Continuous time limit of the stochastic ensemble Kalman inversion: strong convergence analysis

Dirk Blömker, Claudia Schillings, Philipp Wacker, Simon Weissmann

Angaben zur Veröffentlichung / Publication details:

Blömker, Dirk, Claudia Schillings, Philipp Wacker, and Simon Weissmann. 2022. "Continuous time limit of the stochastic ensemble Kalman inversion: strong convergence analysis." *SIAM Journal on Numerical Analysis* 60 (6): 3181–3215. https://doi.org/10.1137/21M1437561.

SIAM J. NUMER. ANAL.
Vol. 60, No. 6, pp. 3181-3215

CONTINUOUS TIME LIMIT OF THE STOCHASTIC ENSEMBLE KALMAN INVERSION: STRONG CONVERGENCE ANALYSIS*

DIRK BLÖMKER†, CLAUDIA SCHILLINGS‡, PHILIPP WACKER‡, AND SIMON WEISSMANN§

Abstract. The ensemble Kalman inversion (EKI) method is a method for the estimation of unknown parameters in the context of (Bayesian) inverse problems. The method approximates the underlying measure by an ensemble of particles and iteratively applies the ensemble Kalman update to evolve (the approximation of the) prior into the posterior measure. For the convergence analysis of the EKI it is common practice to derive a continuous version, replacing the iteration with a stochastic differential equation. In this paper we validate this approach by showing that the stochastic EKI iteration converges to paths of the continuous time stochastic differential equation by considering both the nonlinear and linear setting, and we prove convergence in probability for the former and convergence in moments for the latter. The methods employed do not rely on the specific structure of the ensemble Kalman method and can also be applied to the analysis of more general numerical schemes for stochastic differential equations.

Key words. Bayesian inverse problems, ensemble Kalman inversion, optimization, numerical discretization of SDEs, stochastic differential equations, Euler–Maruyama

MSC codes. 65N21, 62F15, 65N75, 65C30, 90C56

DOI. 10.1137/21M1437561

1. Introduction.

Inverse problems have a wide range of application in sciences and engineering. The goal is to recover some unknown quantity of interest, which can only be observed indirectly through perturbed observations. These problems are typically ill-posed; in particular, solutions often do not depend on the data in a stable way, and regularization techniques are needed in order to overcome the instability. The Bayesian approach to inverse problems interprets the problem in a statistical framework, i.e., introduces a probabilistic model on the parameters and measurements in order to include the underlying uncertainty. The prior distribution on the unknown parameters reflects the prior knowledge on the parameters and regularizes the problem such that, under suitable assumptions, well-posedness results of the Bayesian problem can be shown. The posterior distribution, the solution to the Bayesian inverse problem, is the conditional distribution of the unknown parameters given the observations. Since the posterior distribution is usually not directly accessible, sampling methods for Bayesian inverse problems have become a very active field of research.

We will focus here on the ensemble Kalman filter (EnKF) for inverse problems, also known as ensemble Kalman inversion (EKI), which is a very popular method for the estimation of unknown parameters in various fields of application. Originally, the EnKF was introduced by Evensen [27, 28] for data assimilation problems, and

* Received by the editors July 30, 2021; accepted for publication (in revised form) September 2, 2022; published electronically December 16, 2022.
https://doi.org/10.1137/21M1437561

† Universität Augsburg, Institut für Mathematik, 86135 Augsburg, Germany (dirk.bloemker@math.uni-augsburg.de).

‡ Freie Universität Berlin, Arnimallee 3, 14195 Berlin, Germany (c.schillings@fu-berlin.de, phkwacker@gmail.com).

§ Universität Heidelberg, Interdisziplinäres Zentrum für Wissenschaftliches Rechnen, 69120 Heidelberg, Germany (simon.weissmann@uni-heidelberg.de).

more recently, it has been considered to solve inverse problems [42]. The EKI has been analyzed in the literature as particle approximation of the posterior distribution as well as a derivative-free optimization method for classical inverse problems. Both the EnKF and the EKI method have been analyzed in a continuous time formulation as a coupled system of stochastic differential equations (SDEs). The main focus of this work is to theoretically verify the convergence of the discrete EKI method to its continuous time formulation.

We will give an introduction to our mathematical setup after a brief overview of the existing literature.

1.1. Literature overview. As stated above the EnKF was introduced by Evensen [28] as a data assimilation method which approximates the filtering distribution based on particles. This method was first applied in the context of Bayesian inverse problems in [15, 23] and analyzed in the large ensemble size limit under linear and Gaussian assumptions [54, 48] as well as nonlinear models [53]. In [50] the authors study the mean field limit of the closely related ensemble square root filter (ESRF). The EnKF has been formulated in various multilevel formulations [34, 16, 35, 10]. A long time and ergodicity analysis are presented in [43, 65, 44], including uniform bounds in time and the incorporation of covariance inflation. Under linear and Gaussian assumptions the accuracy of the EnKF for a fixed ensemble size was studied in [66, 56], the accuracy of the ensemble Kalman–Bucy filter was studied in [19, 18]. Beside the large ensemble size limit, much work has been done in the analysis of the continuous time formulation [5, 6, 60]. Theoretical verification of the continuous time limits of the EnKF [52] and the ensemble square root filter [51] have been theoretically verified. In [52], uniform boundedness and global Lipschitz continuity on the forward and observation model is assumed. In [49], this assumption could be relaxed to general nonlinear functions by working with stopping time arguments introducing cut-offs and controlling the the empirical covariances. The results on the continuous time limits then hold locally in time with bounding constants growing exponentially in time or in convergence in probability, but error estimates are not given in any moments uniformly in time.

The application of the EnKF to inverse problems has been proposed in [42]. It can be viewed as a sequential Monte Carlotype method as well as a derivative-free optimization method. While in the setting of linear forward maps and Gaussian prior assumption the posterior can be approximated in the mean field limit, for nonlinear forward maps this iteration is not consistent with respect to the posterior distribution [26]. In [20, 32] the authors analyze the mean field limit based on the connection to the Fokker–Planck equation, whereas in [22] weights have been incorporated in order to correct the resulting posterior estimate for nonlinear models. Much of the existing theory for EKI is based on the continuous time limit resulting in a system of coupled SDEs which was formally derived in [63] and first analyzed in [7]. Furthermore, in [1] a stabilized continuous time formulation has been proposed. The continuous time formulation can be regarded as a derivative-free optimization method due to its gradient flow structure [63, 47]. In the literature two variants are typically considered: the deterministic formulation, which basically ignores the diffusion of the underlying SDE, and the stochastic formulation including the perturbed observations. In [8] the authors extend the results from [63] by showing well-posedness of the stochastic formulation and deriving first convergence results for linear forward models. The EKI for nonlinear forward models has been studied in [13] in discrete time with nonconstant step size. In [9] the dynamical system resulting from the continuous time limit of the EKI has been described and analyzed by a spectral decomposition.

In the viewpoint of EKI as an optimization method it naturally turns out that one has to handle noise in the data. In [64] the authors propose an early stopping criterion based on the Morozov discrepancy, and in [40, 41] discrete regularization has been considered. Most recently, in [12] the authors include Tikhonov regularization within EKI. Furthermore, adaptive regularization methods within EKI are studied in [59, 39].

In comparison to the EKI method studied in the following, a modified ensemble Kalman sampling method was introduced in [29] and further analyzed in [30, 21, 61]. The basic idea is to shift the noise in the observation to the particle itself and make use of the ergodicity of the resulting SDE related to the Langevin dynamic in order to build a sampling method.

1.2. Mathematical setup. We are interested in solving the inverse problem of recovering the unknown parameter $u \in \mathcal{X}$ from noisy data $y \in \mathbb{R}^K$ described through the underlying forward model

$$y = G(u) + \eta. \tag{1.1}$$

Here $G : \mathcal{X} \to \mathbb{R}^K$ denotes the possibly nonlinear forward map, mapping from a parameter space $\mathcal{X}$ to an observation space $\mathbb{R}^K$, and $\eta \sim \mathcal{N}(0, \Gamma)$ models the noise incorporated in the measurement. Throughout this article we will assume a finite-dimensional parameter space $\mathcal{X} = \mathbb{R}^p$. Due to the subspace property of the EKI (cp. [42]) the EKI ensemble stays in the affine subspace spanned by the initial ensemble, thus rendering the dynamics finite dimensional. Determinstic approaches to inverse problems typically consider the minimization of a regularized loss functional

$$\min_{u \in \mathbb{R}^p} \left\{ \mathcal{L}_{\mathbb{R}^K}(G(u), y) + \mathcal{R}_{\mathbb{R}^p}(u) \right\},$$

where $\mathcal{L}_{\mathbb{R}^K} : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}_+$ describes the discrepancy of the mapped parameter and the data, whereas $\mathcal{R}_{\mathbb{R}^p} : \mathbb{R}^p \to \mathbb{R}_+$ is the regularization function incorporating prior information on the parameter $u \in \mathbb{R}^p$. Classical choices of regularization include Tikhonov regularization [25] and total variation regularization [14, 62]. For more details on the different types of regularization we refer to [24, 4].

In contrast, from a statistical point of view, the Bayesian approach for inverse problems incorporates regularization through prior information of the underlying unknown parameter by introducing a probabilistic model. The unknown parameter u is modeled as an $\mathbb{R}^p$-valued random variable with prior distribution μ_0 which is stochastically independent of the noise η. Hence, we can view (u, y) as a jointly distributed random variable on $\mathbb{R}^K \times \mathbb{R}^p$, and solving the Bayesian inverse problem means conditioning on the event of the realized observation $y \in \mathbb{R}^K$. The solution of the Bayesian inverse problem is then given by the distribution of $u \mid y$, also known as the posterior distribution:

$$\mu(\mathrm{d}u) = \frac{1}{Z} \exp(-\Phi(u; y)) \mu_0(\mathrm{d}u) \tag{1.2}$$

with normalization constant $Z := \int_{\mathbb{R}^p} \exp(-\Phi(u; y)) \mu_0(\mathrm{d}u)$ and least-squares functional $\Phi(\cdot; y) : \mathbb{R}^p \to \mathbb{R}_+$ defined by $\Phi(u; y) = \frac{1}{2}\|y - G(u)\|_\Gamma^2$, where $\|\cdot\|_\Gamma := \|\Gamma^{-1/2} \cdot\|$ and $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^K$. We note that for a linear forward map $G(\cdot) = A \cdot$, $A \in \mathcal{L}(\mathbb{R}^p, \mathbb{R}^K)$ and Gaussian prior assumption $\mu_0 = \mathcal{N}(0, \frac{1}{\lambda} C_0)$ the maximum a posteriori estimate is computed as

$$\min_{u \in \mathbb{R}^p} \ \Phi(u; y) + \frac{\lambda}{2} \|u\|_{C_0}^2$$

which relates the Bayesian approach for inverse problems to the Tikhonov regularization with particular choice

$$\mathcal{L}_{\mathbb{R}^K}(G(u),y)=\frac{1}{2}\|y-G(u)\|_\Gamma^2 \quad \text{and} \quad \mathcal{R}_{\mathbb{R}^p}(u)=\frac{\lambda}{2}\|u\|_{C_0}^2.$$

1.3. EKI: The EnKF applied to inverse problems. The EKI method, as it was originally introduced in [40], can be viewed as a a sequential Monte Carlo method for sampling from the posterior distribution (1.2). The basic idea is to draw an ensemble of samples from the prior distribution and evolve it iteratively through linear Gaussian update steps in order to approximate the posterior distribution. The linear Gaussian update steps are based on the introduced tempered distribution

$$\mu_{n+1}(\mathrm{d}u)=\frac{1}{Z_n}\exp(-h\Phi(u;y))\mu_n(\mathrm{d}u) \tag{1.3}$$

with $h=1/N$ for some $N\in\mathbb{N}$ and normalizing constants Z_n. Note that μ_0 corresponds to the prior distribution and μ_N to the posterior distribution.

