversion (EKI) is not consistent with the Bayesian perspective in the nonlinear setting, but can be interpreted as a point estimator of the unknown parameters. We will adopt this viewpoint throughout the paper and analyze the behavior of the EKI as an optimization method of the least-squares misfit functional. However, to motivate the algorithm, we will shortly introduce the Bayesian setting and derive the ensemble Kalman filter for inverse problems. In [22], it was demonstrated that the continuous time limit of the EKI algorithm is an interacting set of gradient flows, see also [23–25] for the continuous time limit of the EnKF in the data assimilation context. In the discrete setting, the connection to deterministic regularisation techniques is established in [26, 27]. In the following, we will interpret the EKI method as a numerical discrete approximation of a stochastic differential equation, see [28] and show well-posedness and asymptotic behaviour of the stochastic differential equation. Our work will extend the results from [22, 29] to the inversion with perturbed observations. Though both methods, i.e. the limit of the EKI with perturbed observations and the deterministic limit from [22], can be analysed from an optimization perspective, the EKI variant with perturbed observation is shown to be second order accurate, whereas the deterministic limit underestimates the covariance in the linear, Gaussian setting, see e.g. [1]. In addition, in the nonlinear setting, methods that add noise to data are reported to be more robust to assumptions about linearity and normality, see e.g. [30] and the references therein. We therefore believe that the EKI with perturbed observations is a good starting point for methods (also of higher accuracy) in the nonlinear, non-Gaussian setting and that the analysis presented here provides valuable insights for the development of these methods.

Our contribution consists of providing a complete analysis of the ensemble Kalman inversion with perturbed observations for linear forward operators. The presented results hold true for arbitrary prior distributions on the unknown parameters, i.e. no Gaussian assumption is invoked. We want to stress that we analyze the algorithm in practical regimes by focusing on results for a fixed ensemble size. We study the continuous time limit of the ensemble Kalman inversion, which allows one to establish well-posedness and accuracy results by exploiting the underlying structure of the limiting coupled stochastic differential equations for the particles. In particular, we make the following main contributions:

- We prove the existence and uniqueness of solutions of the limiting system of stochastic differential equations, thus well-posedness of the algorithm.
- We quantify the ensemble collapse in the observation as well as in the parameter space. The ensemble collapse is characterized in terms of moments and almost sure convergence with given rate.
- In case of exact data, we establish convergence results to the truth using variance inflation. The convergence is characterized in terms of second moments and almost sure convergence with given rate. Under additional assumptions on the forward operator, the results in the data space can be transferred to the parameter space.
- We provide numerical experiments which illustrate the theoretical results studied in this paper.

We do *not* show strong convergence of the discrete EnKF iteration to continuous paths of the corresponding SDE. This would be interesting and there are preliminary results [28], but this is still ongoing research.

The remainder of the article is structured as follows. At the end of this section, we formulate the inverse problem and the Bayesian approach to it. Section 2 is devoted to the ensemble Kalman inversion with perturbed observations. In section 3 we formulate the continuous time limit of the algorithm, introduce the assumptions on the forward problem and prove the well-posedness of the method, i.e. we show the existence and uniqueness of strong solutions of the limit. Section 4 presents the results on the ensemble collapse, in the data and parameter space. In section 5, we show convergence to the truth using variance inflation techniques. Numerical experiments illustrating the theoretical findings are given in section 6. Finally, in section 7, we conclude with a short summary of the main results and discussion of future work. In appendix A auxiliary results are presented and and appendix B contains the proof on the higher-order ensemble collapse.

Let $\mathcal{G} \in \mathcal{C}(\mathcal{X}, \mathbb{R}^K)$ denote the forward response operator mapping the unknown parameters $u \in \mathcal{X}$ to the data space $\mathbb{R}^K$, where $\mathcal{X}$ is a separable Hilbert space and $K \in \mathbb{N}$ denotes the number of observations. We consider the inverse problem of recovering unknown parameters $u \in \mathcal{X}$ from noisy observation $y \in \mathbb{R}^K$ given by

$$y = \mathcal{G}(u) + \eta,$$

where $\eta \sim \mathcal{N}(0, \Gamma)$ is a Gaussian with mean zero and covariance matrix Γ, which models the noise in the observations and in the model.

Following the Bayesian approach, for fixed $y \in \mathbb{R}^K$ we introduce the least-squares functional $\Phi(\cdot; y) : \mathcal{X} \to \mathbb{R}$ by

$$\Phi(u; y) = \frac{1}{2} |(y - \mathcal{G}(u))|_\Gamma^2 ,$$

with $|\cdot|_\Gamma := |\Gamma^{-\frac{1}{2}} \cdot|$ denoting the weighted Euclidean norm in $\mathbb{R}^K$. The unknown parameter u is modeled as a $\mathcal{X}$-valued random variable with prior distribution μ_0. Thus, the pair (u, y) is a jointly varying random variable on $\mathcal{X} \times \mathbb{R}^K$. We assume for the observational noise that $\eta \sim \mathcal{N}(0, \Gamma)$ is independent of $u \sim \mu_0$.

By Bayes' theorem, the solution to the inverse problem is the $\mathcal{X}$-valued random variable $u \mid y \sim \mu$ where the law μ is given by

$$\mu(du) = \frac{1}{Z} \exp(-\Phi(u;y))\mu_0(du),$$

with the normalization constant Z, where

$$Z := \int_{\mathcal{X}} \exp(-\Phi(u;y))\mu_0(du).$$

Note that evaluation of the posterior requires evaluation of the forward model via $\Phi(u;y)$.

2. The EnKF for inverse problems

The Ensemble Kalman methodology consists of choosing an *ensemble* of 'particles' by drawing from the prior which are then transformed to a new set of particles via a linear Gaussian update. A good idea (see [22, 27] for details) is to do this not in one big leap but in an iteration of steps. This amounts to interpolating the step from prior μ_0 to the posterior μ by choosing an artificial time index n and defining a sequence of measures $\mu_0, \mu_1, \mu_2, \dots, \mu_N$ where μ_0 is the prior and $\mu_N = \mu$ is the posterior, i.e.

$$\mu_{n+1}(du) = \frac{1}{Z_n} \exp\big(-h\Phi(u;y)\big)\mu_n(du), \quad n = 0, \dots, N-1,$$

with $h = N^{-1}$ and $Z_n = \int \exp(-h\Phi(u))\mu_n(du)$. The iterative form of the Ensemble Kalman methodology iterates the initial ensemble of particles through this set of intermediate measures. This is the setting we will constrain ourselves. Note again that any 'time' n or (later) t is entirely artificial (transformation) time and independent of any physical time which may be present in the data.

From now on, the set $\{u_0^{(j)}\}_j$ denotes the initial ensemble of particles with each particle living in parameter space: $u_0^{(j)} \in \mathcal{X}$. The iteration of particles consists of the set $\{u_n^{(j)}\}_{j,n}$ where j is the ensemble index and n the artificial time index. Each particle again is an element of parameter space: $u_n^{(j)} \in \mathcal{X}$. The measures μ_n will be approximated by an equally weighted sum of Dirac measures

$$\mu_n \simeq \frac{1}{J} \sum_{j=1}^{J} \delta_{u_n}^{(j)} \tag{1}$$

via the ensemble of particles. The initial ensemble is constructed based on the prior distribution and then mapped to the next iteration via a Gaussian approximation, i.e. given $\{u_n^{(j)}\}_{j,n}$, the transformed ensemble $\{u_{n+1}^{(j)}\}_{j,n}$ satisfies

$$\bar{u}_{n+1} = \bar{u}_n + K_n(y - \overline{\mathcal{G}}(u_n)) \qquad C(u_{n+1}) = C(u_n) - K_n C^{pu}(u_n)$$

with $K_n = C^{up}(u_n)(C^{pp}(u_n) + \frac{1}{h}\Gamma)^{-1}$. The operators C^{pp}, C^{up} and C^{pu} are the empirical covariances defined on $\mathcal{X}^J$ by

$$C^{pp}(u) = \frac{1}{J}\sum_{j=1}^{J}(\mathcal{G}(u^{(j)}) - \overline{\mathcal{G}}) \otimes (\mathcal{G}(u^{(j)}) - \overline{\mathcal{G}}),$$

$$C^{up}(u) = \frac{1}{J}\sum_{j=1}^{J}(u^{(j)} - \overline{u}) \otimes (\mathcal{G}(u^{(j)}) - \overline{\mathcal{G}}),$$

$$C^{pu}(u) = \frac{1}{J}\sum_{j=1}^{J}(\mathcal{G}(u^{(j)}) - \overline{\mathcal{G}}) \otimes (u^{(j)} - \overline{u}),$$

where u is short for the multiindex vector $(u^{(j)})_j \in \mathcal{X}^J$ and $\otimes$ denotes the tensor product (or rank one operator) given by

$$z_1 \otimes z_2 : \mathcal{H}_2 \to \mathcal{H}_1 \text{ with } h \mapsto z_1 \otimes z_2(h) := \langle z_2, h\rangle_{\mathcal{H}_2} \cdot z_1$$

for Hilbert spaces $(\mathcal{H}_1, \langle\cdot,\cdot\rangle_{\mathcal{H}_1}), (\mathcal{H}_2, \langle\cdot,\cdot\rangle_{\mathcal{H}_2})$ and $z_1 \in \mathcal{H}_1, z_2 \in \mathcal{H}_2$. The empirical means are given by

$$\overline{u} = \frac{1}{J}\sum_{j=1}^{J} u^{(j)}, \qquad \overline{\mathcal{G}} = \frac{1}{J}\sum_{j=1}^{J}\mathcal{G}(u^{(j)}).$$

The transformation of the ensemble from iteration n to $n+1$ is not uniquely determined via the Kalman update formula. For the EKI with perturbed observations, the update formula is shown to be satisfied in the mean, see e.g. [1].

Although we consider a Gaussian approximation for the measures μ_n, for our theoretical results we do not require any assumption on Gaussian prior distributions.

For a given artificial step-size $h > 0$ and $J \geqslant 2$ particles, the EnKF iteration for the jth particle is given by

$$u_{n+1}^{(j)} = u_n^{(j)} + C^{up}(u_n)(C^{pp}(u_n) + h^{-1}\Gamma)^{-1}(y_{n+1}^{(j)} - \mathcal{G}(u_n^{(j)})), \quad j = 1, \ldots, J, \tag{2}$$

where the initial particles $u_0^{(j)}$, $j = 1, \ldots, J$ are draws from the prior distribution. In each step, we consider artificially perturbed data

$$y_{n+1}^{(j)} = y + \xi_{n+1}^{(j)},$$

where the perturbations $\xi_{n+1}^{(j)}$, with respect to both j and n, are i.i.d. random variables distributed according to $\mathcal{N}(0, h^{-1}\Gamma)$. For a derivation of the EnKF for inverse problems, we refer to [31].

3. Continuous time limit

The continuous time limit of the discrete EnKF inversion (2) is formally a time discretization of the following SDE:

$$\mathrm{d}u_t^{(j)} = C^{up}(u_t)\Gamma^{-1}(y - \mathcal{G}(u_t^{(j)}))\,\mathrm{d}t + C^{up}(u_t)\Gamma^{-1/2}\,\mathrm{d}W_t^{(j)}. \tag{3}$$

Using the definition of the empirical covariance, (3) can be formulated equivalently as

$$\mathrm{d}u_t^{(j)} = \frac{1}{J}\sum_{k=1}^{J}\left\langle \mathcal{G}(u_t^{(k)}) - \overline{\mathcal{G}}_t, (y - \mathcal{G}(u_t^{(j)}))\,\mathrm{d}t + \sqrt{\Gamma}\,\mathrm{d}W_t^{(j)}\right\rangle_{\Gamma}(u_t^{(k)} - \overline{u}_t), \tag{4}$$

with $\langle \cdot, \cdot \rangle_\Gamma = \langle \Gamma^{-\frac{1}{2}} \cdot, \Gamma^{-\frac{1}{2}} \cdot \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean inner-product on $\mathbb{R}^K$. The processes $W^{(j)}$ are independent Brownian motions on $\mathbb{R}^K$. We further denote by $\mathcal{F}_t = \sigma(u_s, s \leqslant t)$ the filtration introduced by the particle dynamics. Most of the time we will write u_t or $u(t)$ to emphasize the dependence on time t. If the time dependence is clear from the context, we simplify the notation to u.

The formulation (4) reveals that solutions satisfy a generalization of the subspace property of [31, theorem 2.1] to continuous time.

Lemma 3.1. *Assume that $\mathcal{G}$ is locally Lipschitz and let $\mathcal{S}$ be the linear span of $\{u_0^{(j)}\}_{j=1}^J$, then $u_t^{(j)} \in \mathcal{S}$ for all $(t, j) \in [0, \infty) \times \{1, \ldots, J\}$ almost surely.*

We do not give all the technical details of the proof. First one needs to show that the initial value problem related to (4) has a unique $\mathcal{X}$-valued solution, which is assured by a local Lipschitz-property of the drift and diffusion. As the vector field on the right hand side of (4) maps $\mathcal{S}$ in $\mathcal{S}$, we can show by the same argument that there is also a unique $\mathcal{S}$-valued solution. Thus by the uniqueness in any $\mathcal{X}$ both solutions coincide, and all solution must be $\mathcal{S}$-valued. The subspace property reveals the regularization effect of the ensemble of particles in the inverse setting. Due to lemma 3.1, the EKI estimate lies in the subspace spanned by the initial ensemble, which is usually a much smaller space than the original parameter $\mathcal{X}$. Thus, the discretization via the ensemble of particles can be interpreted as a regularization or stabilization of the inverse problem.

3.1. The linear problem

For the whole paper, we will assume that the forward response operator is linear, i.e. $\mathcal{G}(\cdot) = A\cdot$ with $A \in \mathcal{L}(\mathcal{X}, \mathbb{R}^K)$. Then the continuous time limit (4) reads as

$$\mathrm{d}u_t^{(j)} = \frac{1}{J} \sum_{k=1}^{J} \left\langle A(u_t^{(k)} - \bar{u}_t), (y - Au_t^{(j)})\,\mathrm{d}t + \sqrt{\Gamma}\,\mathrm{d}W_t^{(j)} \right\rangle_\Gamma (u_t^{(k)} - \bar{u}_t). \tag{5}$$

We simplify notation by defining the empirical covariance operator

$$C(u) = \frac{1}{J} \sum_{k=1}^{J} (u^{(k)} - \bar{u}) \otimes (u^{(k)} - \bar{u}). \tag{6}$$

Thus equation (5) can be rewritten in the form

$$\mathrm{d}u_t^{(j)} = C(u_t)A^*\Gamma^{-1}(y - Au_t^{(j)})\,\mathrm{d}t + C(u_t)A^*\Gamma^{-1/2}\,\mathrm{d}W_t^{(j)}. \tag{7}$$

3.2. Well-posedness of the EnKF inversion

This section is devoted to proving the existence and uniqueness of global solutions of the set of coupled SDEs (7). Again, the local existence and uniqueness of $\mathcal{X}$-valued local solutions to (7) is straightforward by the local Lipschitz-property of the drift and diffusion on the right-hand side. Thus we rely on the subspace property of lemma 3.1, and first show that we can reduce the $\mathcal{X}$-valued setting without loss of generality to a finite-dimensional setting.

Lemma 3.2. *Without loss of generality we assume that the initial ensemble $(u_0^{(j)})_{j \in \{1, \ldots, J\}}$ is linearly independent almost surely and spans a J-dimensional vector space $\mathcal{S}$.*

Then there exists a linear operator $\tilde{A}:\mathbb{R}^J \to \mathbb{R}^K$ *such that equation (7) restricted to* $\mathcal{S}$ *is equivalent to*

$$\mathrm{d}v_t^{(j)} = \frac{1}{J}\sum_{k=1}^{J}\left\langle \tilde{A}v_t^{(k)} - \tilde{A}\overline{v}_t, (y-\tilde{A}v_t^{(j)})\,\mathrm{d}t + \Gamma^{\frac{1}{2}}\,\mathrm{d}W_t^{(j)}\right\rangle_{\Gamma}(v_t^{(k)}-\overline{v}_t) \tag{8}$$

for $v_t^{(j)} \in \mathbb{R}^J$, $\overline{v}_t := \frac{1}{J}\sum_{k=1}^{J} v_t^{(k)}$, *in the following sense: for* $u_t^{(j)} = \sum_{k=1}^{J}(v_t^{(j)})_k \cdot u_0^{(k)}$ *one has that* u_t *is a* $\mathcal{S}$*-valued solution of (7) if and only if* v_t *is a solution of (8).*

Proof. By lemma 3.1, any $\mathcal{S}$-valued process $u(t)$ can be uniquely expanded as a linear combination $u^{(j)}(t) = \sum_{l=1}^{J} v_l^{(j)}(t)\cdot u^{(l)}(0)$ for every $j\in\{1,\dots,J\}$, $t\geqslant 0$ and coordinates $v_l^{(j)}(t)\in\mathbb{R}$. Let $\Phi^{-1}:\mathbb{R}^J\to\mathcal{S}$ denote the basis isomorphism, i.e. $\Phi:\mathcal{S}\to\mathbb{R}^J$ with $u=\sum_{l=1}^{J} v_l u^{(l)}(0) \overset{\Phi}{\mapsto} (v_1,\dots,v_J)^\top$. Since Φ is a linear isomorphism, (7) can be equivalently transformed to

$$\begin{aligned}\mathrm{d}\Phi(u^{(j)}(t)) &= \Phi(\mathrm{d}u^{(j)}(t))\\ &= \frac{1}{J}\sum_{k=1}^{J}\langle A(u^{(k)}(t)-\overline{u}(t)), (y-Au^{(j)}(t))\,\mathrm{d}t+\Gamma^{\frac{1}{2}}\,\mathrm{d}W_t^{(j)}\rangle_\Gamma(\Phi(u^{(k)}(t))-\Phi(\overline{u}(t))).\end{aligned}$$

Thus, with $\tilde{A}=A\Phi^{-1}$, we obtain

$$\begin{aligned}\mathrm{d}\Phi(u^{(j)}(t)) = \frac{1}{J}\sum_{k=1}^{J}&\langle \tilde{A}\Phi(u^{(k)}(t)-\overline{u}(t)), (y-\tilde{A}\Phi(u^{(j)}(t)))\,\mathrm{d}t+\Gamma^{\frac{1}{2}}\,\mathrm{d}W_t^{(j)}\rangle_\Gamma\\ &\cdot(\Phi(u^{(k)}(t))-\Phi(\overline{u}(t))).\end{aligned}$$

The assertion follows with $v^{(j)} := \Phi(u^{(j)})$. □

Remark 3.3. By the previous lemma solving equation (7) is equivalent to solving the finite dimensional equation (8). Thus, to simplify notation we will assume without loss of generality that $\mathcal{X}=\mathbb{R}^I$, $I\in\mathbb{N}$, $I\leqslant J$. In the case of linearly independent initial ensemble we can assume $I=J$.

For the study of the dynamical behavior of the ensemble, we will sometimes require the following assumption for results in the parameters space:

$$\text{The linear operator } \tilde{A} \text{ defined above is one-to-one.} \tag{9}$$

Note that assumption (9) seems to be a rather strict assumption: it requires that the forward operator ‘sees everything’ and secondly, this means that $(\tilde{A}\Phi(u_0^{(j)}))_{j\in\{1,\dots,J\}}\in\mathbb{R}^K$ is linearly independent. This implies the restriction on the number of particles $J\leqslant K$. However, note that this assumption is on the operator $\tilde{A}$, i.e. we do not assume that A is one-to-one. The discretization of the parameter space via the ensemble of particles acts as a regularization of the inverse problem in this setting. We will need assumption (9) only when we want to prove dynamical properties in parameter space. This makes sense as we cannot hope for convergence to the true parameter if the forward operator is indifferent with respect to some components of this parameter value. Our convergence results in the observation space hold without assumption (9).

In order to prove the existence and uniqueness of global solutions we rewrite the set of coupled SDEs (7) as a single SDE of the following form:

$$\mathrm{d}u_t = F(u_t)\,\mathrm{d}t + G(u_t)\,\mathrm{d}W_t,$$

with $u_t = (u_t^{(j)})_{j\in\{1,\dots,J\}} \in \mathbb{R}^{IJ\times 1}$, $W_t = (W_t^{(j)})_{j\in\{1,\dots,J\}} \in \mathbb{R}^{J^2\times 1}$ and

$$F(x) = (C(x)A^*\Gamma^{-1}(y - Ax^{(j)}))_{j\in\{1,\dots,J\}} \in \mathbb{R}^{IJ\times 1},$$
$$G(x) = \mathrm{diag}(C(x)A^*\Gamma^{-\frac{1}{2}})_{j\in\{1,\dots,J\}} \in \mathbb{R}^{IJ\times J^2},$$

where $x = (x^{(j)})_{j\in\{1,\dots,J\}} \in \mathbb{R}^{IJ\times 1}$ and $\mathrm{diag}(B_j)_{j\in\{1,\dots,J\}}$ is a diagonal block matrix with matrices $(B_j)_{j\in\{1,\dots,J\}}$ on the diagonal. For a given matrix $B = (b_{ij})_{ij} \in \mathbb{R}^{n\times m}$, $m, n \in \mathbb{N}$, the Frobenius norm $\|B\|_F$ is defined by $\|B\|_F^2 = \mathrm{trace} B^\top B = \sum_{i,j} b_{ij}^2 \geqslant \|B\|_2^2$.

We will now formulate and prove the main result of this section on the well-posedness of the EnKF inversion.

Theorem 3.4. *Let $u_0 = (u_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $u_0^{(j)} : \Omega \to \mathcal{X}$ which are linearly independent almost surely. Then for all $T \geqslant 0$ there exists a unique strong solution $(u_t)_{t\in[0,T]}$ (up to $\mathbb{P}$-indistinguishability) of the set of coupled SDEs (7).*

Proof. For the proof we will assume without loss of generality that $\mathcal{X} = \mathbb{R}^I$ for I sufficiently large, as discussed before. The proof of existence and uniqueness of local strong solutions for (7) (up to a stopping-time) is standard, due to the local Lipschitz property of the drift F and the diffusion G. Note that both are polynomials.

The global existence of a strong solution is based on stochastic Lyapunov theory. See for example theorem 4.1 of [32]. We only need to construct a function $V \in C^2(\mathcal{X}; \mathbb{R}_+)$ such that for some constant $c > 0$

$$LV(x) := \nabla V(x)\cdot F(x) + \frac{1}{2}\mathrm{trace}(G^T(x)\mathrm{Hess}[V](x)G(x)) \leqslant cV(x) \tag{10}$$

and

$$\inf_{|x|>R} V(x) \to \infty \text{ as } R \to \infty \tag{11}$$

hold true.

We can uniquely decompose $y \in \mathbb{R}^K$ as $y = y_1 + y_2$, with $y_1 \in \mathcal{R}(\Gamma^{-\frac{1}{2}}A)$ and $y_2 \in \mathcal{R}(\Gamma^{-\frac{1}{2}}A)^\perp$, where $\mathcal{R}(\Gamma^{-\frac{1}{2}}A)$ denotes the image of $\Gamma^{-\frac{1}{2}}A$. We fix $\tilde{u} \in \mathbb{R}^J$ such that $\Gamma^{-\frac{1}{2}}A\tilde{u} = y_1$ and define the Lyapunov function

$$V(u) := V_1(u) + V_2(u) = \frac{1}{J}\sum_{j=1}^J \|u^{(j)} - \bar{u}\|^2 + \|\bar{u} - \tilde{u}\|^2.$$

Obviously, (11) is satisfied.

The generator L applied to V is given by $LV = LV_1 + LV_2$ with

$$LV_1(u) = -\frac{J+1}{J^3}\sum_{j,l=1}^{J}\langle u^{(j)}-\bar{u}, u^{(l)}-\bar{u}\rangle\langle \Gamma^{-\frac{1}{2}}A(u^{(l)}-\bar{u}), \Gamma^{-\frac{1}{2}}A(u^{(j)}-\bar{u})\rangle$$

$$\begin{aligned} LV_2(u) &= -\frac{2}{J}\sum_{l=1}^{J}\langle \bar{u}-\tilde{u}, u^{(l)}-\bar{u}\rangle\langle \Gamma^{-\frac{1}{2}}A(u^{(l)}-\bar{u}), \Gamma^{-\frac{1}{2}}A(\bar{u}-\tilde{u})\rangle \\ &\quad + \frac{1}{J^3}\sum_{j,l=1}^{J}\langle u^{(j)}-\bar{u}, u^{(l)}-\bar{u}\rangle\langle \Gamma^{-\frac{1}{2}}A(u^{(l)}-\bar{u}), \Gamma^{-\frac{1}{2}}A(u^{(j)}-\bar{u})\rangle, \end{aligned}$$

where we used $\langle \Gamma^{-\frac{1}{2}}A(u^{(l)}-\bar{u}), y_2\rangle = 0$ for all $l \in \{1, \dots, J\}$ wich is true by construction. Thus, as $A^\star\Gamma^{-1}A$ is a symmetric non-negative matrix by lemma A.2 the nonnegativity of the generator follows:

$$\begin{aligned} LV(u) &= -\frac{2}{J}\sum_{l=1}^{J}\langle \bar{u}-\tilde{u}, u^{(l)}-\bar{u}\rangle\langle \Gamma^{-\frac{1}{2}}A(u^{(l)}-\bar{u}), \Gamma^{-\frac{1}{2}}A(\bar{u}-\tilde{u})\rangle \\ &\quad - \frac{1}{J^2}\sum_{j,l=1}^{J}\langle u^{(j)}-\bar{u}, u^{(l)}-\bar{u}\rangle\langle \Gamma^{-\frac{1}{2}}A(u^{(l)}-\bar{u}), \Gamma^{-\frac{1}{2}}A(u^{(j)}-\bar{u})\rangle \\ &\leqslant 0. \end{aligned}$$

Thus (10) holds true, for all $c > 0$. □

4. Quantification of the ensemble collapse

The dynamics of the ensemble Kalman filter as presented here can be decomposed into two parts:

- Ensemble collapse. This means convergence of all ensemble members to their joint mean (‘The estimator becomes more confident’).
- Convergence of the ensemble mean. This means that the ensemble mean will tend to a parameter value which is consistent with the data.

Those two notions are totally different in concept but also strongly intertwined in the dynamics of the EnKF (see also the discussion at the beginning of section 5).

We start by quantifying the ensemble collapse. We will present results in the data (or observation) space as well as in the parameter space.

For the further analysis, we introduce the centered quantities

$$e^{(j)} = u^{(j)} - \bar{u}, \qquad r^{(j)} = u^{(j)} - u^\dagger,$$

where $e^{(j)}$ denotes difference of each particles to the mean and $r^{(j)}$ denotes the residuals. Here, the data y is the perturbed image of a truth $u^\dagger \in \mathcal{X}$ under A, i.e. $y = Au^\dagger + \eta$. The quantities satisfy the following equations (note that $C(u) = C(e)$ as the mean of the $e^{(j)}$ vanishes):

$$\mathrm{d}e_t^{(j)} = -C(e_t)A^*\Gamma^{-1}Ae_t^{(j)}\,\mathrm{d}t + C(e_t)A^*\Gamma^{-\frac{1}{2}}\,\mathrm{d}(W_t^{(j)} - \overline{W}_t), \tag{12}$$

$$\mathrm{d}r_t^{(j)} = \mathrm{d}u_t^{(j)} = C(u_t)A^*\Gamma^{-1}(y - Au_t^{(j)})\,\mathrm{d}t + C(u_t)A^*\Gamma^{-1/2}\,\mathrm{d}W_t^{(j)}, \tag{13}$$

with $\overline{W}_t := \frac{1}{J}\sum_{j=1}^{J} W^{(j)}$. The dynamical behavior of the empirical mean is given by

$$\mathrm{d}\bar{u}_t = \frac{1}{J}\sum_{k=1}^{J}(u_t^{(k)} - \bar{u}_t)\langle A(u_t^{(k)} - \bar{u}_t), (y - A\bar{u}_t)\,\mathrm{d}t + \Gamma^{\frac{1}{2}}\,\mathrm{d}\overline{W}_t\rangle_\Gamma.$$

To simplify notation, we also introduce the transformed quantities

$$\mathfrak{r}^{(j)} := \Gamma^{-\frac{1}{2}}Ar^{(j)}, \qquad \mathfrak{e}^{(j)} := \Gamma^{-\frac{1}{2}}Ae^{(j)} = \mathfrak{r}^{(j)} - \bar{\mathfrak{r}}$$

denoting the residuals in observation space and the mapped difference of each particle to the empirical mean.