To be more concrete, we introduce the initial ensemble $(u_0^{(j)})_{j\in\{1,\dots,J\}}$ of size J as an independent and identically distributed (i.i.d.) sample from the prior $u_0^{(j)}\sim\mu_0$. The particle system in the current iteration is used as empirical approximation of the tempering distribution defined in (1.3):

$$\mu_n(\mathrm{d}u)\approx\frac{1}{J}\sum_{j=1}^J\delta_{u_n^{(j)}}(\mathrm{d}u).$$

Given the current particle system $(u_n^{(j)})_{j\in\{1,\dots,J\}}$ we compute the EnKF update for each particle accordingly to obtain a Gaussian approximation on the distribution μ_{n+1}. We define the following empirical means and covariances:

$$\begin{aligned}
C(u_n)&=\frac{1}{J}\sum_{j=1}^J(u_n^{(j)}-\bar{u}_n)(u_n^{(j)}-\bar{u}_n)^\top, \quad \bar{u}_n=\frac{1}{J}\sum_{j=1}^J u_n^{(j)},\\
C^{up}(u_n)&=\frac{1}{J}\sum_{j=1}^J(u_n^{(j)}-\bar{u}_n)(G(u_n^{(j)})-\bar{G}_n)^\top, \quad \bar{G}_n=\frac{1}{J}\sum_{j=1}^J G(u_n^{(j)}),\\
C^{pp}(u_n)&=\frac{1}{J}\sum_{j=1}^J(G(u_n^{(j)})-\bar{G}_n)(G(u_n^{(j)})-\bar{G}_n)^\top.
\end{aligned}$$

The ensemble Kalman iteration in discrete time is then given by

$$u_{n+1}^{(j)}=u_n^{(j)}-C^{up}(u_n)(C^{pp}(u_n)+h^{-1}\Gamma)^{-1}(G(u_n^{(j)})-y_{n+1}^{(j)}), \quad j=1,\dots,J, \tag{1.4}$$

where $h>0$ is the given artificial step size and $y_{n+1}^{(j)}$ is the artificially perturbed observation $y_{n+1}^{(j)}=y+\xi_{n+1}^{(j)}$, where $\xi_{n+1}^{(j)}$ are i.i.d. samples according to $\mathcal{N}(0,\frac{1}{h}\Gamma)$. The above perturbation can also be viewed acting on the mapped particles $G(u_n^{(j)})$ with possible interpretation of randomization. In this context, the scheme might be connected to sampling via randomized likelihood [2] and sampling via randomize-then-optimize [3]. Considering the EKI iteration in (1.4) we find the two parameters $h>0$, denoting the artificial step size, and $J\geq 2$, denoting the number of particles. To analyze the EKI method, typically at least one of the limits $h\to 0$ or $J\to\infty$

is applied. While the limit $J \to \infty$ refers to the mean field limit, the limit $h \to 0$ corresponds to the continuous time limit of the EKI.

Our aim is to give a rigorous verification of the continuous time limit for fixed ensemble size $2 \leq J < \infty$. Therefore, we first rewrite the discrete EKI formulation (1.4) as

$$\begin{aligned} u_{n+1}^{(j)} = u_n^{(j)} &- h\, C^{up}(u_n)(h\, C^{pp}(u_n) + \Gamma)^{-1}(G(u_n^{(j)}) - y) \\ &+ \sqrt{h}\, C^{up}(u_n)(h\, C^{pp}(u_n) + \Gamma)^{-1}\Gamma^{\frac{1}{2}}\zeta_{n+1}^{(j)}, \end{aligned}$$

where $\zeta_{n+1}^{(j)}$ are i.i.d. samples according to $\mathcal{N}(0, E_{K\times K})$, where $E_{K\times K}$ is the identity matrix in $\mathbb{R}^{K\times K}$. Taking the limit $h \to 0$ leads to $(h\, C^{pp}(u_n) + \Gamma)^{-1} \to \Gamma^{-1}$, and the continuous time limit of the discrete EKI can formally be written as the system of coupled SDEs

$$\mathrm{d}u_t^{(j)} = C^{up}(u_t)\Gamma^{-1}(y - G(u_t^{(j)}))\,\mathrm{d}t + C^{up}(u_t)\Gamma^{-\frac{1}{2}}\,\mathrm{d}W_t^{(j)}, \quad j = 1, \dots, J, \tag{1.5}$$

where $W^{(j)} = (W_t^{(j)})_{t\geq 0}$ are independent Brownian motions in $\mathbb{R}^p$. We denote by $\widetilde{\mathcal{F}}_t = \sigma(W_s^{(j)}, s \leq t)$ the filtration introduced by the Brownian motions and the particle system resulting from the continuous time limit, respectively. Furthermore, we denote by $\mathcal{F}_n = \sigma(\zeta_k^{(j)},\ j = 1, \dots, J,\ k \leq n)$ the filtration introduced by the increments of the Brownian motion and the particle system resulting from the discrete EKI formulation, respectively. In particular, for the rest of this article we will consider the filtered probability space $(\Omega, \mathcal{F}, \widetilde{\mathcal{F}} = (\widetilde{\mathcal{F}}_t)_{t\in[0,T]}, \mathbb{P})$ and $(\Omega, \mathcal{F}, \mathcal{F} = (\mathcal{F}_n)_{n=1}^N, \mathbb{P})$, respectively.

We are going to analyze the discrepancy between the discrete EKI formulation and its continuous time limit. Therefore, we introduce a continuous time interpolation of the discrete scheme denoted as $Y(t)$, and we describe the error by $E(t) = Y(t) - u(t)$. We will provide convergence in probability of the discrete EKI for general nonlinear forward maps, whereas in the linear setting we will provide strong convergence under suitable assumptions.

1.4. Outline of the paper. The contribution of this paper is a rigorous theoretical verification of the continuous time limit of the EKI. We provide two very general results independent of the structure of the ensemble Kalman method, which can then be applied to the EKI. In particular, we formulate the strong convergence result in such a way that it applies to various variants of the EKI. It is only required to verify the existence of moments up to a certain order.

We make the following contributions:

- We present approximation results for a general class of SDEs. Based on localization we are able to bound the error of the discretization up to a stopping time. Removing the stopping time leads to our two main results:
  1. convergence in probability with given rate function and
  2. convergence in L^θ with given rate function.
- We apply the general approximation results to the EKI method in a general nonlinear setting, where we can verify convergence in probability under very weak assumptions on the underlying forward model.
- In the linear setting we are able to prove strong convergence in L^θ of the discrete EKI method. While for general linear forward maps we obtain L^θ convergence for $\theta \in (0, 1)$, we provide various modifications of the scheme in order to ensure L^θ convergence for $\theta \in (0, 2)$.

With this manuscript we resolve the question posed in [7]: It is indeed the case that the specific form of the discrete EKI iteration (in particular the additional term $(hC^{pp}(u_n)+\Gamma)^{-1}$ vanishing in the continuous-time limit $h \to 0$) can be thought of as a time discretization for the SDE (1.5) specifically enforcing strong convergence, which cannot be said for a simple Euler–Maruyama-type iteration of form

$$u_{n+1}^{(j)} = u_n^{(j)} - h\, C^{up}(u_n)(G(u_n^{(j)}) - y) + \sqrt{h}\, C^{up}(u_n)\Gamma^{\frac{1}{2}}\zeta_{n+1}^{(j)}.$$

Indeed, numerical simulations (not presented in this manuscript, but easily implemented) show that the Euler–Maruyama discretization does not exhibit strong convergence (as already demonstrated for a similar SDE in [37]) due to rare events resulting in exploding iteration paths. There are connections to taming schemes (which have a similar effect of cutting off exploding iterations), as in [36], although the specific form of EKI is not a taming scheme in the narrow sense.

The remainder of this article is structured as follows. In section 2 we present our general numerical approximation results for SDEs which are then applied to the solution of general nonlinear inverse problems with the EKI method in section 3. The application to linear inverse problems is presented in section 4. We close the main part of the article with a brief conclusion in section 5 discussing possible further directions to take. Most of our proofs are shifted to the appendix in order to keep the focus on the key contribution presented in this article.

2. General approximation results for SDEs. In this section we discuss a general approximation result for SDEs, which is then applied to the EKI. We consider local solutions (i.e., up to a stopping time) of the following general SDE in $\mathbb{R}^n$ in integral notation:

$$x(t) = x_0 + \int_0^t f(x(s))dt + \int_0^t g(x(s))dW(s). \tag{2.1}$$

Given a step size $h > 0$, we assume that f and g are approximated by f_h and g_h, respectively, and we consider the following discrete numerical approximation to $x(t)$:

$$Y(t) = x_0 + \int_0^t f_h(Y(\lfloor s \rfloor))ds + \int_0^t g_h(Y(\lfloor s \rfloor))dW(s), \tag{2.2}$$

where we round down to the grid $\lfloor s \rfloor = \max\{kh \le s \ : \ k \in \mathbb{N}\}$ and suppress the index h in this notation.

It may seem strange to allow for the flexibility of approximating f and g by f_h and g_h (instead of just using pointwise evaluations of f and g), but this is exactly the case for the continuous and discrete version of EKI; see section 3. One can check that Y is a continuous time interpolation of the following discrete scheme:

$$Y_{n+1} = Y_n + hf_h(Y_n) + g_h(Y_n)[W(h(n+1)) - W(nh)], \qquad Y_0 = x(0).$$

We assume that both the discrete and the continuous scheme start at the same initial value, i.e., $x(0) = Y(0) = x_0$. Note that for every fixed $h > 0$ the discrete scheme exists for all times and cannot blow up in finite time. For the nonlinearities we assume that the limiting drift terms f and the limiting diffusion matrix g are locally Lipschitz and that the nonlinearities f_h and g_h have a local uniform bound in h on the growth and approximate f and g again in a local sense. To be more precise we formulate the following assumption.

Assumption 2.1. Assume that the functions $f, f_h : \mathbb{R}^p \to \mathbb{R}^p$ and $g, g_h : \mathbb{R}^p \to \mathbb{R}^{p\times K}$, $h \in (0,1)$, are locally Lipschitz such that for all radii $R > 0$ there exist constants C_a, L, and B such that for all $u, v \in \mathbb{R}^n$ with norm less than R the following properties hold:

1. uniform approximation on compact sets

$$\|f_h(u) - f(u)\| \le C_a(R,h), \qquad \|g_h(u) - g(u)\|_{\mathrm{HS}} \le C_a(R,h)$$

with $C_a(R,h) \to 0$ for $h \to 0$;

2. local Lipschitz continuity

$$\|f(u) - f(v)\| \le L(R)\|u - v\|, \qquad \|g(u) - g(v)\|_{\mathrm{HS}} \le L(R)\|u - v\|;$$

3. growth condition

$$\|f_h(u)\| \le B(R), \qquad \|g_h(u)\|_{\mathrm{HS}} \le B(R).$$

Moreover, we can assume without loss of generality that all R -dependent constants are nondecreasing in R.

Remark 2.2. Note that item 3 of Assumption 2.1 just means local boundedness, but we will use the specific growth factor $B(R)$ in the proofs later and have to compute the dependence of B on R. To be more precise, we will fix R depending on h such that various terms depending on h, $B(R)$, $L(R)$, and $C_a(R,h)$ are small. See, for example, (2.7).

Here and in the following we used $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ for the norm and the standard inner product in $\mathbb{R}^p$, while $\|\cdot\|_{\mathrm{HS}}$ is the standard Hilbert–Schmidt norm on matrices in $\mathbb{R}^{p\times p}$ which appears in Itos formula.

Using the Lipschitz property, we immediately obtain the following statement regarding the one-sided Lipschitz property.

LEMMA 2.3. *Under Assumption* 2.1 *we have for a small* $\epsilon > 0$ *that*

$$2\langle f(u) - f(y), u - v\rangle + c\|g(u) - g(v)\|_{\mathrm{HS}}^2 \le (\delta(R) - \epsilon)\|u - v\|^2$$

for all $u, v \in \mathbb{R}^n$ *with norm less than* R *with*

$$\delta(R) := 2L(R) + cL(R)^2 + \epsilon. \tag{2.3}$$

This estimate with $c = 1 + \epsilon$ is needed if we want to bound second moments of the error. The higher the moment we want to bound, the higher c has to be.

Convergence of the Euler–Maruyama scheme for SDEs was postulated under condition of finite exponential moment bounds of the discretization in [33], but this condition was soon after proven to be too restrictive: Divergence of the vanilla Euler–Maruyama scheme for non-Lipschitz-continuous coefficients was demonstrated in [37] due to an exponentially rare (in h) family of events with biexponentially bad behavior, which is why standard textbooks about numerical approximations of SDEs [46, 55, 57] generally assume globally Lipschitz-continuous coefficients. This led to the development of "taming schemes" in [38, 36] which are able to cut off the rare tail events leading to exploding moment bounds. The idea is to replace the Euler–Maruyama iteration for an SDE of form $dx = \mu(x)dt + \sigma(x)dW$ of type

$$x_{n+1} = x_n + h \cdot \mu(x_n) + \sigma(x_n)\Delta W_n$$

by something of the form

$$x_{n+1} = x_n + \frac{h \cdot \mu(x_n) + \sigma(x_n)\Delta W_n}{1 + |h \cdot \mu(x_n) + \sigma(x_n)\Delta W_n|}.$$

The denominator is close to 1 for small (well-behaving) increments, and it bounds large deviations (which have very small probability anyway) to avoid exploding paths. Our method of using stopping times to bound (stopped) moments and then remove the stopping times is based on ideas in [33].

2.1. Residual. We want to bound the error $E(t) = x(t) - Y(t)$ solving

$$dE = [f(x) - f(x-E)]dt + [g(x) - g(x-E)]dW + d\mathrm{Res}, \tag{2.4}$$

where we define the residual Res, which is an $\mathbb{R}^p$-valued process solving

$$d\mathrm{Res}(t) = [-f_h(Y(\lfloor t \rfloor)) + f(Y(t))]dt + [-g_h(Y(\lfloor t \rfloor)) + g(Y(t))]dW. \tag{2.5}$$

Note that the scheme is set up in such a way that $E(0) = 0$. Our strategy of proof is to first bound the error assuming that E, x, and Y are not too large. Later we will show that this is true with high probability.

DEFINITION 2.4 (cut-off). *For a fixed time $T > 0$ and sufficiently large radius R (which will depend on h later) we define the stopping time*

$$\tau_{R,h} = T \wedge \inf\{t > 0 : \|x(t)\| > R - 1, \text{ or } \|E(t)\| > 1\}.$$

Obviously, we have

$$\sup_{[0,\tau_{R,h}]} \|x(t)\| \le R \quad \text{and} \quad \sup_{[0,\tau_{R,h}]} \|Y(t)\| \le R.$$

Moreover, $\tau_{R,h} > 0$ a.s. if $\|x(0)\| < R - 1$, as both x and E are stochastic processes with continuous paths. We first bound the residual in (2.5).

LEMMA 2.5. *For $t \in [0, \tau_{R,h}]$ one has $d\mathrm{Res}(t) = \mathrm{Res}_1(t)dt + \mathrm{Res}_2(t)dW$ with $\mathrm{Res}_1 = -f_h(Y(\lfloor t \rfloor)) + f(Y(t))$ and $\mathrm{Res}_2 = -g_h(Y(\lfloor t \rfloor)) + g(Y(t))$. Then for $i = 1, 2$*

$$\mathbb{E} \sup_{t \in [0,\tau_{R,h}]} \|\mathrm{Res}_i(t)\|^p \le C_p K(R,h)^p$$

with a constant $C_p > 0$ depending only on p and

$$K(R,h) := C_a(R,h) + L(R)h^{1/2}B(R)\,. \tag{2.6}$$

As the residual needs to be small in order to prove an approximation result, in the applications we will need to choose a radius $R = R(h)$ with $R(h) \to \infty$ for $h \to 0$ such that

$$K(R(h),h) \to 0 \quad \text{for } h \to 0. \tag{2.7}$$

Proof. For the proof see Appendix A. ∎

2.2. Moment bound of the error. For the error we first prove the following lemma.

LEMMA 2.6. *We have for K from* (2.6)

$$\mathbb{E}\|E(t\wedge\tau_{R,h})\|^2 \le CK(R,h)^2\max\{1,e^{\delta(R)t}\} \quad \textit{for all } t\ge 0.$$

Sketch of the proof. The main idea here is to apply Itô's formula in order to derive

$$\begin{aligned} d\|E\|^2 &= 2\langle E, dE\rangle + \langle dE, dE\rangle \\ &= 2\langle E, [f(x)-f(x-E)] + \mathrm{Res}_1\rangle dt \\ &\quad + 2\langle E, [g(x)-g(x-E)+\mathrm{Res}_2]dW\rangle + \|[g(x)-g(x-E)]+\mathrm{Res}_2\|_{\mathrm{HS}}^2 dt \end{aligned}$$

and imply

$$\mathbb{E}\|E(t\wedge\tau_{R,h})\|^2 \le \|E(0)\|^2 + \delta(R)\mathbb{E}\int_0^t \|E(s\wedge\tau_{R,h})\|^2 dt + CK(R,h)^2.$$

The assertion follows by application of Gronwall's lemma. For full details of the proof see Appendix A. ∎

Remark 2.7. Note that for $\delta(R)\le C$ (which implies global Lipschitz continuity of f_h and g_h by its definition (2.3)), we have a valid error bound as soon as the residuals are small determined by $K(R,h)$. In the contrast to that, in the case $\delta(R)\nearrow\infty$ for $R\to\infty$, we might have an additional exponential in the bound. Thus we will have to take $R(h)$ much smaller in h, and we expect it to be some logarithmic term in h at most.

We could now proceed and extend this result to arbitrarily high moments; i.e., we can do estimates of $\mathbb{E}\|E(t\wedge\tau_{R,h})\|^p$ by using

$$\begin{aligned} d\|E\|^p = d(\|E\|^2)^{p/2} &= p\|E\|^{p-2}\langle E, dE\rangle + \frac{p}{2}\|E\|^{p-2}\langle dE, dE\rangle \\ &\quad + \frac{1}{2}p(p-2)\|E\|^{p-4}\langle E, dE\rangle^2. \end{aligned}$$

Each power is now sort of straightforward but needs a different one-sided Lipschitz condition. To avoid having too many technicalities, we only go up to the fourth power. We obtain as before

$$\begin{aligned} d\|E\|^4 &\le 4\|E\|^2\langle E, f(x)-f(x-E)+\mathrm{Res}_1\rangle dt \\ &\quad + 3\|E\|^2\|g(x)-g(x-E)+\mathrm{Res}_2\|_{\mathrm{HS}}^2 dt \\ &\quad + 2\|E\|^2\langle E, [g(x)-g(x-E)+\mathrm{Res}_2]dW\rangle \\ &\le \left[2\delta(R)\|E\|^4 + C_\epsilon\|\mathrm{Res}_2\|_{\mathrm{HS}}^4 + C_\epsilon\|\mathrm{Res}_1\|^4\right] dt \\ &\quad + 2\|E\|^2\langle E, [g(x)-g(x-E)+\mathrm{Res}_2]dW. \end{aligned}$$

Note that for the fourth power we need a slightly different one-sided Lipschitz condition than for the square. This would yield a different $\delta(R)$. Nevertheless, we slightly abuse notation and consider the same $\delta(R)$, i.e., the larger one, for both cases. Finally from Lemma 2.5, using the martingale property of the stopped integrals,

$$\begin{aligned} \mathbb{E}\|E(t\wedge\tau_{R,h})\|^4 &\le 2\delta(R)\mathbb{E}\int_0^{t\wedge\tau_{R,h}} \|E\|^4 dt + C_\epsilon T K(R,h)^4 \\ &\le 2\delta(R)\mathbb{E}\int_0^t \|E(s\wedge\tau_{R,h})\|^4 dt + CK(R,h)^4, \end{aligned}$$

and again Gronwall's lemma implies:

LEMMA 2.8. *We have for K from* (2.6)

$$\mathbb{E}\|E(t\wedge\tau_{R,h})\|^4 \le CK(R,h)^4 \max\{1, e^{2\delta(R)t}\} \quad \textit{for all } t\ge 0.$$

2.3. Uniform moment bound of the error. With our moment bounds we now obtain a bound on $\mathbb{E}\sup_{[0,\tau_{R,h}]}\|E\|^2$. Recall that $\tau_{R,h}\le T$ by definition.

LEMMA 2.9. *For all $T>0$ there is a constant $C>0$ such that for K from* (2.6)

$$\mathbb{E}\sup_{[0,\tau_{R,h}]}\|E\|^2 \le CK(R,h)^2(L(R)^2+1)\max\{1, e^{\delta(R)T}\}.$$

Proof. For the proof see Appendix A. ∎

Now we can finally fix in applications $R(h)\to\infty$ for $h\to 0$ (but sufficiently slow) such that

$$\mathbb{E}\sup_{[0,\tau_{R(h),h}]}\|E\|^2 \to 0 \text{ for } h\to 0.$$

Let us remark that we could also treat $\mathbb{E}\sup_{[0,\tau_{R,h}]}\|E\|^p$ for $p>2$, but this will be quite technical and lengthy, using Burkholder–Davis–Gundy.

2.4. Removing the stopping time. We present two results depending on how good our bounds are on x and Y.

Convergence in probability. For convergence in probability we only need stopped moments of x, as we do not control the error beyond the stopping time. Moreover, these moments can be very weak, for example, logarithmic.

THEOREM 2.10. *Assume that there is a radius $R(h)\to\infty$ and a $\gamma(h)\to 0$ such that*

$$\gamma(h)^{-2}\mathbb{E}\sup_{[0,\tau_{R(h),h}]}\|E\|^2 \to 0 \quad \text{for } h\to 0.$$

Moreover suppose that for an unbounded monotone growing function $\varphi:[0,\infty)\to[0,\infty)$ we have uniformly in $h\in(0,1)$ that $\mathbb{E}\varphi(\|x(\tau_{R(h),h})\|)\le C$. Then

$$\mathbb{P}\left(\sup_{[0,T]}\|E\|>\gamma(h)\right)\to 0 \quad \text{for } h\to 0.$$

Proof. Consider first using the definition of $\tau_{R,h}$ (where $R=R(h)\to\infty$ for $h\to 0$):

$$\begin{aligned}\mathbb{P}\left(\tau_{R,h}<T\right) &\le \mathbb{P}\left(\|E(\tau_{R,h})\|\ge 1 \text{ or } \|x(\tau_{R,h})\|\ge R-1\right)\\ &\le \mathbb{P}\left(\|E(\tau_{R,h})\|\ge 1\right)+\mathbb{P}\left(\|x(\tau_{R,h})\|\ge R-1\right)\\ &\le \mathbb{E}\|E(\tau_{R,h})\|^2+\mathbb{P}\left(\|x(\tau_{R,h})\|\ge R-1\right).\end{aligned}$$

Using this we thus obtain

$$\begin{aligned}\mathbb{P}\left(\sup_{[0,T]}\|E\|>\gamma(h)\right) &\le \mathbb{P}\left(\sup_{[0,T]}\|E\|>\gamma(h);\ \tau_{R,h}=T\right)\\ &\quad+\mathbb{P}\left(\sup_{[0,T]}\|E\|>\gamma(h);\ \tau_{R,h}<T\right)\\ &\le \mathbb{E}\sup_{[0,\tau_{R,h}]}\|E\|^2(1+\gamma(h)^{-2})+\mathbb{P}(\|x(\tau_{R,h})\|\ge R-1).\end{aligned}$$