We will now make a first step towards proving ensemble collapse. As we work with SDEs, any dynamical property can only hold in some probabilistic sense. The following lemma shows that we have ensemble collapse in the L^p sense, with the upper bound for valid parameters p being dependent on the number of particles J.

Lemma 4.1. *Let $p \in [2, J+3)$ and $u_0 = (u_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $u_0^{(j)} : \Omega \to \mathcal{X}$ such that $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_0^{(j)}|^p] < \infty$. Then*

$$t \in [0,\infty) \mapsto \|\mathfrak{e}_t\|_{\mathcal{L}_p(\Omega,\mathbb{R}^K)} := \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_t^{(j)}|^p\right]^{\frac{1}{p}}$$

is monotonically decreasing in t. Furthermore there exists a constant $C > 0$ such that for all $t \geqslant 0$

$$\int_0^t \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_s^{(j)}|^{p+2}\right]\mathrm{d}s < C.$$

Proof. We will prove the assertion in the case $p = 2$, in order to give the key ideas. The case $p > 2$ is very similar, but much more technical. We postpone all details in that case to the appendix.

Applying $\Gamma^{-\frac{1}{2}}A$ to $e^{(j)}$ implies that the quantity $\mathfrak{e}^{(j)}$ satisfies (see (12) and (6))

$$\begin{aligned}\mathrm{d}\mathfrak{e}_t^{(j)} &= -C(\mathfrak{e}_t)\mathfrak{e}_t^{(j)}\,\mathrm{d}t + C(\mathfrak{e}_t)\,\mathrm{d}(W_t^{(j)} - \overline{W}_t)\\ &= -\frac{1}{J}\sum_{k=1}^{J}\mathfrak{e}_t^{(k)}\langle\mathfrak{e}_t^{(k)},\mathfrak{e}_t^{(j)}\rangle\,\mathrm{d}t + \frac{1}{J}\sum_{k=1}^{J}\mathfrak{e}_t^{(k)}\langle\mathfrak{e}_t^{(k)},\mathrm{d}(W_t^{(j)} - \overline{W}_t)\rangle.\end{aligned}$$

Itô's formula gives

$$\begin{aligned}\mathrm{d}|\mathfrak{e}_t^{(j)}|^2 &= 2\langle\mathfrak{e}_t^{(j)},\mathrm{d}\mathfrak{e}_t^{(j)}\rangle + \langle\mathrm{d}\mathfrak{e}_t^{(j)},\mathrm{d}\mathfrak{e}_t^{(j)}\rangle\\ &= -\frac{2}{J}\sum_{k=1}^{J}\left\langle\mathfrak{e}_t^{(j)},\mathfrak{e}_t^{(k)}\right\rangle^2\mathrm{d}t + 2\mathfrak{e}_t^{(j)T}C(\mathfrak{e}_t)\,\mathrm{d}(W_t^{(j)} - \overline{W}_t)\\ &\quad + \frac{1}{J^2}\sum_{k,l=1}^{J}\left\langle\mathfrak{e}_t^{(k)},\mathfrak{e}_t^{(l)}\right\rangle\left\langle\mathfrak{e}_t^{(k)},\mathrm{d}(W_t^{(j)} - \overline{W})\right\rangle\left\langle\mathfrak{e}_t^{(l)},\mathrm{d}(W_t^{(j)} - \overline{W})\right\rangle,\end{aligned}$$

and with lemma A.1 to evaluate the Itô correction we get

$$\mathrm{d}|\mathfrak{e}_t^{(j)}|^2 = -\frac{2}{J}\sum_{k=1}^{J}\langle \mathfrak{e}_t^{(j)}, \mathfrak{e}_t^{(k)}\rangle^2\,\mathrm{d}t + 2\mathfrak{e}_t^{(j)T}C(\mathfrak{e}_t)\,\mathrm{d}(W_t^{(j)} - \overline{W}_t) + \frac{J-1}{J^3}\sum_{k,l=1}^{J}\langle \mathfrak{e}_t^{(k)}, \mathfrak{e}_t^{(l)}\rangle^2\,\mathrm{d}t.$$

Summing over all particles leads to

$$\begin{aligned}\mathrm{d}\left(\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_t^{(j)}|^2\right) &= -\frac{J+1}{J^3}\sum_{j,k=1}^{J}\left\langle \mathfrak{e}_t^{(j)}, \mathfrak{e}_t^{(k)}\right\rangle^2\,\mathrm{d}t + \frac{2}{J}\sum_{j=1}^{J}\mathfrak{e}_t^{(j)\top}C(\mathfrak{e}_t)\,\mathrm{d}(W_t^{(j)} - \overline{W}_t)\\ &= -\frac{J+1}{J^3}\sum_{j,k=1}^{J}\left\langle \mathfrak{e}_t^{(j)}, \mathfrak{e}_t^{(k)}\right\rangle^2\,\mathrm{d}t + \frac{2}{J}\sum_{j=1}^{J}\mathfrak{e}_t^{(j)\top}C(\mathfrak{e}_t)\,\mathrm{d}W_t^{(j)}.\end{aligned}$$

The last step follows from $\sum_j \mathfrak{e}^{(j)} = 0$. This yields

$$\begin{aligned}&\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_t^{(j)}|^2 - \frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_0^{(j)}|^2\\ &\qquad = -\frac{J+1}{J^3}\int_0^t\sum_{j,k=1}^{J}\langle \mathfrak{e}_t^{(j)}, \mathfrak{e}_t^{(k)}\rangle^2\,\mathrm{d}t + \frac{2}{J}\int_0^t\sum_{j=1}^{J}\mathfrak{e}_t^{(j)\top}C(\mathfrak{e}_t)\,\mathrm{d}W_t^{(j)}.\end{aligned} \tag{14}$$

Now we cannot simply take the expectation, as we do not know that the stochastic integral is a martingale. We need a localization. Set $t, s \geqslant 0$ and let $(\tau_n)_{n\in\mathbb{N}}$ with $\tau_n \xrightarrow{n} \infty$ a.s. be a sequence of deterministically bounded stopping times, such that

$$\int_s^{s+(t\wedge\tau_n)}\mathfrak{e}_s^{(j)T}C(\mathfrak{e}_s)\,\mathrm{d}W_s^{(j)}$$

is a martingale for every $j \in \{1, \cdots, J\}$. This is possible by definition of local martingales, with any stochastic integral being one. For example we can take for τ_n the minimum of n and the first exit time of $\mathfrak{e}_s$ at radius n. Then, for all $n \in \mathbb{N}$, from (14) (after rebasing the integration interval from $[0, t]$ to $[s, s+t]$) we obtain

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_{s+(t\wedge\tau_n)}^{(j)}|^2\right] - \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_s^{(j)}|^2\right] = -\mathbb{E}\left[\int_s^{s+(t\wedge\tau_n)}\frac{J+1}{J^3}\sum_{j,k=1}^{J}\langle \mathfrak{e}_r^{(j)}, \mathfrak{e}_r^{(k)}\rangle^2\,\mathrm{d}r\right]$$

As $\tau_n \to \infty$, applying Fatou's lemma on the left hand side and applying the monotone convergence theorem on the right hand side gives

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_{s+t}^{(j)}|^2\right] - \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_s^{(j)}|^2\right] \leqslant -\mathbb{E}\left[\int_s^{s+t}\frac{J+1}{J^3}\sum_{j,k=1}^{J}\langle \mathfrak{e}_r^{(j)}, \mathfrak{e}_r^{(k)}\rangle^2\,\mathrm{d}r\right] \leqslant 0, \tag{15}$$

which implies that $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_t^{(j)}|^2]$ is monotonically decreasing in t.

Finally,

$$\int_0^t \mathbb{E}\left[\frac{J+1}{J^3}\sum_{j=1}^J |\mathfrak{e}_s^{(j)}|^4\right] \mathrm{d}s \leqslant \int_0^t \mathbb{E}\left[\frac{J+1}{J^3}\sum_{j,k=1}^J \langle \mathfrak{e}_s^{(j)}, \mathfrak{e}_s^{(k)}\rangle^2\right] \mathrm{d}s \leqslant \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^J |\mathfrak{e}_0^{(j)}|^2\right],$$

where the first inequality is trivial by inserting non-negative terms in the sum and the second inequality is (15) with $s = 0$. This proves the second claim.

Let us finally remark that $\tau_n \to \infty$ necessarily holds. If we assume that $\tau_n \to \tau_*$ then the previous argument with $s = 0$ and arbitrary $T > 0$ gives $\mathbb{E}[\frac{1}{J}\sum_{j=1}^J |\mathfrak{e}_{t\wedge\tau_*}^{(j)}|^2] < \infty$. Thus $t < \tau_*$ for our choice of stopping time. □

The main obstacle in quantifying the ensemble collapse is proving that the stochastic integral in (14) is actually a true martingale. See lemma A.4 for details.

Theorem 4.2. *Let $u_0 = (u_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable random variables $u_0^{(j)} : \Omega \to \mathcal{X}$ such that $C_0 := \mathbb{E}[\frac{1}{J}\sum_{j=1}^J |\mathfrak{e}_0^{(j)}|^2] < \infty$. Then, the ensemble collapse is quantified by*

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^J |\mathfrak{e}_t^{(j)}|^2\right] \leqslant \frac{1}{\frac{J+1}{J^2}t + \frac{1}{C_0}}. \tag{16}$$

Proof. By lemma A.4 we can directly take expectations in (14) to obtain

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^J |\mathfrak{e}_t^{(j)}|^2\right] = \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^J |\mathfrak{e}_0^{(j)}|^2\right] - \frac{J+1}{J^3}\int_0^t \mathbb{E}\left[\sum_{j,k=1}^J \langle \mathfrak{e}_s^{(j)}, \mathfrak{e}_s^{(k)}\rangle^2\right] \mathrm{d}s.$$

Note that by dropping the non-negative mixed terms $j \neq k$ and by using Jensen's and Young's inequality

$$\frac{J+1}{J^3}\mathbb{E}\left[\sum_{j,k=1}^J \langle \mathfrak{e}_s^{(j)}, \mathfrak{e}_s^{(k)}\rangle^2\right] \geqslant \frac{J+1}{J^2}\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^J |\mathfrak{e}_s^{(j)}|^2\right]^2.$$

Thus setting $t \mapsto h(t) := \mathbb{E}[\frac{1}{J}\sum_{j=1}^J |\mathfrak{e}_t^{(j)}|^2]$ we can write

$$h(t) = h(0) - \frac{J+1}{J^2}\int_0^t h^2(s)\,\mathrm{d}s - \int_0^t p(s)\mathrm{d}s$$

for a non-negative function $p \geqslant 0$. Hence, we can differentiate to obtain the differential inequality

$$h' \leqslant -\frac{J+1}{J^2}h^2,$$

from which by a comparison argument for scalar ODE it follows that

$$h(t) = \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_t^{(j)}|^2\right] \leqslant \frac{1}{\frac{J+1}{J^2}t + \frac{1}{h(0)}}.$$

□

Corollary 4.3. *Under the same assumptions as in theorem 4.2 and under assumption (9) it holds true that*

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|e_t^{(j)}|^2\right] \leqslant \frac{1}{\sigma_{\min}}\frac{1}{\frac{J+1}{J^2}t + \frac{1}{C_0}},$$

where $\sigma_{\min}$ is the smallest eigenvalue of the positive definite operator $A^*\Gamma^{-1}A$.

Proof. The assertion follows directly from the inequality

$$|\mathfrak{e}^{(j)}|^2 = |\Gamma^{-\frac{1}{2}}Ae^{(j)}|^2 = \langle e^{(j)}, A^*\Gamma^{-1}Ae^{(j)}\rangle \geqslant \sigma_{\min}|e^{(j)}|^2,$$

since $A^*\Gamma^{-1}A$ is positive definite. □

Remark 4.4. Note that the bound in (16) deteriorates with growing number of particles J, i.e. the result does not quantify the ensemble collapse in the large ensemble size limit. However, the presented analysis is tailored for fixed ensemble size and we will demonstrate in the numerical experiments that the derived bound (16) can be efficiently used to quantify the collapse in this setting.

4.1. Higher-order ensemble collapse

Here we state the result for higher moments and postpone the proof to the appendix, as they are very similar to but technically more involved than the case $p = 2$.

Theorem 4.5. *Let $p \in (2, \frac{J+3}{2})$ and let $u_0 = (u_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $u_0^{(j)} : \Omega \to \mathcal{X}$ such that $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_0^{(j)}|^p] < \infty$. Then it holds true that*

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_t^{(j)}|^p\right] \leqslant \frac{J^{\frac{p}{2}}}{\left(\frac{2}{p}C(p,J)K^{-\frac{2}{p}}J^{1-\frac{2}{p}}t + \left(K^{\frac{p-1}{2}}\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_0^{(j)}|^p\right]\right)^{-\frac{2}{p}}\right)^{\frac{p}{2}}},$$

with $C(p,J) := \frac{p}{J^2}(1 - \frac{(p-2+J)\cdot(J-1)}{2J^2} - \frac{p-2}{2J^2})$.

Proof. The proof based on Itô's formula and a comparison principle for ODEs is very similar to the case $p = 2$. Details can be found in the appendix. □

Remark 4.6. Note that a larger ensemble seems to regularize the dynamics. The higher the ensemble number J, the larger is the highest moment of ensemble collapse we can bound.

The restriction $2p < J + 3$ comes from the fact that we need the martingale property of the stochastic integral, which we obtain from the bounds in lemma 4.1.

Corollary 4.7. *Under the same assumptions as in theorem 4.5 and under assumption (9) it holds true that*

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|e_t^{(j)}|^p\right] \leqslant \frac{J^{\frac{p}{2}}}{\left(\sigma_{\min}\cdot\frac{2}{p}C(p,J)K^{-\frac{2}{p}}J^{1-\frac{2}{p}}t + \left(K^{\frac{p-1}{2}}\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_0^{(j)}|^p\right]\right)^{-\frac{2}{p}}\right)^{\frac{p}{2}}},$$

where $\sigma_{\min}$ is the smallest eigenvalue of the positive definite operator $A^*\Gamma^{-1}A$ and $C(p,J)$ is defined in theorem 4.5.

4.2. Almost sure ensemble collapse

We have proven conditions for ensemble collapse in pth moments, but a stronger measure of stochastic convergence is almost sure convergence. This is the focus of this section.