Using the first assumption of the theorem for the first term, together with

$$\mathbb{P}(\|x(\tau_{R,h})\| \geq R-1) \leq \varphi(R-1)^{-1}\mathbb{E}\varphi(\|x(\tau_{R,h})\|) \to 0 \quad \text{as } R = R(h) \to \infty$$

by Markov inequality and the second assumption, finishes the proof. ∎

Convergence in moments. In order to bound the moments, we need control of the error beyond the stopping time $\tau_{R,h}$. Thus, we need a control on the moments of x and Y. Consider, for $\theta > 0$ to be fixed later, and $p > 1$,

$$\begin{aligned}\mathbb{E}\|E(t)\|^\theta &= \int_{\{\tau_{R,h}\geq t\}} \|E(t)\|^\theta d\mathbb{P} + \int_{\{\tau_{R,h}<t\}} \|E(t)\|^\theta d\mathbb{P} \\ &= \int_{\{\tau_{R,h}\geq t\}} \|E(t\wedge\tau_{R,h})\|^\theta d\mathbb{P} + \mathbb{E}\chi_{\{\tau_{R,h}<t\}}\|E(t)\|^\theta d\mathbb{P} \\ &\leq \mathbb{E}\|E(t\wedge\tau_{R,h})\|^\theta + \mathbb{P}\{\tau_{R,h}<t\}^{(p-1)/p}\Big(\mathbb{E}\|E(t)\|^{p\theta}\Big)^{1/p}.\end{aligned}$$

Now we use first $\Big(\mathbb{E}\|E(t)\|^{p\theta}\Big)^{1/p} \leq C\Big(\Big(\mathbb{E}\|x(t)\|^{p\theta}\Big)^{1/p\theta} + \Big(\mathbb{E}\|Y(t)\|^{p\theta}\Big)^{1/p\theta}\Big)^\theta$.

Secondly, we already saw (here $t \in [0,T]$)

$$\mathbb{P}\left(\tau_{R,h} < t\right) \leq \mathbb{P}\left(\tau_{R,h} < T\right) \leq \mathbb{E}\sup_{[0,\tau_{R,h}]}\|E\|^2 + \mathbb{P}(\|x(\tau_{R,h})\| \geq R-1).$$

We fix $q > \theta$, $\theta \leq 2$ and $p = q/\theta$ to obtain the following theorem.

THEOREM 2.11. *Assume that there is a radius $R(h) \to \infty$ such that*

$$\mathbb{E}\sup_{[0,\tau_{R(h),h}]}\|E\|^2 \to 0.$$

Moreover suppose that for an unbounded monotone growing function $\varphi : [0,\infty) \to [0,\infty)$ we have, uniformly in $h \in (0,1)$,

$$\mathbb{E}\varphi(\|x(\tau_{R(h),h})\|) \leq C$$

and suppose the following moment bounds for some $q > 0$:

$$\sup_{t\in[0,T]}\mathbb{E}\|x(t)\|^q + \sup_{t\in[0,T]}\mathbb{E}\|Y(t)\|^q \leq C.$$

Then we have for any $\theta \in (0,q)\cap(0,2]$ that $\lim_{h\searrow 0}\sup_{t\in[0,T]}\mathbb{E}\|E(t)\|^\theta = 0$.

Remark 2.12. In Lemma 2.3 we have provided a weak one-sided Lipschitz property which is enough to prove convergence of the error. Nevertheless, we remark without proof that all the error terms are much smaller if $\delta(R)$ is negative or at least bounded uniformly in R. We are even able to obtain rates of convergence in that case. For optimal rates we would also need arbitrarily high moments of both x and Y leading to quite technical estimates. See, for example, [33]. Since determining an optimal $\delta(R)$ is quite delicate in theapplication we have in mind, we postpone these questions to further research.

Remark 2.13. Furthermore, we remark without proof that we expect to be able to exchange the $\sup_{t\in[0,T]}$ and expectation in the statements. Actually, many strong convergence results are formulated as $\mathbb{E}\sup_t\|E(t)\|^\theta \to 0$.

For this we anyway have to first prove the result that we stated in the theorem above and then, in a second step, improve the estimate by using Burkholder inequality. As this would add further technical details and usually halves the order of convergence, we refrain from giving further details here.

3. Application to EKI: The nonlinear setting. After deriving approximation results for a general class of SDEs, we want to apply the proposed methods in order to quantify the convergence of the discrete EKI algorithm to its continuous version. We start the discussion by recalling our general nonlinear inverse problem $y = G(u) + \eta$, where $u \in \mathbb{R}^p$, $\eta \sim \mathcal{N}(0, \Gamma)$ for $\Gamma \in \mathbb{R}^{K\times K}$ and $y \in \mathbb{R}^K$. We suppose for simplicity that the forward model $G : \mathbb{R}^p \to \mathbb{R}^K$ is differentiable and grows at most polynomially. To be more precise we assume that there is an $m > 0$ and a constant such that for all u

$$\|G(u)\| \le C(1 + \|u\|^m) \quad \text{and} \quad \|DG(u)\| \le C(1 + \|u\|^{m-1}). \tag{3.1}$$

Recall that the discrete algorithm of the EKI is given by

$$\begin{aligned} u_{n+1}^{(j)} = u_n^{(j)} &- hC^{up}(u_n)(hC^{pp}(u_n) + \Gamma)^{-1}(G(u_n^{(j)}) - y) \\ &+ h^{1/2}C^{up}(u_n)(hC^{pp}(u_n) + \Gamma)^{-1}\Gamma^{1/2}W_{n+1}^{(j)}), \end{aligned}$$

while the continuous time limit is given by the system of coupled SDEs

$$\mathrm{d}u_t^{(j)} = C^{up}(u_t)\Gamma^{-1}(y - G(u_t^{(j)}))\,\mathrm{d}t + C^{up}(u_t)\Gamma^{-\frac{1}{2}}\mathrm{d}W_t^{(j)}, \tag{3.2}$$

where the sample covariances are defined in section 1.3 with ensemble size $J \ge 2$ and $W_n^{(j)}$ are i.i.d. $\mathcal{N}(0,1)$ random variables in both j and n. Now consider $u \in \mathbb{R}^{pJ}$ as $u = (u^{(1)}, \dots, u^{(J)})^T$ with $u^{(j)} \in \mathbb{R}^p$, and define the drift $f : \mathbb{R}^{pJ} \to \mathbb{R}^{pJ}$ and the diffusion $g : \mathbb{R}^{pJ} \to \mathbb{R}^{pJ\times pJ}$ by

$$f^{(j)}(u) = C^{up}(u)\Gamma^{-1}(y - G(u^{(j)})) \quad \text{and} \quad [g(u)z]_j = C^{up}(u)\Gamma^{-\frac{1}{2}}z_j.$$

The drift and diffusion in the discrete model aregiven by $f_h^{(j)}(u) = C^{up}(u)(hC^{pp}(u) + \Gamma)^{-1}(G(u^{(j)}) - y)$ and $[g_h(u)z]_j = C^{up}(u)(hC^{pp}(u)+\Gamma)^{-1}\Gamma^{1/2}z_j$, while the continuous interpolation Y is defined in (2.2) such that $Y(nk) = u_n$. Consider as before the error $E = u - Y$ between the continous solution u and the continuous interpolation Y of u_n.

We first observe that Assumption 2.1 is satisfied:

1. Obviously, both nonlinear terms are locally Lipschitz, since G is.
2. The matrix $(hC^{pp}(u)+\Gamma)^{-1}$ is uniformly bounded such that we have $B(R) = C(R^{1+2m} + 1)$ in Assumption 2.1 (f contains G twice).
3. Similarly, by computing the derivative we obtain that $L(R) = C(R^{2m} + 1)$ in Assumption 2.1.
4. For the approximation, we mainly have to bound

$$\begin{aligned} \|(hC^{pp}(u) + \Gamma)^{-1} - \Gamma^{-1}\|_{\mathrm{HS}} &= \|h\Gamma^{-1}C^{pp}(u)(hC^{pp}(u) + \Gamma)^{-1}\|_{\mathrm{HS}} \\ &\le Ch(R^{2m} + 1). \end{aligned}$$

Therefore we can choose $C_a(R, h) = Ch(R^{4m+1} + 1)$ in Assumption 2.1.

Thus we obtain for $h \in (0, 1)$

$$K(R, h) := Ch(R^{4m+1} + 1) + h^{1/2}C(R^{2m} + 1)C(R^{1+2m} + 1) \le Ch^{1/2}(R^{4m+1} + 1).$$

Moreover, we can choose a trivial bound with $\delta(R) = C(R^{4m} + 1)$. Thus, for any fixed $\gamma \in (0, 1/2)$, we can fix a radius $R(h) \nearrow \infty$ growing very slowly (logarithmically) in h such that using Lemma 2.6 (for small $h \to 0$)

$$h^{-2\gamma}\mathbb{E}\|E(t\wedge\tau_{R,h})\|^2 \le Ch^{-2\gamma}K(R(h),h)^2e^{\delta(R(h))T}$$
$$\le Ch^{1-2\gamma}R(h)^{8m+2}e^{CR(h)^{4m}} \to 0 \quad \text{for } h\to 0.$$

We are now ready to rewrite Theorem 2.10 for the EKI.

THEOREM 3.1. *Consider for the EKI with G satisfying* (3.1). *Define the error $E = u - Y$ as above, and fix $R(h) \nearrow \infty$ as above. Suppose that for a monotone growing function $\varphi : [0,\infty) \to [0,\infty)$ and every $T > 0$ in the definition of the stopping time $\tau_{R,h}$ we have uniformly for $h \in (0,1)$ that $\mathbb{E}\varphi(\|u(\tau_{R(h),h})\|) \le C$. Then for any fixed $\gamma \in (0,1/2)$ and $T > 0$*

$$\lim_{h\searrow 0}\mathbb{P}\left(\sup_{[0,T]}\|E\| > h^\gamma\right) = 0\,.$$

In the nonlinear setting based on Theorem 3.1 we will now prove the following main theorem for globally Lipschitz G. Later in the next section, we will use Theorem 3.5 in the case when G is linear.

THEOREM 3.2. *Consider for the EKI with G satisfying* (3.1) *with $m = 1$. Let $u_0 = (u_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$ -measurable maps $\Omega \to \mathbb{R}^p$ such that $\mathbb{E}[\|u_0^{(j)}\|^2] < \infty$, and suppose $\|y\|\|\Gamma^{-1/2}\|_{\mathrm{HS}} \le C$. For the error $E = u - Y$ as above we have for any fixed $\gamma \in (0,1/2)$ and $T > 0$*

$$\lim_{h\searrow 0}\mathbb{P}\left(\sup_{[0,T]}\|E\| > h^\gamma\right) = 0\,.$$

Proof. For the proof see Appendix B. ∎

We note that the above result can be used to verify unique strong solutions of the coupled SDEs (3.2). The proposed function $\varphi(\|\bar u\|^2) = \ln(1+\|\bar u\|^2)$ can be used as a stochastic Lyapunov function. With the computations given in the proof of Theorem 3.2, it is easy to verify that for $V(u) = \varphi(\|\bar u\|^2)$ it holds true that $LV(u) \le CV(u)$ for some constant $C > 0$, where LV denotes the V defined as

$$LV(u) := \nabla V(u)\cdot f(u) + \frac{1}{2}\mathrm{Tr}(g^\top(u)\nabla^2[V](u)g(u)).$$

Thus, by Theorem 3.5 in [45] we obtain global existence of unique strong solutions.

COROLLARY 3.3. *Under the same assumptions of Theorem* 3.2 *for all $T \ge 0$ there exists a unique strong solution $(u_t)_{t\in[0,T]}$ (up to $\mathbb{P}$-indistinguishability) of the set of coupled SDEs* (3.2).

Remark 3.4. We note that assuming that the forward map G takes values $G(u) = 0$ for $\|u\| \ge M$, where M is a certain tolerance value, we can directly apply Theorem 2.11 in order to prove strong convergence of the EKI iteration. This assumption forces the particle system in discrete and continuous time to be bounded, and is reasonable if it is known that proper solutions of the underlying inverse problem should be bounded. This assumption can be implemented by modifying the underlying forward map with a smooth shift to 0 close to the boundary of $\|u\| \in (-M, M)$. The EKI has been analyzed under this assumption, for example, in [13, 12].

Moreover, we can rewrite Theorem 2.11.

THEOREM 3.5. *Under the setting of Theorem* 3.1 *suppose we additionally have for $p > 0$ uniform bounds on the p-th moments of u and Y, i.e., there exists a $C > 0$*

such that for all $h \in (0,1)$

$$\sup_{t\in[0,T]} \mathbb{E}\|u(t)\|^p + \sup_{t\in[0,T]} \mathbb{E}\|Y(t)\|^p \le C;$$

then we have for any $\theta \in (0, \min\{2,p\})$

$$\lim_{h\searrow 0} \sup_{[0,T]} \mathbb{E}\|E(t)\|^\theta = 0.$$

We note that we only need to prove $\sup_{n\in\{0,\lfloor T/h\rfloor\}} \mathbb{E}\|u_n\|^2 \le C$ in the linear case later. As we have

$$Y(t) = \int_0^t f_h(Y(\lfloor s\rfloor))ds + \int_0^t g_h(Y(\lfloor s\rfloor))dW(s)$$

with $d\|Y(t)\|^2 = 2\langle Y(t), f_h(Y\lfloor t\rfloor)\rangle dt + \|g_h(Y\lfloor t\rfloor)\|^2_{\mathrm{HS}}dt + \langle Y(t), g_h(Y\lfloor t\rfloor)dW\rangle$, we provide the following interpolation result verifying that it is sufficient to derive bounds on the discretization scheme $\mathbb{E}\|u_n\|^2$.

LEMMA 3.6 (an interpolation lemma for lower moments). *Let* $u(t) = u_0 + t\cdot f(u_0) + g(u_0)W_t$ *with* u_0, W_t *independent and* $p \in (0,2)$. *Assume further that* $\mathbb{E}\|u_0\|^p < C$ *and* $\mathbb{E}\|u(1)\|^p < C$; *then* $\mathbb{E}\|u(t)\|^p < C_p\left[\mathbb{E}\|u_0\|^p + \mathbb{E}\|u(1)\|^p\right]$ *for all* $t \in [0,1]$.

Proof. The proof for this statement can be found in the appendix. ∎

We note that we can extend the above result to the whole time interval $[0,T]$ by a shift in time. We leave the details to reader.

4. Application to EKI: The linear setting. We consider the linear inverse problem of recovering an unknown parameter $u \in \mathbb{R}^p$, given noisy observations $y = Au + \eta \in \mathbb{R}^K$, where $\eta \sim \mathcal{N}(0,\Gamma)$ for $\Gamma \in \mathbb{R}^{K\times K}$. The ensemble Kalman iteration in discrete time is then given by

$$\begin{aligned} u^{(j)}_{n+1} &= u^{(j)}_n - C(u_n)A^T(AC(u_n)A^T + h^{-1}\Gamma)^{-1}(Au^{(j)}_n - y^{(j)}_{n+1}) \\ &= u^{(j)}_n - hC(u_n)A^T\Gamma^{-\frac12}(h\Gamma^{-\frac12}AC(u_n)A^T\Gamma^{-\frac12} + I)^{-1}\Gamma^{-\frac12}(Au^{(j)}_n - y^{(j)}_{n+1}), \end{aligned}$$

where we consider perturbed observations $y^{(j)}_{n+1} = y + h^{-\frac12}\Gamma^{\frac12}W^{(j)}_{n+1}$ with $W^{(j)}_{n+1}$ being i.i.d. $\mathcal{N}(0,1)$ random variables, and we denote by $\mathcal{F}_n = \sigma(W^{(j)}_m, m \le n, j = 1,\dots,J)$ the filtration introduced by the pertubation.

Further, we denote the identity matrix $I \in \mathbb{R}^p$, we define the scaled forward model $B := \Gamma^{-\frac12}A$, and we write the ensemble Kalman iteration for simplicity as

$$u^{(j)}_{n+1} = u^{(j)}_n - hC(u_n)B^TM(u_n)(Bu^{(j)}_n - \Gamma^{-\frac12}y) + \sqrt{h}C(u_n)B^TM(u_n)W^{(j)}_{n+1},$$

where we have introduced the notation $M(u_n) = (hBC(u_n)B^T + I)^{-1}$.

We can decompose $\Gamma^{-\frac12}y = \hat{y} + \tilde{y}$, where $\hat{y} \in \mathrm{range}\Gamma^{-\frac12}A$ and $\tilde{y}$ is in the orthogonal complement, such that the iteration reads as

$$\begin{aligned} u^{(j)}_{n+1} = u^{(j)}_n &- hC(u_n)B^TM(u_n)(Bu^{(j)}_n - \hat{y}) + hC(u_n)B^TM(u_n)\tilde{y} \\ &+ \sqrt{h}C(u_n)B^TM(u_n)W^{(j)}_{n+1}. \end{aligned}$$

Our first result states that the EKI dynamic ignores the part of observation which takes place in the orthogonal complement of the range of B.

LEMMA 4.1. *Let* $\tilde{y} \in \text{range}(B)^{\perp}$; *then for all* $n \in \mathbb{N}$ *we have* $C(u_n)B^T M(u_n)\tilde{y} = 0$.

Proof. For the proof see Appendix C. ∎

Our goal is to apply Theorem 3.5 in order to prove strong convergence of the ensemble Kalman iteration. To do so, we have to derive bounds on the moments of the continuous time limit $u(t)$ and on the continuous time interpolation of the discrete iteration $Y(t)$. We formulate our main result in the following theorem.

THEOREM 4.2. *Let* $u_0 = (u_0^{(j)})_{j\in\{1,\dots,J\}}$ *be* $\mathcal{F}_0$*-measurable maps* $\Omega \to \mathbb{R}^p$ *such that* $\mathbb{E}[\|u_0^{(j)}\|^2] < \infty$. *Furthermore, we assume that the discrete ensemble Kalman iteration can be bounded uniformly in* h, *i.e., there exists a* $C > 0$ *such that for all* $j \in \{1, \dots, J\}$ *it holds true that*

$$\sup_{n\in\{1,\dots,T\cdot N\}} \mathbb{E}[\|u_n^{(j)}\|^p] \le C.$$

Then we have strong convergence of the approximation error of the EKI method:

$$\lim_{h\searrow 0} \sup_{[0,T]} \mathbb{E}\|E(t)\|^\theta = 0 \qquad \text{for all } \theta \in (0, \min\{2, p\}).$$

Proof. In order to apply Theorem 3.5 we have to verify that

$$\sup_{t\in[0,T]} \mathbb{E}\|u(t)\|^p + \sup_{t\in[0,T]} \mathbb{E}\|Y(t)\|^p$$

is bounded uniformly in h. Much work has been investigated in the solution of the continuous formulation in [20, 8], where $\sup_{t\in[0,T]} \mathbb{E}\|u(t)\|^p$ can be bounded as the ensemble spread can be bounded in high moments up to $p < J + 3$ and moreover the bound follows by application of Itô's formula and Hölder's inequality [20, Lemma 5]. Note that this can be seen better in the continuous time formulation

$$du_t^{(j)} = \frac{1}{J}\sum_{k=1}^{J} \langle B(u_t^{(j)} - \bar{u}_t), y - Bu_t^{(j)} + dW_t^{(j)}\rangle (u_t^{(k)} - \bar{u}_t).$$

Secondly, we have to bound $\sup_{t\in[0,T]} \mathbb{E}\|Y(t)\|^p$. We apply the interpolation lemma for the p-th moments as the nodes of the interpolation are assumed to be bounded uniformly in h and, hence, $\sup_{t\in[0,T]} \mathbb{E}\|Y(t)\|^p \le C$. ∎

The update of the ensemble mean is governed by

$$\bar{u}_{n+1} = \bar{u}_n - hC(u_n)B^T M(u_n)(B\bar{u}_n - \hat{y}) + \sqrt{h}C(u_n)B^T M(u_n)\bar{W}_{n+1}$$

with $\bar{W}_{n+1} = \frac{1}{J}\sum_{j=1}^{J} W_{n+1}^{(j)}$. Further, we set $e_n^{(j)} := u_n^{(j)} - \bar{u}_n$, the particle deviation from the mean. Here we get the update formula

$$e_{n+1}^{(j)} = e_n^{(j)} - hC(u_n)B^T M(u_n)Be_n^{(j)} + \sqrt{h}C(u_n)B^T M(u_n)(W_{n+1}^{(j)} - \bar{W}_{n+1}).$$

We have seen that the update can be written as

$$u_{n+1}^{(j)} = u_n^{(j)} - hC(u_n)B^T M(u_n)(Bu_n^{(j)} - \hat{y}) + \sqrt{h}C(u_n)B^T M(u_n)W_{n+1}^{(j)}, \tag{4.1}$$

where $\hat{y} \in \text{range}(B)$, i.e., there exists $\hat{u}$ such that $\hat{y} = B\hat{u}$. We define the residuals $r_n^{(j)} = u_n^{(j)} - \hat{u}$, where the update of the residuals can be written as

$$r_{n+1}^{(j)} = r_n^{(j)} - hC(u_n)B^T M(u_n) B r_n^{(j)} + \sqrt{h} C(u_n) B^T M(u_n) W_{n+1}^{(j)}. \tag{4.2}$$

We note that all of the derived auxiliary results below crucially depend on the taming through

$$M(u_n) = (hBC(u_n)B^T + I)^{-1}, \tag{4.3}$$

suggesting that ignoring $hBC(u_n)B^T$ (which corresponds to an Euler–Maruyama scheme) does not lead to a stable discretization scheme.

We note that the result of Theorem 4.2 can be used as a general concept in order to prove the strong convergence for different variants of the EKI method as Tikhonov regularized EKI [12], ensemble Kalman one-shot inversion [31], or EKI under box-constraints [11]. Here, the main task is to derive bounds on the discrete ensemble Kalman iteration. To do so, we present a series of properties which can be used to bound the discrete iteration in moments.

Our first useful auxiliary result is a bound on the ensemble spread. In particular, we prove that the spread of the particle system is monotonically decreasing in time. This property is very useful from various perspectives. First, this property can be used to derive bounds on the particle system itself as we can describe the decrease of the spread through a concrete Lyapunov-type bound. Hence, by adding $\frac{1}{J}\sum_{j=1}^{J} \|e_n^{(j)}\|^2$ to the target value to bound, the increments of the target value decrease. We will see how to apply this approach in Proposition C.4. Second, in the interpretation of EKI as an optimization method we are interested in a convergence of the EKI to a point estimate. Hence, we expect each of the particles to converge to the same point. In particular, we will need to prove the following statements (see Appendix C):

- We provide a bound on the spread of the particles, i.e., we prove $\sup_{n\in\{1,\dots,N\}} \mathbb{E}[\|e_n^{(j)}\|^2] < \text{const}$, where $e_n^{(j)} := u_n^{(j)} - \bar{u}_n$.
- We extend this result by bounding the spread of the particles mapped by B, i.e., we prove $\sup_{n\in\{1,\dots,N\}} \mathbb{E}[\|Be_n^{(j)}\|^2] < \text{const}$.
- We provide a bound on the residuals mapped by B, i.e., we prove that the data misfit is bounded in the sense that $\sup_{n\in\{1,\dots,N\}} \mathbb{E}[\|Br_n^{(j)}\|^2] < \text{const}$.