Theorem 4.8. *Let $u_0 = (u_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $u_0^{(j)} : \Omega \to \mathcal{X}$ and $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$ a positive, monotonically increasing and differentiable function such that $\int_0^\infty \frac{\gamma'(s)^2}{\gamma(s)}\,\mathrm{d}s < \infty$. Then the trivial solution of*

$$\mathrm{d}\mathfrak{e}_t^{(j)} = -C(\mathfrak{e}_t)\mathfrak{e}_t^{(j)}\mathrm{d}t + C(\mathfrak{e}_t)\mathrm{d}(W_t^{(j)} - \overline{W}_t) \tag{17}$$

is almost surely asymptotically stable with rate function $\rho(t) = (\gamma(t))^{-\frac{1}{2}}$. In particular, $(\mathfrak{e}_t^{(j)})_{j=1,\dots,J}$ converges to zero almost surely as $t \to \infty$.

For examples of γ see the remark below.

Proof. The idea of this proof is based on theorem 4.6.2 in [33]. We define the stochastic Lyapunov function

$$V(\mathfrak{e}, t) = \gamma(t)\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}^{(j)}|^2.$$

The generator applied to V fulfills

$$\begin{aligned} LV(\mathfrak{e}, t) &= \frac{\gamma'(t)}{J}\sum_{j=1}^{J}|\mathfrak{e}^{(j)}|^2 - \gamma(t)\frac{J+1}{J^3}\sum_{j,k=1}^{J}\langle\mathfrak{e}^{(k)}, \mathfrak{e}^{(j)}\rangle^2 \\ &\leqslant \frac{\gamma'(t)}{J}\sum_{j=1}^{J}|\mathfrak{e}^{(j)}|^2 - \gamma(t)\frac{J+1}{J^3}\sum_{j=1}^{J}|\mathfrak{e}^{(j)}|^4. \end{aligned}$$

We can maximize this w.r.t. $(|\mathfrak{e}^{(1)}|^2, \dots, |\mathfrak{e}^{(J)}|^2)$ and get the following bound for LV:

$$LV(\mathfrak{e}, t) \leqslant \frac{\gamma'(t)^2}{\gamma(t)}\frac{1}{4}\frac{J^2}{J+1} =: \eta(t).$$

Since $\int_0^\infty \eta(t)\,\mathrm{d}t < \infty$, with theorem 4.6.2 of [33] the trivial solution of (17) is almost surely asymptotically stable with rate function $\rho(t) = (\gamma(t))^{-\frac{1}{2}}$. □

Corollary 4.9. *Under the same assumptions as in theorem 4.8 and assumption (9) it holds true that* $(e_t^{(j)})_{j=1,\dots,J}$ *converges to zero almost surely as* $t \to \infty$ *with rate function* $\rho(t) = (\gamma(t))^{-\frac{1}{2}}$.

Remark 4.10. Let us give two examples of admissible $\gamma(t)$:

- $\gamma(t) = (t+\varepsilon)^\alpha$ for $\alpha \in (0,1)$ and $\varepsilon > 0$ sufficiently small to obtain the rate function $\rho(t) = \dfrac{1}{(t+\varepsilon)^{\frac{\alpha}{2}}}$.
- $\gamma(t) = (t+\varepsilon)\log(t+\varepsilon)^{-\alpha}$ for arbitrarily small $\alpha > \frac{1}{2}$ and $\varepsilon > 0$ to obtain the rate function $\rho(t) = \dfrac{\log(t+\varepsilon)^{\frac{\alpha}{2}}}{(t+\varepsilon)^{\frac{1}{2}}}$.

4.3. Ensemble collapse in the parameter space

The following result holds true without the strong assumption (9). It only shows a monotone decrease, but not the collapse, where we need (9). See also corollaries 4.7 and 4.9.

Proposition 4.11. *Let* $u_0 = (u_0^{(j)})_{j\in\{1,\dots,J\}}$ *be* $\mathcal{F}_0$*-measurable maps* $u_0^{(j)} : \Omega \to \mathcal{X}$ *such that* $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|e_0^{(j)}|^2] < \infty$. *Then it holds true that* $t \mapsto \mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|e_t^{(j)}|^2]^{\frac{1}{2}}$ *is monotonically decreasing for* $t \geqslant 0$.

Proof. Itô's formula leads to

$$\begin{aligned}
\mathrm{d}|e_t^{(j)}|^2 &= 2\langle e_t^{(j)}, \mathrm{d}e_t^{(j)}\rangle + \langle \mathrm{d}e_t^{(j)}, \mathrm{d}e_t^{(j)}\rangle \\
&= -\frac{2}{J}\sum_{k=1}^{J}\langle e_t^{(j)}, e_t^{(k)}\rangle\langle \Gamma^{-\frac{1}{2}}Ae_t^{(k)}, \Gamma^{-\frac{1}{2}}Ae_t^{(j)}\rangle\,\mathrm{d}t \\
&\quad + \frac{2}{J}\sum_{k=1}^{J}\langle e_t^{(j)}, e_t^{(k)}\rangle\langle \Gamma^{-\frac{1}{2}}Ae_t^{(k)}, \mathrm{d}(W_t^{(j)} - \overline{W}_t)\rangle \\
&\quad + \frac{1}{J^2}\sum_{k,l=1}^{J}\frac{J-1}{J}\langle e_t^{(k)}, e_t^{(l)}\rangle\langle \Gamma^{-\frac{1}{2}}Ae_t^{(k)}, \Gamma^{-\frac{1}{2}}Ae_t^{(l)}\rangle\,\mathrm{d}t
\end{aligned}$$

and taking the mean over all particles $j \in \{1,\dots,J\}$ gives

$$\begin{aligned}
\mathrm{d}\Big(\frac{1}{J}\sum_{j=1}^{J}|e_t^{(j)}|^2\Big) &= -\frac{J+1}{J^3}\sum_{j,k=1}^{J}\langle e_t^{(k)}, e_t^{(j)}\rangle\langle \Gamma^{-\frac{1}{2}}Ae_t^{(k)}, \Gamma^{-\frac{1}{2}}Ae_t^{(j)}\rangle\,\mathrm{d}t \\
&\quad + \frac{2}{J^2}\sum_{k,j=1}^{J}\langle e_t^{(k)}, e_t^{(j)}\rangle\langle \Gamma^{-\frac{1}{2}}Ae_t^{(k)}, \mathrm{d}(W_t^{(j)} - \overline{W}_t)\rangle.
\end{aligned}$$

Again, we do not know whether the stochastic integral is a martingale, and we need again a localization. Consider as in lemma 4.1 a sequence of stopping times $(\tau_n)_{n\in\mathbb{N}}$ with $\tau_n \to \infty$ a.s., such that

$$\int_0^{t\wedge\tau_n} \frac{2}{J^2} \sum_{k,j=1}^{J} \langle e_s^{(k)}, e_s^{(j)}\rangle \langle \Gamma^{-\frac{1}{2}} A e_s^{(k)}, \mathrm{d}(W_s^{(j)} - \overline{W}_s)\rangle$$

is a martingale. We obtain for all $n \in \mathbb{N}$

$$\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J} |e_{t\wedge\tau_n}^{(j)}|^2] = \mathbb{E}[\frac{1}{J}\sum_{j=1}^{J} |e_0^{(j)}|^2] - \mathbb{E}[\int_0^{t\wedge\tau_n} \frac{J+1}{J^3} \sum_{j,k=1}^{J} \langle e_s^{(k)}, e_s^{(j)}\rangle \langle \Gamma^{-\frac{1}{2}} A e_s^{(k)}, \Gamma^{-\frac{1}{2}} A e_s^{(j)}\rangle \,\mathrm{d}s]$$

and hence, as we have the positivity of the integrand by lemma A.2, we obtain that $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J} |e_{t\wedge\tau_n}^{(j)}|^2]$ is monotonically decreasing and bounded. Analogously to the proof of lemma 4.1, we can pass to the limit $n \to \infty$ by Fatou's lemma and the monotone convergence theorem. This implies for $t > s \geqslant 0$

$$\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J} |e_{t+s}^{(j)}|^2] \leqslant \mathbb{E}[\frac{1}{J}\sum_{j=1}^{J} |e_s^{(j)}|^2] - \mathbb{E}[\int_s^{s+t} \frac{J+1}{J^3} \sum_{j,k=1}^{J} \langle e_r^{(k)}, e_r^{(j)}\rangle \langle \Gamma^{-\frac{1}{2}} A e_r^{(k)}, \Gamma^{-\frac{1}{2}} A e_r^{(j)}\rangle \,\mathrm{d}r].$$

In particular, it follows that $\mathbb{E}[\frac{1}{J}\sum\limits_{j=1}^{J} |e_t^{(j)}|^2]$ is monotonically decreasing. □

5. Convergence to ground truth

Under the assumption that y is the image of a truth $u^\dagger \in \mathcal{X}$ under A, we are interested now in the analysis of the convergence to the truth. Recall the equation

$$\mathrm{d}\mathfrak{r}_t^{(j)} = -C(\mathfrak{r}_t)\mathfrak{r}_t^{(j)}\,\mathrm{d}t + C(\mathfrak{r}_t)\,\mathrm{d}W_t^{(j)}.$$

The following properties can be shown for the residuals.

Proposition 5.1. *Let y be the image of a truth $u^\dagger \in \mathcal{X}$ under A and $u_0 = (u_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $u_0^{(j)} : \Omega \to \mathcal{X}$ such that $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J} |\mathfrak{r}_0^{(j)}|^2] < \infty$. Then $\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J} |\mathfrak{r}_t^{(j)}|^2]^{\frac{1}{2}}$ is monotonically decreasing.*

Proof. The assertions follow by arguments similar to the proof of proposition 4.11. □

The main issue in showing convergence of the residuals $\mathfrak{r}^{(j)}$ to zero is, seemingly paradoxically, the fact of ensemble collapse. Now obviously, convergence cannot happen without ensemble collapse (as the particles cannot converge on the same point if their distance to their joint mean does not vanish) but ensemble collapse itself actually delays convergence. To see this, consider the following toy model. It is deterministic, but the same effects can be observed by straightforward extension to an SDE setting:

$$\begin{aligned} w'(t) &= -w(t)^3 \\ z'(t) &= -w(t)^2 \cdot z(t). \end{aligned}$$

Now this system of ODEs can be solved explicitly by separation of variables and it can be seen that for any $w(0) \neq 0$, both w and z will converge to 0 for $t \to \infty$. It makes sense to try and apply Lyapunov theory with a straightforward Lyapunov functional $V(w,z) = \frac{1}{2}(w^2 + z^2)$. Then

$$\dot{V}(w,z) = \langle w(t) \cdot (-w(t)^3) + z(t) \cdot (-w(t)^2 \cdot z(t)) = -w(t)^2 \cdot V(w(t), z(t)).$$

But $\dot{V}(w,z)$ is not negatively definite in any neighborhood of $(0,0)$ (the problem being the manifold $w = 0$). This means we cannot prove that 0 is an asymptotically stable equilibrium. It actually is not asymptotically stable: if $w(t)$ happens to become 0 at any time t, the whole dynamics will stop there and will not approach $(0,0)$ any further. The origin is rather 'asymptotically stable if bounded away from $w = 0$' in a double-cone-like manner.

In a similar way, the solution of the ODE $y'(t) = -t^{-\alpha} y(t)$ will only converge to 0 if $\alpha \leqslant 1$, i.e. if the rate function does not converge to 0 too fast.

The EnKF dynamics works like this toy model: if the ensemble collapse (played by w in the first model and the rate function $t^{-\alpha}$ in the second model) happens too fast, we cannot expect convergence. We suspect that it is possible to prove that the ensemble collapse can be bounded from below (in contrast to also being bounded from above by virtue of theorem 4.5) and this is the subject of ongoing work. However, the numerical experiments suggest that the collapse happens too fast. In order to circumvent this issue of 'too quick ensemble collapse' we use artificial inflation of the covariance operator by addition of a positively definite operator (but this is gradually reduced with a certain rate). In addition to solving the problem of counterproductive ensemble collapse, variance inflation stabilizes the convergence in a very suitable manner and is used in practice for this reason, see e.g. [1, 34].

5.1. Variance inflation

In order to correct rank deficiencies of the empirical covariance operator $C(\mathfrak{r})$, we will use variance inflation in the following sense. Let $B \in \mathcal{L}(\mathbb{R}^K, \mathbb{R}^K)$ be a positive definite operator (for example the identity) and consider the equation

$$\mathrm{d}\mathfrak{r}_t^{(j)} = -\left(C(\mathfrak{r}_t) + \frac{1}{t^\alpha + R} B\right) \mathfrak{r}_t^{(j)}\, \mathrm{d}t + C(\mathfrak{r}_t)\, \mathrm{d}W_t^{(j)}, \quad \alpha \in (0,1), R > 0. \tag{18}$$

This modification gives convergence of the mapped residuals. For sufficiently small $\mathfrak{r}_t$, the new term will dominate, and for $\alpha \in (0,1)$ we then expect convergence to 0 at a rate faster than any polynomial. The question is now whether and when this asymptotic for small $\mathfrak{r}_t$ sets in.

Theorem 5.2. *Assume that y is the image of a truth $u^\dagger \in \mathcal{X}$ under A and let $\mathfrak{r}_0 = (\mathfrak{r}_0^{(j)})_{j \in \{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\mathfrak{r}_0^{(j)} : \Omega \to \mathbb{R}^K$ such that $\mathbb{E}[\frac{1}{J}\sum_{j=1}^J |\mathfrak{r}_0^{(j)}|^2] < \infty$, $B \in \mathcal{L}(\mathbb{R}^K, \mathbb{R}^K)$ a positive definite operator and $(\mathfrak{r}_t^{(j)})_{t \geqslant 0, j=1,\dots,J}$ the solution of (18). Then for all $\beta > 0$ it holds true that $\mathbb{E}[\frac{1}{J}\sum_{j=1}^J |\mathfrak{r}_t^{(j)}|^2] \in \mathcal{O}(t^{-\beta})$ and $\mathbb{E}[\frac{1}{J}\sum_{j=1}^J |\mathfrak{r}_t^{(j)}|^2]$ is monotonically decreasing.*

Proof. Let $B \in \mathcal{L}(\mathbb{R}^K, \mathbb{R}^K)$ be a positive definite operator, $\alpha \in (0,1)$, $R > 0$ and assume, that that the smallest eigenvalue of B is $\lambda_{\min} = c > 0$.