Using these auxiliary results we are then able to provide various strong convergence results under certain assumptions, which are summarized in the following:

- In the first main result we do not state specific assumptions on the forward model without being linear. However, the strong convergence in Theorem 3.5 only holds for $\theta \in (0,1)$.
- Our second main result is based on the assumption that the initial ensemble lies outside the kernel of the forward map. While the moments of the dynamical system can be controlled in the image space of B, we are not able to control the unobserved part of the system, which is moving in the kernel of B. We again obtain strong convergence in the sense that Theorem 3.5 holds for all $\theta \in (0,2)$.
- Furthermore, including Tikhonov regularization within EKI we can verify the strong convergence for $\theta \in (0,2)$.

4.1. Strong convergence for general linear forward maps. In this section we consider general linear forward models $B = \mathbb{R}^{K\times p}$. In the following we derive bounds for $\mathbb{E}[\|u_n\|^\theta]$ for any $\theta \in (0,1)$.

LEMMA 4.3. *There exists a constant $C > 0$ independent of h and J but depending on T such that for all $j \in \{1, \dots, J\}$*

$$\sup_{n \in \{1,\dots,T\cdot N\}} \mathbb{E}[\|u_n^{(j)}\|] \leq C. \tag{4.4}$$

Proof. For the proof see Appendix C. ∎

COROLLARY 4.4. *Let $u_0 = (u_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\Omega \to \mathbb{R}^p$ such that $\mathbb{E}[\|u_0^{(j)}\|] < \infty$. Then we have strong convergence of the approximation error of the EKI method*

$$\lim_{h \searrow 0} \sup_{[0,T]} \mathbb{E}\|E(t)\|^\theta = 0 \quad \textit{for all } \theta \in (0,1).$$

Proof. The proof is a direct implication of Lemma 4.3 and Theorem 4.2. ∎

Remark 4.5. We note that the bound on $\theta < 1$ is due to technical reasons and does not come from a fact that there exist no uniform bounds on the moments of the discrete time system. In particular, we expect existence of uniformly bounded moments

$$\sup_{n \in \{1,\dots,T\cdot N\}} \mathbb{E}[\|u_n^{(j)}\|^p] \leq C \tag{4.5}$$

up to $p = 2$ and hence strong convergence up to $\theta < 2$. However, for proving bounds in L^2 one needs to derive bounds on moments of the ensemble spread in discrete time up to power 4, which is a challenging task in itself. Furthermore, we note that the derived bound is increasing in time with $\sqrt{T}$.

4.2. Strong convergence for the particle system initialized in the orthogonal complement of the kernel. The key idea of the following proof is to divide the particles dynamics into the dynamics in the kernel of the forward map B and its orthogonal complement. To do so, we introduce the orthogonal projection onto the orthogonal complement of the kernel $P = B^\top (BB^\top)^- B$, where $(BB^\top)^-$ denotes the generalized Moore–Penrose inverse of $BB^\top$. The idea is to split

$$r_n^{(j)} = P r_n^{(j)} + (I - P) r_n^{(j)}$$

and provide bounds for each term separately. We can verify bounded second moments of the particle system for the discrete EKI iteration initialized in the image space.

LEMMA 4.6. *Let $u_0 = (u_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\Omega \to \mathbb{R}^p$ such that $\mathbb{E}[\|B r_0^{(j)}\|^2] < \infty$, $\mathbb{E}[\|(I-P) r_0^{(j)}\|^2] < \infty$ and $(I-P) e_0^{(j)} = 0$ for all $j \in \{1, \dots, J\}$. Then there exists a constant $C > 0$ independent of h, J, and T such that for all $j \in \{1, \dots, J\}$*

$$\sup_{n \in \{1,\dots,T\cdot N\}} \mathbb{E}[\|r_n^{(j)}\|^2] \leq C. \tag{4.6}$$

Proof. For the proof see Appendix C. ∎

COROLLARY 4.7. *Let $u_0 = (u_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\Omega \to \mathbb{R}^p$ such that $\mathbb{E}[\|u_0^{(j)}\|^2] < \infty$ and $(I-P) e_0^{(j)} = 0$ for all $j \in \{1, \dots, J\}$. Then we have strong convergence of the approximation error of the EKI method*

$$\lim_{h \searrow 0} \sup_{[0,T]} \mathbb{E}\|E(t)\|^\theta = 0 \quad \textit{for all } \theta \in (0,2).$$

Proof. The proof is a direct implication of Lemma 4.6 and Theorem 4.2. ∎

Remark 4.8. We note that the assumption $(I-P)e_0^{(j)} = 0$ for all $j \in \{1,\dots,J\}$ could, for example, be ensured if the particle system is initialized with $u_0^{(j)} \mapsto Pu_0^{(j)}$. However, we mention that through the projection P much information about the forward map is necessary, which makes this result quite restrictive.

4.3. Strong convergence for Tikhonov regularized EKI and general linear forward maps. Much of the theoretical analysis of EKI is based on viewing it as an optimization method. The analysis is based on the long time behavior of the scheme, which is the study of the system of coupled SDEs (3.2) or the simplified ODE system suppressing the diffusion term for increasing time T. In particular, the aim of EKI in the long time behavior is to solve the minimization problem

$$\min_u \ \frac{1}{2}\|G(u)-y\|_\Gamma^2 \tag{4.7}$$

iteratively. For a linear forward map the motivation behind the EKI as optimization method can be seen by writing the drift term of (1.5) in a preconditioned gradient flow structure:

$$C^{up}(u_t)\Gamma^{-1}(y - Au_t^{(j)}) = C(u_t)A^\top\Gamma^{-1}(y - Au_t^{(j)}) = -C(u_t)\nabla_u\left(\frac{1}{2}\|Au^{(j)} - y\|_\Gamma^2\right).$$

Similarly, in the nonlinear setting, using a second-order approximation, we can view the drift term of (1.5) as approximation of a preconditioned gradient flow [47]:

$$\begin{aligned} C^{up}(u_t)\Gamma^{-1}(y - G(u_t^{(j)})) &\approx C(u_t)(DG(u_t^{(j)}))^\top\Gamma^{-1}(y - G(u_t^{(j)})) \\ &= -C(u_t)\nabla_u\left(\frac{1}{2}\|G(u^{(j)}) - y\|_\Gamma^2\right). \end{aligned}$$

Solving the inverse problem through the optimization problem (4.7) is typically ill-posed, and regularization is needed. In [64] the authors propose an early stopping criterion based on the Morozov discrepancy principle [58], whereas in [12] Tikhonov regularization has been included into the scheme. We will focus on the Tikhonov regularized EKI (TEKI) and prove the strong convergence of the discrete TEKI. While the TEKI can also be formulated for nonlinear forward maps, we we will focus on the linear setting.

The basic idea of the incorporation of Tikhonov regularization into EKI is to extend the underlying inverse problem (1.1) by prior information. This extension reads as follows:

$$y = Au + \eta, \qquad 0 = u + \xi,$$

where $\eta \sim \mathcal{N}(0,\Gamma)$ and $\xi \sim \mathcal{N}(0, \frac{1}{\lambda}C_0)$. Introducing the variables

$$\tilde{A} = \begin{pmatrix} A \\ I \end{pmatrix}, \quad \tilde{y} = \begin{pmatrix} y \\ 0 \end{pmatrix}, \quad \tilde{\eta} \sim \mathcal{N}\left(0,\tilde{\Gamma}\right), \quad \tilde{\Gamma} = \begin{pmatrix} \Gamma & 0 \\ 0 & \frac{1}{\lambda}C_0 \end{pmatrix}$$

we can write the extended inverse problem as $\tilde{y} = \tilde{A}u + \tilde{\eta}$.

For TEKI we now apply EKI to the extended inverse problem which then reads as

$$u_{n+1}^{(j)} = u_n^{(j)} - C(u_n)\tilde{A}^T(\tilde{A}C(u_n)\tilde{A}^T + h^{-1}\tilde{\Gamma})^{-1}(\tilde{A}u_n^{(j)} - \tilde{y}_{n+1}^{(j)})$$

with corresponding continuous time limit

$$\mathrm{d}u_t^{(j)} = C(u_t)\tilde{A}^T\tilde{\Gamma}^{-1}(\tilde{y} - \tilde{A}u_t^{(j)})\,\mathrm{d}t + C(u_t)\tilde{A}^T\tilde{\Gamma}^{-\frac{1}{2}}\mathrm{d}W_t^{(j)}, \tag{4.8}$$

where $W^{(j)}$ are independent Brownian motions in $\mathbb{R}^K \times \mathcal{X}$. In the long time behavior TEKI can be viewed as optimizer of the regularized objective function

$$\Phi_R(u, y) = \frac{1}{2}\|\tilde{A}u - \tilde{y}\|^2_{\mathbb{R}^K\times\mathcal{X}} = \frac{1}{2}\|Au - y\|^2 + \frac{\lambda}{2}\|C_0^{-1/2}u\|^2_{\mathcal{X}}.$$

The motivation behind this viewpoint can be seen by writing out the drift term of (4.8) as

$$\begin{aligned} C(u_t)\tilde{A}^T\tilde{\Gamma}^{-1}(\tilde{y} - \tilde{A}u^{(j)}) &= C(u_t)\left(A^T\Gamma^{-1}(y - Au_t^{(j)}) - \lambda C_0^{-1}u_t^{(j)}\right) \\ &= -C(u_t)\nabla_u\left(\frac{1}{2}\|Au_t^{(j)} - y\|^2_\Gamma + \frac{\lambda}{2}\|u_t^{(j)}\|^2_{C_0}\right). \end{aligned}$$

For a detailed convergence analysis of the TEKI as optimization method we refer to [12]. Since $\tilde{A}$ and $\tilde{B} := \tilde{\Gamma}^{-1/2}\tilde{A}$, respectively, are linear operators, we can directly apply the resultspresented above. In particular, we are going to apply Proposition C.4 in order to verify the strong convergence of the discrete TEKI to its continuous time formulation. We prove that the second moments of the particle system are bounded.

LEMMA 4.9. *Let $u_0 = (u_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\Omega \to \mathbb{R}^p$ such that $\mathbb{E}[\|u_0^{(j)}\|^2] < \infty$ for all $j \in \{1,\dots,J\}$. Then there exists a constant $C > 0$ independent of h, J, and T such that for all $j \in \{1,\dots,J\}$*

$$\sup_{n\in\{1,\dots,T\cdot N\}} \mathbb{E}[\|u_n^{(j)}\|^2] \le C. \tag{4.9}$$

Proof. For the proof see Appendix C. ∎

As we can ensure the bound on the second moments of the particle system we are ready to formulate our main result of strong convergence for the discrete TEKI iteration.

COROLLARY 4.10. *Let $u_0 = (u_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\Omega \to \mathbb{R}^p$ such that $\mathbb{E}[\|u_0^{(j)}\|^2] < \infty$. Then we have strong convergence of the approximation error of the TEKI method: $\lim_{h\searrow 0}\sup_{[0,T]}\ \mathbb{E}\|E(t)\|^\theta = 0$ for all $\theta \in (0, 2)$.*

5. Conclusion. We have shown that on finite time scales $[0, T]$, the discrete EKI dynamics can be used to approximate the continuous EKI. Or, the other way around, we have established the legitimacy of analyzing the EKI dynamics with a time-continuous model and draw conclusions about the discrete EKI dynamics implemented in practice. For the general nonlinear model, we were able to prove convergence of the discretization in probability, while for the linear setting, we were even able to prove convergence in the L^θ sense, for $\theta \in (0, 1)$, with higher exponents in more favorable settings. We note that the constant derived in the proof still depends on time in the form of $\sqrt{T}$. Due to the fact that we were able to eliminate dependence on T in the other settings considered (TEKI, convergence in probability for the nonlinear model), we believe that this can be done similarly in the linear setting as well, maybe under additional assumptions, and we leave this as a task for future work.