We derive an equation for $\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}_t^{(j)}|^2$ by using Itô's formula:

$$\begin{aligned}
\mathrm{d}|\mathfrak{r}_t^{(j)}|^2 = &-2\left\langle \mathfrak{r}_t^{(j)}, \left(C(\mathfrak{r}_t)+\frac{1}{t^\alpha+R}B\right)\mathfrak{r}_t^{(j)}\right\rangle \mathrm{d}t + 2\langle \mathfrak{r}_t^{(j)}, C(\mathfrak{r}_t)\mathrm{d}W_t^{(j)}\rangle \\
&+\frac{1}{J}\sum_{j=1}^{J}\left\langle \mathfrak{r}_t^{(k)}-\overline{\mathfrak{r}}_t, C(\mathfrak{r}_t)(\mathfrak{r}_t^{(k)}-\overline{\mathfrak{r}}_t)\right\rangle \mathrm{d}t.
\end{aligned}$$

Taking the empirical mean over all particles yields

$$\begin{aligned}
\mathrm{d}\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}_t^{(j)}|^2 = &-\frac{2}{J}\sum_{j=1}^{J}\left\langle \mathfrak{r}_t^{(j)}, \left(C(\mathfrak{r}_t)+\frac{1}{t^\alpha+R}B\right)\mathfrak{r}_t^{(j)}\right\rangle \mathrm{d}t + \frac{2}{J}\sum_{j=1}^{J}\langle \mathfrak{r}_t^{(j)}, C(\mathfrak{r}_t)\mathrm{d}W^{(j)}\rangle \\
&+\frac{1}{J}\sum_{k=1}^{J}\langle \mathfrak{r}_t^{(k)}-\overline{\mathfrak{r}}_t, C(\mathfrak{r}_t)(\mathfrak{r}_t^{(k)}-\overline{\mathfrak{r}}_t)\rangle \mathrm{d}t.
\end{aligned}$$

Thus, for all $t, s \geqslant 0$, it follows similarly to the proof of lemma 4.1 that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}_{t+s}^{(j)}|^2\right] \leqslant\ &\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}_s^{(j)}|^2\right] - \frac{2}{J}\int_s^{s+t}\mathbb{E}\left[\sum_{j=1}^{J}\langle \mathfrak{r}_r^{(j)}, C(\mathfrak{r}_r)\mathfrak{r}_r^{(j)}\rangle\right]\mathrm{d}r \\
&-\frac{2}{J}\int_s^{s+t}\frac{1}{r^\alpha+R}\mathbb{E}\left[\sum_{j=1}^{J}\langle \mathfrak{r}_r^{(j)}, B\mathfrak{r}_r^{(j)}\rangle\right]\mathrm{d}r \\
&+\frac{1}{J}\int_s^{s+t}\mathbb{E}\left[\sum_{j=1}^{J}\langle \mathfrak{r}_r^{(j)}-\overline{\mathfrak{r}}_r, C(\mathfrak{r}_r)(\mathfrak{r}_r^{(j)}-\overline{\mathfrak{r}}_r)\rangle\right]\mathrm{d}r \\
\leqslant\ &\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}_s^{(j)}|^2\right] - \frac{1}{J}\int_s^{s+t}\mathbb{E}\left[\sum_{j=1}^{J}\langle \mathfrak{r}_r^{(j)}, \left(C(\mathfrak{r}_r)+\frac{1}{r^\alpha+R}B\right)\mathfrak{r}_r^{(j)}\rangle\right]\mathrm{d}r,
\end{aligned}$$

where we used lemma A.2 and the non-negativity of B. This yields the monotonicity, as both the covariance $C(\mathfrak{r}_r)$ as well as B are non-negative matrices.

Now we will improve the estimate to obtain the asymptotic rate. Consider $S(t) = \frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}_t^{(j)}|^2$, then

$$\mathrm{d}(t^\beta S(t)) = \beta t^{\beta-1}S(t)\mathrm{d}t + t^\beta \mathrm{d}S(t).$$

Now we can use all the previous estimates for the terms in $\mathrm{d}S$ together with the non-negativity of the covariance matrix $C(\mathfrak{r}_t)$ and $B \geqslant \lambda_{\min} > 0$ to obtain

$$\begin{aligned}
t^\beta\mathbb{E}S(t) &\leqslant \beta\int_0^t \tau^{\beta-1}\mathbb{E}S(\tau)\mathrm{d}\tau - \frac{2}{J}\int_0^t \tau^\beta\frac{\lambda_{\min}}{\tau^\alpha+R}\mathbb{E}S(\tau)\mathrm{d}\tau \\
&\leqslant \int_0^t \tau^{\beta-1}\left[\beta - \frac{2\lambda_{\min}}{J}\frac{\tau}{\tau^\alpha+R}\right]\mathbb{E}S(\tau)\mathrm{d}\tau.
\end{aligned}$$

There is a time $T > 0$ such that the integrand in the equation above is negative for all $t > T$ and thus using the monotonicity of $\mathbb{E}S(\tau)$ we obtain for all $t > T$

$$t^{\beta}\mathbb{E}S(t) \leqslant \int_0^T \tau^{\beta-1}\Big[\beta - \frac{2\lambda_{\min}}{J}\frac{\tau}{\tau^{\alpha}+R}\Big]\mathrm{d}\tau\mathbb{E}S(0),$$

which yields the asymptotic rate $t^{-\beta}$ for $\mathbb{E}S(t)$. □

Remark 5.3. In case of a positive semidefinite matrix B, the convergence of the residuals will then take place in the image space of the matrix B. The proof can be straightforwardly generalized to this setting by projections of the quantities to the corresponding subspace.

We can also verify almost sure convergence faster than any polynomial rate.

Theorem 5.4. *Assume that y is the image of a truth $u^{\dagger} \in \mathcal{X}$ under A and let $\mathfrak{r}_0 = (\mathfrak{r}_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\mathfrak{r}_0^{(j)} : \Omega \to \mathbb{R}^K$ and $B \in \mathcal{L}(\mathbb{R}^K, \mathbb{R}^K)$ a positive definite operator. Then the solution of (18) is almost surely asymptotically stable with rate function $\rho(t) = t^{-\frac{\beta}{2}}$ for all $\beta > 0$. In particular, $(\mathfrak{r}_t^{(j)})_{j=1,\dots,J}$ converges to zero almost surely as $t \to \infty$.*

Proof. We define the Lyapunov function

$$V(\mathfrak{r}, t) = t^{\beta}\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}^{(j)}|^2$$

and obtain

$$LV(\mathfrak{r}, t) \leqslant \frac{\beta t^{\beta-1}}{J}\sum_{j=1}^{J}|\mathfrak{r}^{(j)}|^2 - t^{\beta}\frac{1}{J}\sum_{j=1}^{J}\langle\mathfrak{r}^{(j)}, \left(C(\mathfrak{r}) + \frac{1}{t^{\alpha}+R}B\right)\mathfrak{r}^{(j)}\rangle.$$

Thus,

$$LV(\mathfrak{r}, t) \leqslant \frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}^{(j)}|^2\left(\beta - \frac{\lambda_{\min}t}{t^{\alpha}+R}\right)t^{\beta-1}.$$

There is a $T > 0$ such that the bracket above is non-positive for all $t \geqslant T$. We obtain $\int_0^{\infty} LV(\mathfrak{r}, t)\,\mathrm{d}t \leqslant \int_0^T LV(\mathfrak{r}, t)\,\mathrm{d}t$. Moreover, by neglecting the negative term in the bracket for $t \leqslant T$ we obtain

$$\mathbb{E}[\int_0^T LV(\mathfrak{r}_t, t)\,\mathrm{d}t] \leqslant \mathbb{E}[\int_0^T \beta s^{\beta-1}\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{r}_s^{(j)}|^2\,\mathrm{d}s] \leqslant \frac{T^{\beta}}{J}\mathbb{E}[\sum_{j=1}^{J}|\mathfrak{r}_0^{(j)}|^2] < \infty,$$

by using the monotonicity of the sum. Hence, $\int_0^{\infty} LV(\mathfrak{r}_t, t)\,\mathrm{d}t < \infty$ and thus $\mathfrak{r}_t$ is almost surely asymptotically stable with rate function $\rho(t) = t^{-\frac{\beta}{2}}$. □

Remark 5.5. Note that the convergence rate is faster than any polynomial rate. However, the proof reveals that the constant in the convergence result will grow w.r.t. the rate β and $\alpha \in (0, 1)$, which is consistent with the numerical experiments presented in section 6.

Our aim is to use variance inflation in the parameter space, such that we can apply theorem 5.2. We will use variance inflation in the finite dimensional system of SDEs of the coordinates in the parameter space.

Let $y \in A\mathcal{S}$ where $A\mathcal{S}$ is the linear span of $\{Au_0^{(1)}, \dots, Au^{(J)}\}$ and consider the equation

$$\mathrm{d}u_t^{(j)} = (C(u_t) + \frac{1}{t^\alpha + R}B)A^*\Gamma^{-1}(y - Au_t^{(j)})\,\mathrm{d}t + C(u_t)A^*\Gamma^{-\frac{1}{2}}\,\mathrm{d}W_t^{(j)}, \quad (19)$$

$j = 1, \dots, J$, for B positive definite, $R > 0$ and $\alpha \in (0, 1)$. Since $y \in A\mathcal{S}$, the subspace property still holds, i.e. $u_t^{(j)} \in \mathcal{S}$ for all $(t,j) \in [0, \infty) \times \{1, \dots, J\}$. The following result transfers the results of theorem 5.2 to the parameter space:

Corollary 5.6. *Let $y \in A\mathcal{S}$ and assume that y is the image of a truth $u^\dagger \in \mathcal{X}$ under A, A^* is assumed to be one-to-one and let $(u_t^{(j)})_{t\geqslant 0, j=1,\dots,J}$ be the solution of (19). Then*

(i) $\lim\limits_{t\to\infty} \mathbb{E}[\frac{1}{J}\sum\limits_{j\neq 1}^{J} |\mathfrak{e}_t^{(j)}|^2] = 0.$

(ii) $\lim\limits_{t\to\infty} \mathbb{E}[\frac{1}{J}\sum\limits_{j=1}^{J} |\mathfrak{r}_t^{(j)}|^2] = 0.$

(iii) $(\mathfrak{r}_t^{(j)})_{t\geqslant 0}$ *converges almost surely to zero with rate function* $\rho(t) = t^{-\frac{\beta}{2}}$ *for all* $\beta > 0$.

Proof. Let $R > 0$ and $\alpha \in (0, 1)$ and observe

$$\mathrm{d}\mathfrak{r}_t^{(j)} = -(C(\mathfrak{r}_t) + \frac{1}{t^\alpha + R}\Gamma^{-\frac{1}{2}}AB(\Gamma^{-\frac{1}{2}}A)^*)\mathfrak{r}_t^{(j)}\,\mathrm{d}t + C(\mathfrak{r}_t)\,\mathrm{d}W_t^{(j)}.$$

Since $\Gamma^{-\frac{1}{2}}AB(\Gamma^{-\frac{1}{2}}A)^*$ is positive definite the second and third assertion follow directly from theorems 5.2 and 5.4. The proof of the first assertion is similar to the proof of theorem 4.2. □

6. Numerical results

We consider the problem of recovering the unknown data $u^\dagger$ from noise-free observations

$$y^\dagger = A(u^\dagger),$$

where $p = \mathcal{A}^{-1}(u)$ is the solution of the one-dimensional elliptic equation

$$\begin{aligned} -\frac{\mathrm{d}^2 p}{\mathrm{d}x^2} + p &= u \qquad \text{in } D := (0, \pi), \\ p &= 0 \qquad \text{on } \partial D. \end{aligned} \quad (20)$$

The forward response operator is defined by

$$A = \mathcal{O} \circ \mathcal{A}^{-1} \quad \text{with} \quad \mathcal{A} = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \mathrm{id} \quad \text{on} \quad \mathcal{D}(\mathcal{A}) = H^2 \cap H_0^1$$

and with operator $\mathcal{O}$ observing the dynamical system at $K = 2^4 - 1$ equispaced observation points $x_k = \frac{k}{2^4}$, $k = 1, \dots, K$. We approximate the forward-problem (20) numerically on a uniform mesh with meshwidth $h = 2^{-8}$ by a finite element method with continuous, piecewise linear ansatz functions.

We choose the initial ensemble of particles based on the eigenvalue and eigenfunctions $\{\lambda_j, z_j\}_{j\in\mathbb{N}}$ of the covariance operator C_0, defined by $C_0 = \beta(\mathcal{A} - \mathrm{id})^{-1}$ for $\beta = 10$.

From the Bayesian perspective we may interpret this as prior distributed by $\mu_0 = \mathcal{N}(0, C_0)$. We set our j^{th} initial particle to $u^{(j)}(0) = \sqrt{\lambda_j}\zeta_j z_j$ with $\zeta_j \sim \mathcal{N}(0,1)$, i.e. we use the Karhunen–Loève expansion to generate draws from μ_0.

The EnKF continuous time limit

$$\mathrm{d}u_t^{(j)} = C(u_t)A^*\Gamma^{-1}(y - Au_t^{(j)})\,\mathrm{d}t + C(u_t)A^*\Gamma^{-\frac{1}{2}}\,\mathrm{d}W_t^{(j)}$$

is discretized by equation (2) for the following simulations.

6.1. Ensemble collapse

In the following we illustrate the results from section 4, in particular the bounds on the ensemble collapse derived in theorems 4.2 and 4.5.

Figure 1 shows that the Monte Carlo approximation of the expected value $\hat{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_t^{(j)}|^2]$ is bounded from above by $(\frac{J+1}{J^2}t + C)^{-1}$ with $C = (\hat{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_0^{(j)}|^2])^{-1}$, as derived in theorem 4.2.

Similarly figure 2 demonstrates that the approximated higher moments $\hat{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_t^{(j)}|^p]^{-\frac{1}{p}}$ are bounded by $J^{\frac{1}{2}}(\frac{2}{p}C(p,J)J^{1-\frac{2}{p}}K^{-\frac{2}{p}}t + C)^{-\frac{1}{2}}$ with $C = (K^{\frac{p-1}{2}}\hat{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_0^{(j)}|^p])^{\frac{2}{p}}$, compare theorem 4.5.

In order to verify the almost sure ensemble collapse numerically, we have simulated $Q = 10$ paths.

From theorem 4.8 we know, that $\mathfrak{e}(t)$ converges almost surely to zero with rate function $\rho(t) = t^{-\frac{\alpha}{2}}$ for every $\alpha \in (0,1)$. Figure 3 illustrates this behavior, the expected convergence rates can be observed in this example.

6.2. Convergence to ground truth

We compare simulations of the ensemble Kalman inversion without variance inflation with simulations of the ensemble Kalman inversion with variance inflation. The variance inflation is used in the following setting: we set $\alpha \in \{\frac{1}{2}, \frac{3}{4}\}$ and $R = 1$ in equation (19). The number of particles is $J = 15$, i.e. the forward response operator is bijective as a mapping from the subspace spanned by the initial ensemble to the data space.

Figure 4 shows the differences of the EnKF estimation in the parameter space as well as in the observation space. We observe that the simulations with variance inflation giving a better estimation in the observation space as well as in the parameter space. If we reduce the variance inflation in time faster, i.e. we increase the parameter α from $\frac{1}{2}$ to $\frac{3}{4}$, the effect of the variance inflation decreases. The following figures demonstrate the effect on the ensemble collapse and the residuals.

The idea of the variance inflation was to slow down the convergence of the particles to the ensemble mean, i.e. to control the rate of the ensemble collapse, in order to ensure the convergence of the residuals in the observation space. Figure 5 illustrates that we can ensure a higher spread of the ensemble in the simulations with variance inflation in comparison to the simulations without variance inflation in the observation space.

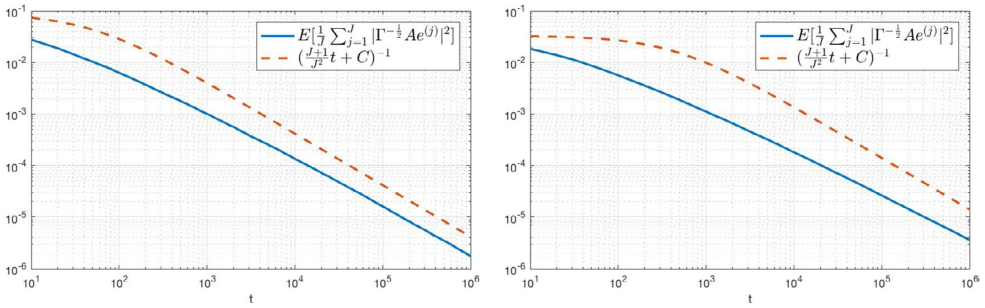

Figure 1. $\hat{E}(\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}^{(j)}(t)|^2)$ with w.r. of time. $Q=1000$ paths with $J=5$ (left) and $J=15$ (right) particles has been simulated.

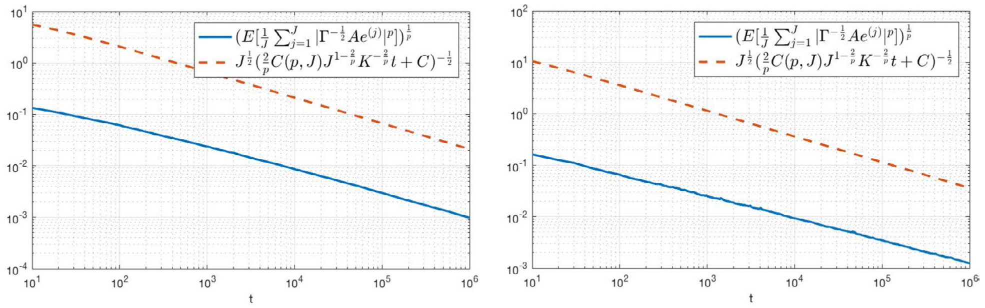

Figure 2. $\hat{E}(\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}^{(j)}(t)|^p)^{-\frac{1}{p}}$, $p=\lfloor\frac{J+3}{2}\rfloor-1$, w.r. of time. $Q=1000$ paths with $J=5$ (left) and $J=15$ (right) particles has been simulated.

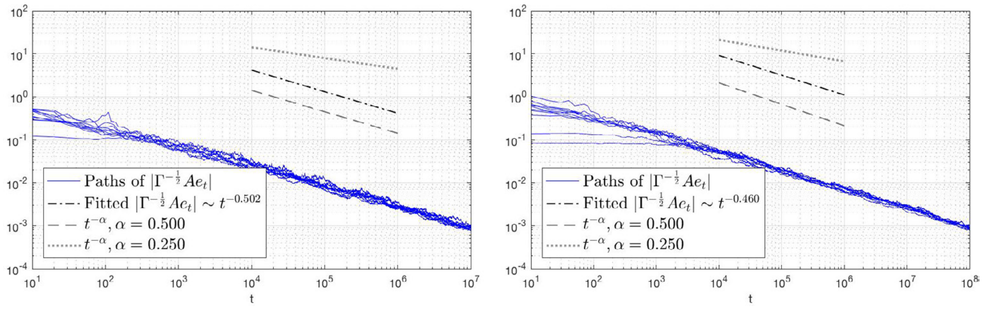

Figure 3. Paths of $|\mathfrak{e}(t)|_2$ w.r. of time. $Q=10$ paths with $J=5$ (left) and $J=15$ (right) particles has been simulated.

Figure 6 points out that we end up with convergence of the residuals in the observation and parameter space in case of variance inflation. Without variance inflation the simulations show a slight increase of the residuals in the parameter space, suggesting that the convergence of the residuals will slow down in the observation space as well.

To emphasize this result, we reduce the dimension of the example and we set $h=2^4$ with $K=3$ equispaced observation points. Furthermore, we set again $R=1$ and $\alpha=\frac{1}{2}$ and we use

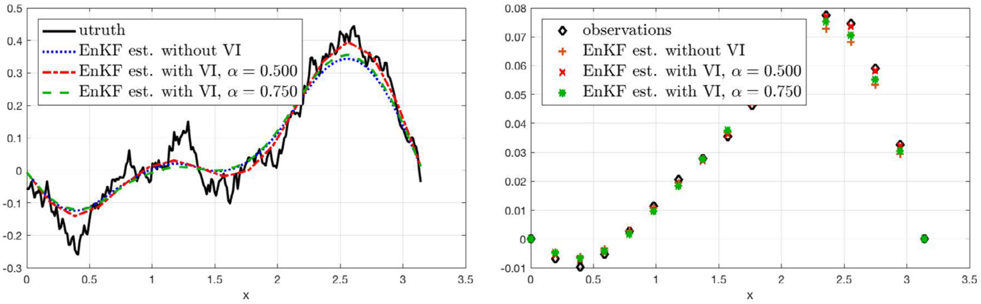

Figure 4. EnKF estimation without VI versus EnKF estimation with VI. $J = 15$ particles and $Q = 1000$ paths has been simulated.

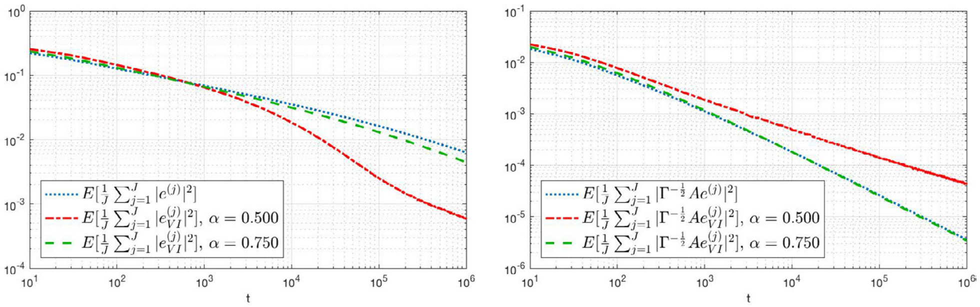

Figure 5. Comparison of the spread of the ensemble w.r. to time with VI and without VI.

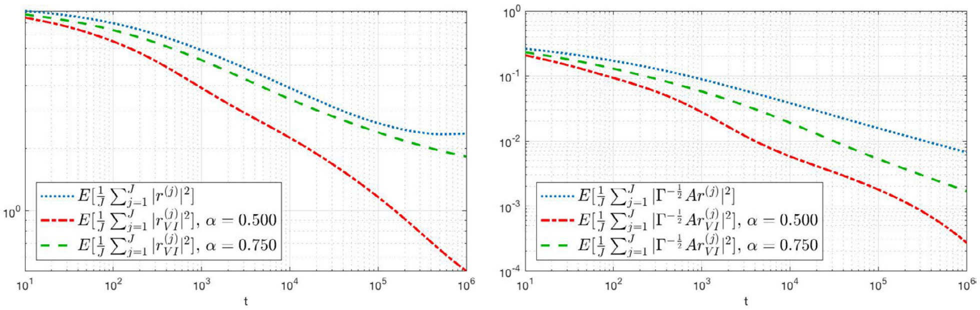

Figure 6. Comparison of the residuals w.r. to time with VI and without VI.

$J = 3$ particles, such that the forward response operator is again bijective as mapping from the subspace spanned by the initial ensemble to the observation space.

Figure 7 shows again the difference of the EnKF estimation with and without variance inflation.

Figure 8 points out the effect of the variance inflation. While the residuals in the observation space without variance inflation diverge, we obtain convergence of the residuals in the observation space using variance inflation. In addition, in figure 9 we can see that the ensemble of particles still collapse in the parameter space as well as in the observation space.

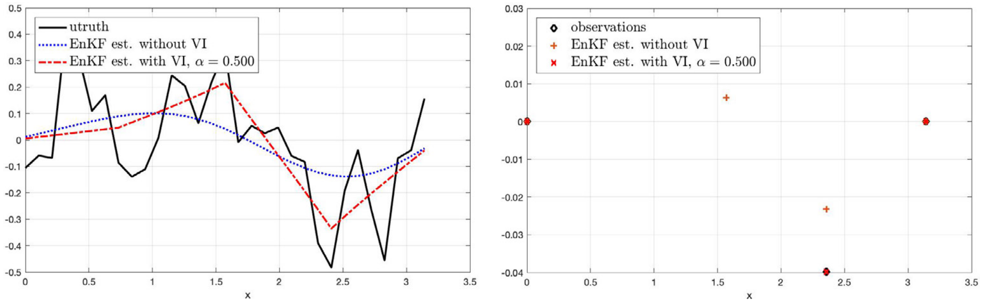

Figure 7. EnKF estimation without VI versus EnKF estimation with VI. $J = 3$ particles and $Q = 10\,000$ paths has been simulated.

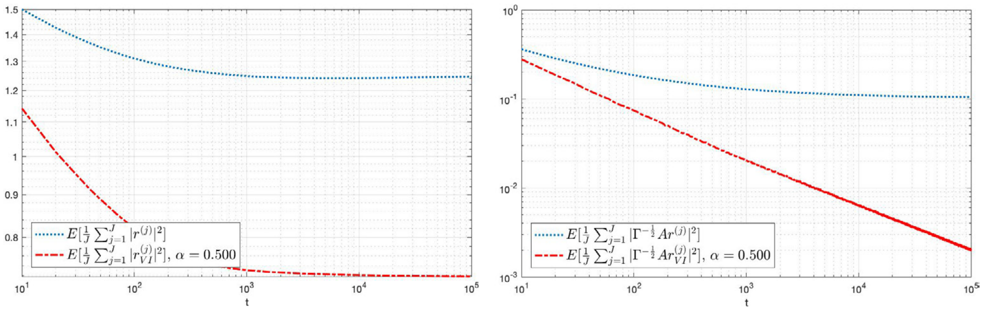

Figure 8. Comparison of the residuals w.r. to time with VI and without VI.

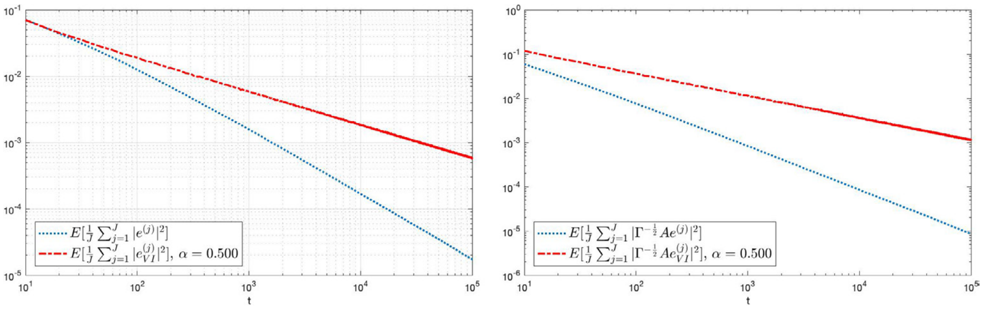

Figure 9. Comparison of the ensemble spread w.r. to time with VI and without VI.

7. Conclusions

Our analysis of the ensemble Kalman inversion shows the well-posedness and accuracy of the method in the case of linear forward operators. The results are based on the continuous time limit of the algorithm consisting of a coupled system of stochastic differential equations. Due to the subspace property of the ensemble Kalman inversion, the theory of finite-dimensional stochastic differential equations could be applied to establish existence and uniqueness of solutions, i.e. to show the well-posedness of the method. The ensemble collapse has been quantified in terms of moments as well as almost sure convergence of the particles to the empirical mean. Furthermore, we suggest a time-adaptive variance inflation to stabilize the convergence of the empirical mean to the truth in the noise free case. The inflation can be interpreted as model error delaying the ensemble collapse. The presented numerical experiments confirm the theoretical results and indicate that the ensemble collapse can be bounded from below for the original iteration scheme without variance inflation. However, the rate seems to be too small to achieve convergence. This will be subject to future work. In addition, the next steps include the generalization of the presented results to case of noisy observations in the inverse problem and the development of appropriate stopping criteria in the noisy case. Even though the presented analysis relies on the linearity of the forward operator, the statements hold true for non-Gaussian priors and can guide the analysis of the nonlinear setting.