The methods which we have employed can be used very generally in an SDE setting and can be applied to the analysis of discretization schemes for SDEs in different contexts.

Appendix A. Proofs of section 2.

Proof of Lemma 2.5. We start by bounding the error between $Y(t)$ for $t \in [kh,(k+1)]h$ and Y_k. By the SDE for the approximation,

$$\begin{aligned}\|Y(\lfloor t\rfloor) - Y(t)\| &= \| \int_{\lfloor t\rfloor}^{t} f_h(Y(\lfloor s\rfloor))ds + \int_{\lfloor t\rfloor}^{t} g_h(Y(\lfloor s\rfloor))dW(s)\| \\ &\le hB(R) + B(R)\|W(t) - W(\lfloor t\rfloor)\|.\end{aligned}$$

Thus, by the Burkholder–Davis–Gundy inequality and by merging the higher-order term h^p into the lower-order term $h^{p/2}$ with an appropriate constant,

$$\mathbb{E} \sup_{[0,\tau_{R,h}]} \|Y(\lfloor t\rfloor) - Y(t)\|^p \le C_p h^{p/2} B(R)^p.$$

In order to bound the residual, we consider $t \in [0,\tau_{R,h}]$ and thus $\lfloor t\rfloor \in [0,\tau_{R,h}]$ with $\|Y(\lfloor t\rfloor)\| \le R$. Now $\|f_h(Y(\lfloor t\rfloor)) - f(Y(\lfloor t\rfloor))\| \le C_a(R,h)$ and $\|f(Y(\lfloor t\rfloor)) - f(Y(t))\| \le L(R)\|Y(\lfloor t\rfloor) - Y(t)\|$. Thus we have

$$\mathbb{E} \sup_{[0,\tau_{R,h}]} \|\mathrm{Res}_1(t)\|^p \le C_p \left[C_a(R,h) + L(R)h^{1/2}B(R)\right]^p.$$

The bound for Res_2 follows in a similar way. ∎

Proof of Lemma 2.6. Recall that the error is given by

$$dE = [f(x) - f(x-E)]dt + [g(x) - g(x-E)]dW + \mathrm{Res}_1 dt + \mathrm{Res}_2 dW.$$

Thus, using Itos-formula we obtain for some constant C_ϵ depending only on ϵ

$$\begin{aligned} d\|E\|^2 &= 2\langle E, dE\rangle + \langle dE, dE\rangle \\ &= 2\langle E, [f(x) - f(x-E)] + \mathrm{Res}_1\rangle dt + 2\langle E, [g(x) - g(x-E) + \mathrm{Res}_2]dW\rangle \\ &\quad + \|[g(x) - g(x-E)] + \mathrm{Res}_2\|_{\mathrm{HS}}^2 dt \\ &\le 2\langle E, [f(x) - f(x-E)]\rangle dt + (1+\epsilon)\|g(x) - g(x-E)\|_{\mathrm{HS}}^2 dt + \epsilon\|E\|^2 dt \\ &\quad + \left[C_\epsilon\|\mathrm{Res}_2\|_{\mathrm{HS}}^2 + C_\epsilon\|\mathrm{Res}_1\|^2\right] dt + 2\langle E, [g(x) - g(x-E) + \mathrm{Res}_2]dW\rangle \\ &\le \left[\delta(R)\|E\|^2 + C_\epsilon\|\mathrm{Res}_2\|_{\mathrm{HS}}^2 + C_\epsilon\|\mathrm{Res}_1\|^2\right] dt \\ &\quad + 2\langle E, [g(x) - g(x-E) + \mathrm{Res}_2]dW\rangle. \end{aligned} \tag{A.1}$$

This yields from Lemma 2.5 using the martingale property of the stopped integrals

$$\begin{aligned}\mathbb{E}\|E(t\wedge\tau_{R,h})\|^2 &\le \|E(0)\|^2 + \delta(R)\mathbb{E}\int_0^{t\wedge\tau_{R,h}} \|E\|^2 dt + C_\epsilon T K(R,h)^2 \\ &\le \|E(0)\|^2 + \delta(R)\mathbb{E}\int_0^{t} \|E(s\wedge\tau_{R,h})\|^2 dt + CK(R,h)^2,\end{aligned}$$

where the constant depends on T and the choice of ϵ. Assume first that $\delta(R) > 0$. Using Gronwall's lemma and $E(0) = 0$ we obtain the bound

$$\mathbb{E}\|E(t\wedge\tau_{R,h})\|^2 \le Ce^{\delta(R)t}K(R,h)^2.$$

Assume now that $\delta(R) \le 0$. This yields from (A.1) using the martingale property of the stopped integrals

$$\mathbb{E}\|E(t\wedge\tau_{R,h})\|^2 + |\delta(R)|\mathbb{E}\int_0^{t\wedge\tau_{R,h}} \|E\|^2 dt \le CK(R,h)^2. \qquad ∎$$

Proof of Lemma 2.9. Recall from (A.1) for $t \le \tau_{R,h}$

$$\|E(t)\|^2 \le \int_0^t \left[\delta(R)\|E\|^2 + C\|\mathrm{Res}_2\|_{\mathrm{HS}}^2 + C\|\mathrm{Res}_1\|^2\right] dt + 2\int_0^t \langle E, [g(x) - g(x-E) + \mathrm{Res}_2] dW\rangle.$$

Thus, using Burkholder–Davis–Gundy (recall $\tau_{R,h} \in [0,T]$) assuming $\delta(R) > 0$

$$\begin{aligned}\mathbb{E}\sup_{[0,\tau_{R,h}]}\|E\|^2 &\le \mathbb{E}\int_0^{\tau_{R,h}} \left[\delta(R)\|E\|^2 + C\|\mathrm{Res}_2\|_{\mathrm{HS}}^2 + C\|\mathrm{Res}_1\|^2\right] ds \\ &\quad + 2\mathbb{E}\Big(\int_0^{\tau_{R,h}} \left[L(R)^2\|E\|^4 + \|E\|^2\|\mathrm{Res}_2\|_{\mathrm{HS}}^2\right] dt\Big)^{1/2} \\ &\le \delta(R)\int_0^T \mathbb{E}\|E(s\wedge\tau_{R,h})\|^2 ds + C\mathbb{E}\sup_{[0,\tau_{R,h}]}\|\mathrm{Res}_2\|_{\mathrm{HS}}^2 + C\mathbb{E}\sup_{[0,\tau_{R,h}]}\|\mathrm{Res}_1\|^2 \\ &\quad + C\Big((L(R)^2+1)\int_0^T \mathbb{E}\|E(s\wedge\tau_{R,h})\|^4 ds + \mathbb{E}\sup_{[0,\tau_{R,h}]}\|\mathrm{Res}_2\|_{\mathrm{HS}}^4\Big)^{1/2}.\end{aligned}$$

Using Lemma 2.5 we obtain

$$\begin{aligned}\mathbb{E}\sup_{[0,\tau_{R,h}]}\|E\|^2 &\le \delta(R)\int_0^T \mathbb{E}\|E(s\wedge\tau_{R,h})\|^2 ds \\ &\quad + C(L(R)^2+1)\Big(\int_0^T \mathbb{E}\|E(s\wedge\tau_{R,h})\|^4 ds\Big)^{1/2} + CK(R,h)^2.\end{aligned}$$

Moreover in the case $\delta(R) \le 0$ we have similarly

$$\mathbb{E}\sup_{[0,\tau_{R,h}]}\|E\|^2 \le C(L(R)^2+1)\Big(\int_0^T \mathbb{E}\|E(s\wedge\tau_{R,h})\|^4 ds\Big)^{1/2} + CK(R,h)^2.$$

We obtain the assertion by using Lemmas 2.6 and 2.8. ∎

Appendix B. Proofs of section 3.

LEMMA B.1 (an interpolation lemma for second moments). *Let $u(t) = u_0 + t\cdot f(u_0) + g(u_0)W_t$ with u_0, W_t independent. Assume further that $\mathbb{E}\|u_0\|^2 < C$ and $\mathbb{E}\|u(1)\|^2 < C$; then $\mathbb{E}\|u(t)\|^2 < C$ for all $t \in [0,1]$.*

Proof. Note first that by independence, $\mathbb{E}[h(u_0)W_t] = 0$ and

$$\mathbb{E}[h(u_0)^2 W_t^2] = \mathbb{E}[h(u_0)]^2\mathbb{E}[W_t]^2 = \mathbb{E}[h(u_0)]^2 t$$

for (suitably integrable) functions h. We compute first

$$\mathbb{E}[u(1)]^2 = \mathbb{E}[u_0 + f(u_0) + g(u_0)W_1]^2 = \mathbb{E}[u_0 + f(u_0)]^2 + 2\cdot 0 + \mathbb{E}[g(u_0)]^2\cdot 1$$

.

Thus,

$$\begin{aligned}\mathbb{E}[u_0 + f(u_0)t + g(u_0)W_t]^2 &= \mathbb{E}[u_0 + f(u_0)t]^2 + 2\cdot 0 + \mathbb{E}[g(u_0)W_t]^2 \\ &= \mathbb{E}[(1-t)u_0 + t(u_0 + g(u_0))]^2 + \mathbb{E}[g(u_0)]^2 t.\end{aligned}$$

Now we note that $((1-t)a+tb)^2 \le (1-t)a^2 + t(a+b)^2$ by Jensen's inequality

$$\begin{aligned}\mathbb{E}[u_0 + f(u_0)t + g(u_0)W_t]^2 &\le (1-t)\mathbb{E}u_0^2 + t\mathbb{E}[u_0 + f(u_0)]^2 + t\mathbb{E}[g(u_0)]^2 \\ &= (1-t)\mathbb{E}u_0^2 + t\mathbb{E}[u(1)]^2\end{aligned}$$

from which the statement follows. ∎

We will need the following fundamental lemmata.

LEMMA B.2. *Let $W \sim N(0,\sigma^2)$ be a centered Gaussian random variable. Then $\mathbb{E}|W|^p = C_p \cdot \left(\mathbb{E}|W|^2\right)^{\frac{p}{2}}$.*

Proof. See Proposition 2.19 in [17]. ∎

For non-centered Gaussian random variables we can show the following lemma.