Acknowledgments

ClS would like to thank the Isaac Newton Institute for Mathematical Sciences for support and hospitality during the programme Uncertainty quantification for complex systems: theory and methodologies when work on this paper was undertaken. This work was supported by: EPSRC Grant Nos. EP/K032208/1 and EP/R014604/1. SW is grateful to the DFG RTG1953 'Statistical Modeling of Complex Systems and Processes' for funding of this research. The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

Appendix A. Auxiliary results

In order to use Itô's formula we have to calculate the following quadratic covariation in many cases:

Lemma A.1. *Let* $(W^{(j)})_{j=1,\dots,J}$ *be independent Brownian motions in* $\mathbb{R}^K$, $u, v \in \mathbb{R}^K$ *and let* $l \neq j \in \{1, \dots, J\}$. *Then with* $\bar{W} = \frac{1}{J}\sum_{k=1}^{J} W^{(k)}$,

$$\langle u, \mathrm{d}(W^{(j)} - \overline{W})\rangle\langle v, \mathrm{d}(W^{(j)} - \overline{W})\rangle = \frac{J-1}{J}\langle u, v\rangle\, \mathrm{d}t,$$

$$\langle u, \mathrm{d}(W^{(j)} - \overline{W})\rangle\langle v, \mathrm{d}(W^{(l)} - \overline{W})\rangle = -\frac{1}{J}\langle u, v\rangle\, \mathrm{d}t.$$

Proof. Observe

$$W^{(j)} - \overline{W} = -\frac{1}{J}\sum_{k=1,k\neq j}^{J} W^{(k)} + \frac{J-1}{J}W^{(j)}.$$

Since $W^{(k)}$ are independent Brownian motions it follows

$$\begin{aligned}\langle u, \mathrm{d}(W^{(j)} - \overline{W})\rangle\langle v, \mathrm{d}(W^{(j)} - \overline{W})\rangle &= \frac{1}{J^2}\sum_{k=1,k\neq j}^{J}\langle u, \mathrm{d}W^{(k)}\rangle\langle v, \mathrm{d}W^{(k)}\rangle \\ &\quad + \frac{(J-1)^2}{J^2}\langle u, \mathrm{d}W^{(j)}\rangle\langle v, \mathrm{d}W^{(j)}\rangle \\ &= \frac{J-1}{J}\langle u, v\rangle\, \mathrm{d}t.\end{aligned}$$

Similarly,

$$\begin{aligned}\langle u, \mathrm{d}(W^{(j)} - \overline{W})\rangle\langle v, \mathrm{d}(W^{(l)} - \overline{W})\rangle &= -\frac{1}{J}\sum_{k=1}^{J}(\langle u, \mathrm{d}W^{(j)}\rangle\langle v, \mathrm{d}W^{(k)}\rangle + \langle u, \mathrm{d}W^{(k)}\rangle\langle v, \mathrm{d}W^{(l)}\rangle)\\ &\quad + \frac{1}{J^2}\sum_{i,k=1}^{J}\langle u, \mathrm{d}W^{(i)}\rangle\langle v, \mathrm{d}W^{(k)}\rangle\\ &= -\frac{1}{J}\langle u, v\rangle\,\mathrm{d}t.\end{aligned}$$

□

Lemma A.2. *Let M be a symmetric and nonnegative $d \times d$-matrix, then for all choices of vectors $(z^{(k)})_{k=1,\dots,J}$ in $\mathbb{R}^n$ we have*

$$\sum_{k,l=1}^{J}\langle z^{(k)}, z^{(l)}\rangle\langle z^{(k)}, Mz^{(l)}\rangle \geqslant 0.$$

Proof. Let $(v^{(m)})_{m=1,\dots,d}$ be an orthonormal basis of eigenvectors such that $Mv^{(m)} = \lambda_m v^{(m)}$ with $\lambda_m \geqslant 0$. Then $z^{(l)} = \sum_{m=1}^{d} z_m^{(l)} v^{(m)}$ and thus

$$\sum_{k,l=1}^{J}\langle z^{(k)}, z^{(l)}\rangle\langle z^{(k)}, Mz^{(l)}\rangle = \sum_{k,l=1}^{J}\sum_{m,n=1}^{d} z_n(k) z_n^{(l)} z_m^{(k)} z_m^{(l)}\lambda_m = \sum_{n,m=1}^{d}\lambda_m(\sum_{k=1}^{J} z_n^{(k)} z_m^{(k)})^2 \geqslant 0.$$

□

Lemma A.3. *Let $(x^{(j)})_{j=1,\dots,J}$ be vectors in $\mathbb{R}^n$ and let $C(x)$ denote the sample covariance matrix*

$$C(x) = \frac{1}{J}\sum_{k=1}^{J}(x^{(k)} - \overline{x}) \otimes (x^{(k)} - \overline{x}), \qquad \overline{x} = \frac{1}{J}\sum_{j=1}^{J} x^{(j)}.$$

Then it holds true that

$$\sum_{j=1}^{J}\langle x^{(j)} - \overline{x}, C(x)(x^{(j)} - \overline{x})\rangle \leqslant \sum_{j=1}^{J}\langle x^{(j)}, C(x)x^{(j)}\rangle$$

Proof. By expanding the non-centered quadratic form we obtain

$$\sum_{j=1}^{J}\langle x^{(j)} - \overline{x}, C(x)(x^{(j)} - \overline{x})\rangle = \sum_{j=1}^{J}\langle x^{(j)}, C(x)x^{(j)}\rangle - J\langle \overline{x}, C(x)\overline{x}\rangle,$$

which yields the claim by the non-negativity of the covariance matrix. □

Lemma A.4. *For all $j \in \{1, \dots, J\}$ the process*

$$(M(t))_{t\geqslant 0} := \Big(\int_0^t \mathfrak{e}_s^{(j)T} C(\mathfrak{e}_s)\,\mathrm{d}W_s^{(j)}\Big)_{t\geqslant 0}$$

is a (global) martingale.

Proof. The local martingale given by the stochastic integral is a true martingale by Itô-isometry if we show that following second moment is finite (see [35, theorem 2.4])

$$\|\mathfrak{e}_\cdot^{(j)T}C(\mathfrak{e}_\cdot)\|_{\Lambda_2;T} := \mathbb{E}[\int_0^T \|\mathfrak{e}_s^{(j)T}C(\mathfrak{e}_s)\|_F^2\,\mathrm{d}s] = \int_0^T \mathbb{E}[\|\mathfrak{e}_s^{(j)T}C(\mathfrak{e}_s)\|_F^2]\,\mathrm{d}s < \infty$$

for all $T \geqslant 0$. For this, we first estimate the Frobenius norm by

$$\begin{aligned}\|\mathfrak{e}_s^{(j)T}C(\mathfrak{e}_s)\|_F^2 := \mathrm{trace}\mathfrak{e}^{(j)T}C(\mathfrak{e})(\mathfrak{e}^{(j)T}C(\mathfrak{e}))^T &= \frac{1}{J^2}\sum_{k,l=1}^J \langle\mathfrak{e}^{(l)},\mathfrak{e}^{(k)}\rangle\langle\mathfrak{e}^{(j)},\mathfrak{e}^{(k)}\rangle\langle\mathfrak{e}^{(l)},\mathfrak{e}^{(j)}\rangle \\ &\leqslant \frac{1}{J^2}\sum_{k,l=1}^J |\mathfrak{e}^{(l)}|^2|\mathfrak{e}^{(j)}|^2|\mathfrak{e}^{(k)}|^2.\end{aligned}$$

Thus, it holds true that

$$\frac{1}{J}\sum_{j=1}^J \|\mathfrak{e}_s^{(j)T}C(\mathfrak{e}_s)\|_F^2 \leqslant \frac{1}{J^3}\sum_{j,k,l=1}^J |\mathfrak{e}^{(l)}|^2|\mathfrak{e}^{(j)}|^2|\mathfrak{e}^{(k)}|^2 = (\frac{1}{J}\sum_{j=1}^J |e^{(j)}|^2)^3 \leqslant \frac{1}{J}\sum_{j=1}^J |e^{(j)}|^6$$

and with lemma 4.1 it follows

$$\frac{1}{J}\sum_{j=1}^J \|\mathfrak{e}_\cdot^{(j)T}C(\mathfrak{e}_\cdot)\|_{\Lambda_2;T} \leqslant \int_0^T \mathbb{E}[\frac{1}{J}\sum_{j=1}^J |e^{(j)}|^6]\,\mathrm{d}s \leqslant C,$$

since $p + 2 := 6 \leqslant J + 4$. □

Lemma A.5. *For all $k \in \{1, \dots, J\}$ and $p \in (2, \frac{J+3}{2})$ the process*

$$(M(t))_{t\geqslant 0} := \left(\int_0^t \frac{p}{J^2}\sum_{m=1}^K ((\sum_{k=1}^J |\mathfrak{e}_m^{(k)}|^2)^{\frac{p}{2}-1}\sum_{j,l=1}^J \mathfrak{e}_m^{(l)}\mathfrak{e}_m^{(j)})\mathfrak{e}^{(l)\top}\,\mathrm{d}W^{(k)}\right)$$

is a (global) martingale.

Proof. Similarly to the proof of lemma A.4 we estimate the Frobenius norm of the integrand by

$$\begin{aligned}\|\sum_{m=1}^K ((\sum_{k=1}^J |\mathfrak{e}_m^{(k)}|^2)^{\frac{p}{2}-1}\sum_{j,l=1}^J \mathfrak{e}_m^{(l)}\mathfrak{e}_m^{(j)})\mathfrak{e}^{(l)\top}\|_F^2 &\leqslant C_1(J)\sum_{m=1}^K (\sum_{k=1}^J |\mathfrak{e}_m^{(k)}|^2)^{p-2}\sum_{j,l=1}^J (\mathfrak{e}_m^{(l)})^2(\mathfrak{e}_m^{(j)})^2|\mathfrak{e}^{(l)}|^2 \\ &\leqslant C_2(J,K)\sum_{k=1}^J |\mathfrak{e}^{(k)}|^{2(p-2)}\sum_{j,l=1}^J |\mathfrak{e}^{(l)}|^4|\mathfrak{e}^{(j)}|^2 \\ &\leqslant C_3(J,K)\sum_{l=1}^J |\mathfrak{e}^{(l)}|^{2p+2},\end{aligned}$$

where we have used Jensen's inequality and the fact $|\mathfrak{e}_m^{(j)}|^2 \leqslant \sum_{n=1}^K |\mathfrak{e}_n^{(j)}|^2 = |\mathfrak{e}^{(j)}|^2$. The assertion follows by the bound (B.2) in the proof of theorem 4.5, which we obtained by localization and Fatou's lemma without martingale property. □

Appendix B. Higher-order ensemble collapse: proof of theorem 4.5

We will use the following auxiliary result in order to prove theorem 4.5. It is a well known statement of the equivalence of norms, but we need the precise constants.

Lemma B.1. *For $a_{m,j} \in \mathbb{R}$, $m = 1, \dots, d$, $j = 1, \dots, J$ and $p \in \mathbb{N}$,*

$$\sum_{j=1}^{J} (\sum_{m=1}^{d} |a_{m,j}|^2)^{\frac{p}{2}} \leqslant d^{(p-1)/2} \cdot \sum_{m=1}^{d} \sum_{j=1}^{J} |a_{m,j}|^p$$

and

$$\sum_{m=1}^{d} \sum_{j=1}^{J} |a_{m,j}|^p \leqslant J^{p/2} \cdot \sum_{m=1}^{d} (\sum_{j=1}^{J} |a_{m,j}|^2)^{\frac{p}{2}}.$$

By symmetry we also have

$$\sum_{m=1}^{d} (\sum_{j=1}^{J} |a_{m,j}|^2)^{\frac{p}{2}} \leqslant J^{p/2} \cdot \sum_{m=1}^{d} \sum_{j=1}^{J} |a_{m,j}|^p \quad \text{and} \quad \sum_{m=1}^{d} \sum_{j=1}^{J} |a_{m,j}|^p \leqslant d^{(p-1)/2} \cdot \sum_{j=1}^{J} (\sum_{m=1}^{d} |a_{m,j}|^2)^{\frac{p}{2}}.$$

Proof. We start with the first claim and write

$$\sum_{j=1}^{J} (\sum_{m=1}^{d} |a_{m,j}|^2)^{\frac{p}{2}} = \sum_{j=1}^{J} T_j$$

with $T_j^2 = (\sum_{m=1}^d |a_{m,j}|^2)^p$. We continue by expressing T_j^2 using the multinomial theorem and Young's inequality

$$\begin{aligned} T_j^2 &= \sum_{k_1+\cdots+k_d=p} \binom{p}{k_1, \dots, k_d} \cdot \prod_{m=1}^{d} |a_{m,j}|^{2 \cdot k_m} \\ &= \sum_{k_1+\cdots+k_d=p} \binom{p}{k_1, \dots, k_d} \cdot \prod_{m=1, k_m \neq 0}^{d} |a_{m,j}|^{2 \cdot k_m} \\ &\leqslant \sum_{m=1}^{d} |a_{m,j}|^{2p} \cdot \sum_{l_1+\cdots+l_d=p-1} \binom{p-1}{l_1, \dots, l_d} = \sum_{m=1}^{d} |a_{m,j}|^{2p} \cdot d^{p-1}. \end{aligned}$$

This means that

$$\sum_{j=1}^{J} (\sum_{m=1}^{d} |a_{m,j}|^2)^{\frac{p}{2}} \leqslant d^{\frac{p-1}{2}} \cdot \sum_{j=1}^{J} \sqrt{\sum_{m=1}^{d} |a_{m,j}|^{2p}} \leqslant d^{\frac{p-1}{2}} \cdot \sum_{j=1}^{J} \sum_{m=1}^{d} |a_{m,j}|^p,$$

which proves the first statement. For the second claim we can write by concavity of the square root

$$\sum_{m=1}^{d}(\sum_{j=1}^{J}|a_{m,j}|^2)^{\frac{p}{2}} = \sum_{m=1}^{d}(\sqrt{J}\cdot\sqrt{\sum_{j=1}^{J}\frac{|a_{m,j}|^2}{J}})^p \geqslant J^{-\frac{p}{2}}\sum_{m=1}^{d}\sum_{j=1}^{J}|a_{m,j}|^p,$$

i.e.

$$\sum_{m=1}^{d}\sum_{j=1}^{J}|a_{m,j}|^p \leqslant J^{\frac{p}{2}}\cdot\sum_{m=1}^{d}(\sum_{j=1}^{J}|a_{m,j}|^2)^{\frac{p}{2}}.$$

□

Proof of theorem 4.5. Recall the equation of $\mathfrak{e}^{(j)}$

$$\mathrm{d}\mathfrak{e}^{(j)} = -\frac{1}{J}\sum_{l=1}^{J}\mathfrak{e}^{(l)}\langle\mathfrak{e}^{(l)},\mathfrak{e}^{(j)}\rangle\mathrm{d}t + \frac{1}{J}\sum_{l=1}^{J}\mathfrak{e}^{(l)}\langle\mathfrak{e}^{(l)},\mathrm{d}(W^{(j)}-\overline{W})\rangle.$$

And (recall that $\mathfrak{e}^{(j)}\in\mathbb{R}^K$) componentwise

$$\mathrm{d}\mathfrak{e}_m^{(j)} = -\frac{1}{J}\sum_{l=1}^{J}\mathfrak{e}_m^{(l)}\langle\mathfrak{e}^{(l)},\mathfrak{e}^{(j)}\rangle\mathrm{d}t + \frac{1}{J}\sum_{l=1}^{J}\mathfrak{e}_m^{(l)}\langle\mathfrak{e}^{(l)},\mathrm{d}(W^{(j)}-\overline{W})\rangle.$$

We define the Lyapunov function (for equivalent notions of 'p-norms' of the ensemble, see lemma B.1)

$$V_p(\mathfrak{e}) = \frac{1}{J}\sum_{m=1}^{K}(\sum_{j=1}^{J}|\mathfrak{e}_m^{(j)}|^2)^{\frac{p}{2}}$$

and according to Ito's lemma it holds that

$$\mathrm{d}V_p(\mathfrak{e}) = \sum_{m=1}^{K}\sum_{j=1}^{J}\frac{\partial V_p}{\partial\mathfrak{e}_m^{(j)}}\mathrm{d}\mathfrak{e}_m^{(j)} + \frac{1}{2}\sum_{m,m'=1}^{K}\sum_{j,j'=1}^{J}\mathrm{d}\mathfrak{e}_m^{(j)}\frac{\partial^2 V_p}{\partial\mathfrak{e}_m^{(j)}\partial\mathfrak{e}_{m'}^{(j')}}\mathrm{d}\mathfrak{e}_{m'}^{(j')}.$$

Analogously to the proof of theorem 4.2 the expectation is given by

$$\begin{aligned}\mathbb{E}[V_p(\mathfrak{e}_{s+t})] =& \mathbb{E}[V_p(\mathfrak{e}_s)]\\ &- C(p,J)\mathbb{E}[\int_s^{s+t}\sum_{m=1}^{K}[\{(\sum_{k=1}^{J}|\mathfrak{e}_m^{(k)}|^2)^{\frac{p}{2}-1}\}[\sum_{n=1}^{K}(\sum_{l=1}^{J}\mathfrak{e}_m^{(l)}\mathfrak{e}_n^{(l)})^2]]\,\mathrm{d}r]\\ &+\mathbb{E}[\int_0^t\frac{p}{J^2}\sum_{m=1}^{K}((\sum_{k=1}^{J}|\mathfrak{e}_m^{(k)}|^2)^{\frac{p}{2}-1}\sum_{j,l=1}^{J}\mathfrak{e}_m^{(l)}\mathfrak{e}_m^{(j)}\langle\mathfrak{e}^{(l)},\mathrm{d}(W^{(j)}-\frac{1}{J}\sum_{r=1}^{J}W^{(r)})\rangle)]\end{aligned} \tag{B.1}$$

by defining $C(p,J) := \frac{p}{J^2}(1-\frac{(p-2+J)\cdot(J-1)}{2J^2}-\frac{p-2}{2J^2})$.

Thus, similarly to lemma 4.1 we obtain by setting $s=0$ and using Fatou's lemma

$$\mathbb{E}[V_p(\mathfrak{e}_0)] \geqslant C(p,J)\mathbb{E}[\int_0^t\sum_{m=1}^{K}[\{(\sum_{k=1}^{J}|\mathfrak{e}_m^{(k)}|^2)^{\frac{p}{2}-1}\}[\sum_{n=1}^{K}(\sum_{l=1}^{J}\mathfrak{e}_m^{(l)}\mathfrak{e}_n^{(l)})^2]]\,\mathrm{d}s].$$

Note that

$$\mathbb{E}[\int_0^t \sum_{m=1}^K [\{(\sum_{k=1}^J |\mathfrak{e}_m^{(k)}|^2)^{\frac{p}{2}-1}\}[\sum_{n=1}^K (\sum_{l=1}^J \mathfrak{e}_m^{(l)} \mathfrak{e}_n^{(l)})^2]]\,\mathrm{d}s] < C.$$

Now we bound the integrand by below by

$$\sum_{m=1}^K ((\sum_{k=1}^J |\mathfrak{e}_m^{(k)}|^2)^{\frac{p}{2}-1})(\sum_{n=1}^K (\sum_{l=1}^J \mathfrak{e}_m^{(l)} \mathfrak{e}_n^{(l)})^2) \geqslant \sum_{m=1}^K (\sum_{k=1}^J |\mathfrak{e}_m^{(k)}|^2)^{\frac{p}{2}+1} = JV_{p+2}(\mathfrak{e}).$$

Thus, we also have

$$\mathbb{E}[\int_0^t V_{p+2}(\mathfrak{e}_s)\,\mathrm{d}s] < C \tag{B.2}$$

for all $p < J+3$.

Note, that with (B.2) one can prove similar to lemma A.4, that the stochastic integral

$$\int_0^t \frac{p}{J^2} \sum_{m=1}^K ((\sum_{k=1}^J |\mathfrak{e}_m^{(k)}|^2)^{\frac{p}{2}-1} \sum_{j,l=1}^J \mathfrak{e}_m^{(l)} \mathfrak{e}_m^{(j)} \langle \mathfrak{e}^{(l)}, \mathrm{d}(W^{(j)} - \frac{1}{J}\sum_{r=1}^J W^{(r)})\rangle)$$

is a martingale for all $p \in (2, \frac{J+3}{2})$. For details see lemma A.5.

By (B.1) we get that $\mathbb{E}[V_p(\mathfrak{e}_t)]$ is monotonically decreasing and it follows

$$\mathbb{E}[V_p(\mathfrak{e}_t)] \leqslant \mathbb{E}[V_p(\mathfrak{e}_0)] - C(p,J)J\int_0^t \mathbb{E}[V_{p+2}(\mathfrak{e}_s)]\,\mathrm{d}s.$$

By Jensen's inequality it follows

$$V_{p+2}(\mathfrak{e}) = \frac{1}{J}\sum_{m=1}^K (\sum_{j=1}^J |\mathfrak{e}_m^{(j)}|^2)^{\frac{p}{2}\frac{p+2}{p}} \geqslant K^{-\frac{2}{p}} J^{-\frac{2}{p}} (V_p(\mathfrak{e}))^{\frac{p+2}{p}}$$

and we obtain

$$\mathbb{E}[V_p(\mathfrak{e}_t)] \leqslant \mathbb{E}[V_p(\mathfrak{e}_0)] - C(p,J)J^{1-\frac{2}{p}}K^{-\frac{2}{p}}\int_0^t \mathbb{E}[V_p(\mathfrak{e}_s)]^{\frac{p+2}{p}}\,\mathrm{d}s.$$

Similarly to the proof of theorem 4.2 we get

$$h' \leqslant -C(p,J)J^{1-\frac{2}{p}}K^{-\frac{2}{p}}h^{\frac{p+2}{p}},$$

by defining $h(t) := \mathbb{E}[V_p(\mathfrak{e}_t)]$, from which it follows that

$$h(t) \leqslant (\frac{2}{p}C(p,J)K^{-\frac{2}{p}}J^{1-\frac{2}{p}}t + (h(0))^{-\frac{2}{p}})^{-\frac{p}{2}}.$$

Finally, we conclude with

$$\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_t^{(j)}|^p] \leqslant J^{\frac{p}{2}}(\frac{2}{p}C(p,J)K^{-\frac{2}{p}}J^{1-\frac{2}{p}}t+(K^{\frac{p-1}{2}}\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}|\mathfrak{e}_0^{(j)}|^p])^{-\frac{2}{p}})^{-\frac{p}{2}}$$

by using lemma B.1. □

ORCID iDs

Philipp Wacker https://orcid.org/0000-0001-8718-4313
Simon Weissmann https://orcid.org/0000-0002-5111-6658

References

[1] Evensen G 2003 The ensemble Kalman filter: theoretical formulation and practical implementation *Ocean Dyn.* **53** 343–67
[2] Oliver D S, Reynolds A C and Liu N 2008 *Inverse Theory for Petroleum Reservoir Characterization and History Matching* (Cambridge: Cambridge University Press)
[3] Schneider T, Lan S, Stuart A and Teixeira J 2017 Earth system modeling 2.0: a blueprint for models that learn from observations and targeted high-resolution simulations *Geophys. Res. Lett.* **44** 12
[4] Hu J, Fennel K, Mattern J P and Wilkin J 2012 Data assimilation with a local ensemble Kalman filter applied to a three-dimensional biological model of the middle atlantic bight *J. Mar. Syst.* **94** 145–56
[5] Butala M D, Frazin R A, Chen Y and Kamalabadi F 2009 Tomographic imaging of dynamic objects with the ensemble Kalman filter *IEEE Trans. Image Process.* **18** 1573–87
[6] Simon L D, Iglesias M, Jones B and Wood C 2018 Quantifying uncertainty in thermophysical properties of walls by means of bayesian inversion *Energy Build.* **177** 220–45
[7] Iglesias M, Park M and Tretyakov M V 2018 Bayesian inversion in resin transfer molding *Inverse Problems* **34** 105002
[8] Kovachki N and Stuart A M 2018 Ensemble Kalman inversion: a derivative-free technique for machine learning tasks (arXiv:1808.03620)
[9] Le Gland F, Monbet V and Tran V-D 2009 Large sample asymptotics for the ensemble Kalman filter *Research Report* RR-7014, INRIA (https://hal.inria.fr/inria-00409060)
[10] Kwiatkowski E and Mandel J 2015 Convergence of the square root ensemble Kalman filter in the large ensemble limit *SIAM/ASA J. Uncertain. Quantification* **3** 1–17
[11] Law K, Tembine H and Tempone R 2016 Deterministic mean-field ensemble Kalman filtering *SIAM J. Sci. Comput.* **38** A1251–79
[12] Hoel H, Law K and Tempone R 2016 Multilevel ensemble Kalman filtering *SIAM J. Numer. Anal.* **54** 1813–39
[13] Chernov A, Hoel H, Law K, Nobile F and Tempone R 2018 Multilevel ensemble Kalman filtering for spatially extended models *SIAM J. Numer. Anal.* **54** 1813–39
[14] Kelly D, Law K and Stuart A M 2014 Well-posedness and accuracy of the ensemble Kalman filter in discrete and continuous time *Nonlinearity* **27** 2579
[15] Tong X, Majda A and Kelly D 2016 Nonlinear stability of the ensemble Kalman filter with adaptive covariance inflation *Commun. Math. Sci.* **14** 1283–313
[16] Kelly D, Majda A J and Tong X T 2016 Nonlinear stability and ergodicity of ensemble based Kalman filters *Nonlinearity* **29** 657
[17] Majda A J and Tong X T 2018 Performance of ensemble Kalman filters in large dimensions *Commun. Pure Appl. Math.* **71** 892–937
[18] Tong X T 2018 Performance analysis of local ensemble Kalman filter *J. Nonlinear Sci.* **28** 1397–442
[19] Del Moral P and Tugaut J 2018 On the stability and the uniform propagation of chaos properties of ensemble Kalman Bucy filters *Ann. Appl. Probab.* **28** 790–850

[20] de Wiljes J, Reich S and Stannat W 2018 Long-time stability and accuracy of the ensemble Kalman–Bucy filter for fully observed processes and small measurement noise *SIAM J. Appl. Dyn. Syst.* **17** 1152–81

[21] Ernst O G, Sprungk B and Starkloff H-J 2015 Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems *SIAM/ASA J. Uncertain. Quantification* **3** 823–51

[22] Schillings C and Stuart A M 2017 Analysis of the ensemble Kalman filter for inverse problems *SIAM J. Numer. Anal.* **55** 1264–90

[23] Bergemann K and Reich S 2010 A localization technique for ensemble Kalman filters *Q. J. R. Meteorol. Soc.* **136** 701–7

[24] Bergemann K and Reich S 2010 A mollified ensemble Kalman filter *Q. J. R. Meteorol. Soc.* **136** 1636–43

[25] Reich S 2011 A dynamical systems framework for intermittent data assimilation *BIT Numer. Math.* **51** 235–49

[26] Iglesias M A 2015 Iterative regularization for ensemble data assimilation in reservoir models *Comput. Geosci.* **19** 177–212

[27] Iglesias M A 2016 A regularizing iterative ensemble Kalman method for PDE-constrained inverse problems *Inverse Problems* **32** 025002

[28] Blömker D, Schillings C and Wacker P 2018 A strongly convergent numerical scheme from ensemble kalman inversion *SIAM J. Numer. Anal.* **56** 2537–62

[29] Schillings C and Stuart A M 2018 Convergence analysis of ensemble Kalman inversion: the linear, noisy case *Appl. Anal.* **97** 107–23

[30] Zhang Y, Liu N and Oliver D 2010 Ensemble filter methods with perturbed observations applied to nonlinear problems *Comput. Geosci.* **14**

[31] Iglesias M A, Law K and Stuart A M 2013 Ensemble Kalman methods for inverse problems *Inverse Problems* **29** 045001

[32] Khasminskii R Z 1980 *Stochastic Stability of Differential Equations* (*Monographs and Textbooks on Mechanics of Solids and Fluids. Mechanics: Analysis* vol 7) (Alphen aan den Rijn, The Netherlands) (Heidelberg: Springer)

[33] Mao X 2008 *Stochastic Differential Equations and Applications* (*Horwood Series in Mathematics & Applications*) (Chichester: Horwood Pub.)

[34] Kelly D, Majda A J and Tong X T 2016 Nonlinear stability of the ensemble Kalman filter with adaptive covariance inflation *Commun. Math. Sci.* **14** 1283–313

[35] Gawarecki L 2011 *Stochastic Differential Equations in Infinite Dimensions with Applications to Stochastic Partial Differential Equations* (*Probability and its Applications*) (Berlin: Springer)

[36] Law K, Stuart A M and Zygalakis K 2016 *Data Assimilation: a Mathematical Introduction* (*Texts in Applied Mathematics*) (Berlin: Springer)