LEMMA B.3. *Let $Z \sim N(a,\sigma^2)$ be a Gaussian random variable. Then there is a constant $C_p > 0$ such that*

$$\left(\mathbb{E}|Z|^2\right)^{\frac{1}{2}} \le C_p \left(\mathbb{E}|Z|^p\right)^{\frac{1}{p}}.$$

Proof. We can assume $\sigma = 1$ by rescaling and setting $Z = a + W$ with $W \sim N(0,1)$. Now we consider

$$\frac{\left(\mathbb{E}|Z|^p\right)^{\frac{1}{p}}}{\left(\mathbb{E}|Z|^2\right)^{\frac{1}{2}}} = \frac{\left(\mathbb{E}|a+W|^p\right)^{\frac{1}{p}}}{\left(\mathbb{E}|a+W|^2\right)^{\frac{1}{2}}} =: f_p(a)$$

as a function of a. If we can show that $\inf_a f_p(a) > 0$, then the statement follows with $C_p = (\inf_a f_p(a))^{-1}$. Evidently $f_p(a) > 0$ for all $a \in \mathbb{R}$, also $f_p(-a) = f_p(a)$, and f_p is a continuous map. Thus, if we can show that $\lim_{a\to\infty} f_p(a) > 0$, then $\inf_a f_p(a) > 0$. We start by noting that $\mathbb{E}|a+W|^2 = a^2 + 1$. Then

$$\begin{aligned}\lim_{a\to\infty} (f_p(a))^p &= \lim_{a\to\infty} \mathbb{E}\left|\frac{a+W}{\sqrt{1+a^2}}\right|^p = \lim_{a\to\infty} \mathbb{E}\left|\frac{a+W}{a}\right|^p \left|\frac{a}{\sqrt{1+a^2}}\right|^p \\ &= \lim_{a\to\infty} \mathbb{E}\left|1 + a^{-1}W\right|^p \ge \lim_{a\to\infty} (1-\epsilon)^{-p} \cdot \mathbb{P}\left(|1+a^{-1}W| \ge 1-\epsilon\right) \\ &= (1-\epsilon)^{-p},\end{aligned}$$

where we used Chebyshev's inequality and

$$\mathbb{P}\left(|1+a^{-1}W| \ge 1-\epsilon\right) \ge \mathbb{P}\left(1+a^{-1}W \ge 1-\epsilon\right) \ge \mathbb{P}\left(W \ge -a\epsilon\right).$$

As $\lim_{a\to\infty}(f_p(a))^p \ge \sup_{\epsilon\in(0,1)}(1-\epsilon)^{-p} = 1 > 0$, we have shown the statement. ∎

LEMMA B.4. *Let $W \sim N(0,1)$. Then for $a, b \in \mathbb{R}$, $p \in (0,2)$ and $t \in (0,1)$, we have*

$$\mathbb{E}|a + tb + \sqrt{t}cW|^p \le C_p \cdot \left[|a|^p + \mathbb{E}|a + b + cW|^p\right].$$

Proof. We note that for for random variables $\mathbb{E}|X|^p \le \left(\mathbb{E}|X|^2\right)^{\frac{p}{2}}$ by Hölder's inequality. Also, $|w+z|^{\frac{p}{2}} \le |w|^{\frac{p}{2}} + |z|^{\frac{p}{2}}$. Thus, using Lemma B.1,

$$\begin{aligned}\mathbb{E}|a + tb + \sqrt{t}cW|^p &\le \left(\mathbb{E}|a + tb + \sqrt{t}cW|^2\right)^{\frac{p}{2}} \le \left(a^2 + \mathbb{E}|a+b+cW|^2\right)^{\frac{p}{2}} \\ &\le |a|^p + \left(\mathbb{E}|a+b+cW|^2\right)^{\frac{p}{2}} \le |a|^p + C_p\mathbb{E}|a+b+cW|^p\end{aligned}$$

with the last step being due to Lemma B.3. ∎

Proof of Lemma 3.6. We first consider the case where all stochastic processes and random variables involved are one-dimensional. Then the statement is a consequence of Lemma B.4 after seeing that

$$\mathbb{E}|u_0 + tf(u_0) + g(u_0)W_t|^p = \mathbb{E}\left[\mathbb{E}\left[|u_0 + tf(u_0) + g(u_0)W_t|^p|\mathcal{F}_0\right]\right]$$

and identifying $a = u_0$, $b = f(u_0)$, $\sqrt{t}cW = g(u_0)W_t$ (where we can use $\sqrt{t}W = W_t$ in distribution for $W \sim N(0,1)$). The higher-dimensional case then follows from the one-dimensional considerations by seeing that for a random vector Z,

$$\mathbb{E}\|Z\|^p = \mathbb{E}\left(\sum_{i=1}^d |z_i|^2\right)^{\frac{p}{2}} \simeq \left(\sum_{i=1}^d \mathbb{E}|z_i|^p\right),$$
$$\left(\mathbb{E}\|Z\|^2\right)^{\frac{p}{2}} = \left(\mathbb{E}\sum_{i=1}^d |z_i|^2\right)^{\frac{p}{2}} \simeq \sum_{i=1}^d \mathbb{E}\left(|z_i|^2\right)^{\frac{p}{2}},$$

where $x \simeq y$ means that there exist constants $a, A > 0$ such that $ax \le y \le Ax$. ∎

Proof of Theorem 3.2. Recall that by Theorem 3.1, we just have to verify that there exists φ (monotone growing) such that $\mathbb{E}\varphi(\|u(\tau_{R,h})\|) \le C$. We first introduce the shorthand notation $\mathcal{F}(u) = C^{up}(u)\Gamma^{-1/2}$ and rewrite

$$du^{(j)} = -\mathcal{F}(u)\Gamma^{-1/2}(G(u^{(j)}) - y)dt + \mathcal{F}(u)dW.$$

Denote by $\overline{u}$, $\overline{W}$ and $\overline{G}$ the mean values of $u^{(j)}$, $W^{(j)}$, and $G(u^{(j)})$ with respect to j. Thus,

$$d\overline{u} = -\mathcal{F}(u)\Gamma^{-1/2}(\overline{G} - y)dt + \mathcal{F}(u)d\overline{W}$$

and

$$d(u^{(j)} - \overline{u}) = -\mathcal{F}(u)\Gamma^{-1/2}(G(u^{(j)}) - \overline{G})dt + \mathcal{F}(u)d(W^{(j)} - \overline{W}).$$

By Itô-s formula we obtain

$$\begin{aligned} d\|u^{(j)} - \overline{u}\|^2 &= 2\langle u^{(j)} - \overline{u}, d(u^{(j)} - \overline{u})\rangle + \langle d(u^{(j)} - \overline{u}), d(u^{(j)} - \overline{u})\rangle \\ &= -2\langle u^{(j)} - \overline{u}, \mathcal{F}(u)\Gamma^{-1/2}(G(u^{(j)}) - \overline{G})\rangle dt \\ &\quad + 2\langle u^{(j)} - \overline{u}, \mathcal{F}(u)d(W^{(j)} - \overline{W})\rangle \\ &\quad + \langle \mathcal{F}(u)d(W^{(j)} - \overline{W}), \mathcal{F}(u)d(W^{(j)} - \overline{W})\rangle. \end{aligned}$$

Now we use that

$$\begin{aligned} &\frac{1}{J}\sum_j \langle u^{(j)} - \overline{u}, \mathcal{F}(u)\Gamma^{-1/2}(G(u^{(j)}) - \overline{G})\rangle \\ &\qquad = \frac{1}{J}\sum_j \mathrm{Tr}(\mathcal{F}(u)\Gamma^{-1/2}(G(u^{(j)}) - \overline{G})(u^{(j)} - \overline{u})^\top) \\ &\qquad = \mathrm{Tr}(F(u)F(u)^\top) = \|\mathcal{F}(u)\|_{\mathrm{HS}}^2 \end{aligned}$$

and $\langle \mathcal{F}(u)d(W^{(j)} - \overline{W}), \mathcal{F}(u)d(W^{(j)} - \overline{W})\rangle = 2(1 - \frac{1}{J})\|\mathcal{F}(u)\|_{\mathrm{HS}}^2 dt$ to obtain

$$d\frac{1}{J}\sum_j \|u^{(j)} - \overline{u}\|^2 = -\frac{2}{J}\|\mathcal{F}(u)\|_{\mathrm{HS}}^2 dt + 2\frac{1}{J}\sum_j \langle u^{(j)} - \overline{u}, \mathcal{F}(u)d(W^{(j)} - \overline{W})\rangle.$$

The martingale term vanishes in expectation if we intergrate up to stopping times such that u remains bounded. Thus, we obtain the first main result of this proof.

For all $t \in [0,T]$, $R > 1$ and $h \in (0,1)$ we have

(B.1)
$$\mathbb{E}\frac{1}{J}\sum_j \|u^{(j)} - \overline{u}\|^2(t \wedge \tau_{R,h}) + \frac{2}{J}\int_0^{t\wedge\tau_{R,h}} \|\mathcal{F}(u)\|_{\mathrm{HS}}^2 ds \le \mathbb{E}\frac{1}{J}\sum_j \|u^{(j)} - \overline{u}\|^2(0).$$

In this result we did not use any particular property of G. It remains to bound $\overline{u}$ now, which is the crucial point that leads to restrictions. First by Itô-s formula

$$\begin{aligned} d\|\overline{u}\|^2 &= 2\langle \overline{u}, d\overline{u}\rangle + \langle d\overline{u}, d\overline{u}\rangle \\ &= 2\langle \overline{u}, \mathcal{F}(u)\Gamma^{-1/2}(\overline{G} - y)\rangle dt + \frac{1}{J}\|\mathcal{F}(u)\|_{\mathrm{HS}}^2 dt + 2\langle \overline{u}, \mathcal{F}(u) d\overline{W}\rangle. \end{aligned}$$

Here, we cannot use cancellations as in the step before. Therefore, we define for $z \ge 0$ the function

$$\varphi(z) = \ln(1+z) \quad \text{with} \quad 0 < \varphi'(z) \le \min\{1, z^{-1}\} \quad \text{and} \quad |\varphi''(z)| \le \min\{1, z^{-2}\}.$$

Again using Itô-s formula, we have

$$\begin{aligned} d\varphi(\|\overline{u}\|^2) &= \varphi'(\|\overline{u}\|^2) d\|\overline{u}\|^2 + \frac{1}{2}\varphi''(\|\overline{u}\|^2) d\|\overline{u}\|^2 d\|\overline{u}\|^2 \\ &= 2\varphi'(\|\overline{u}\|^2)\langle \overline{u}, \mathcal{F}(u)\Gamma^{-1/2}\overline{G}\rangle dt && \text{(B.2)} \\ &\quad - 2\varphi'(\|\overline{u}\|^2)\langle \overline{u}, \mathcal{F}(u)\Gamma^{-1/2} y\rangle dt && \text{(B.3)} \\ &\quad + \frac{1}{J}\varphi'(\|\overline{u}\|^2)\|\mathcal{F}(u)\|_{\mathrm{HS}}^2 dt && \text{(B.4)} \\ &\quad + 2\varphi'(\|\overline{u}\|^2)\langle \overline{u}, \mathcal{F}(u) d\overline{W}\rangle && \text{(B.5)} \\ &\quad + \frac{2}{J}\varphi''(\|\overline{u}\|^2)\langle \overline{u}, \mathcal{F}(u)\mathcal{F}(u)^T \overline{u}\rangle. && \text{(B.6)} \end{aligned}$$

Now we have to bound all terms separately. The martingale term in (B.5) vanishes in expectation if we integrate up to $t \wedge \tau_{R,h}$. Now (B.4) $\le C\|\mathcal{F}(u)\|_{\mathrm{HS}}^2 dt$, which is integrated up to $t \wedge \tau_{R,h}$ in expectation bounded by (B.1). We bound similarly

$$(B.6) \le \frac{2}{J}|\varphi''(\|\overline{u}\|^2)|\|\overline{u}\|^2\|\mathcal{F}(u)\|_{\mathrm{HS}}^2 dt \le C\|\mathcal{F}(u)\|_{\mathrm{HS}}^2 dt$$

and

$$(B.3) \le 2\varphi'(\|\overline{u}\|^2)\|\overline{u}\|\|y\|\|\mathcal{F}(u)\|_{\mathrm{HS}}\|\Gamma^{-1/2}\|_{\mathrm{HS}} dt \le C(1 + \|\mathcal{F}(u)\|_{\mathrm{HS}}^2) dt.$$

The crucial term is (B.2). Here, we have

$$(B.2) \le 2\frac{\|\overline{u}\|\|\overline{G}\|}{1+\|\overline{u}\|^2}\|\mathcal{F}(u)\|_{\mathrm{HS}}\|\Gamma^{-1/2}\|_{\mathrm{HS}} \le \frac{\|\overline{G}\|^2}{1+\|\overline{u}\|^2} + C\|\mathcal{F}(u)\|_{\mathrm{HS}}^2.$$

Now we need to use that G is Lipschitz to obtain

$$\|\overline{G}\| \le C(\frac{1}{J}\sum_j \|u^{(j)}\| + 1) \le C(\frac{1}{J}\sum_j \|u^{(j)} - \overline{u}\| + \|\overline{u}\| + 1)$$

which implies (for constants depending on J)

$$(B.2) \leq C \left(1 + \frac{1}{J} \sum_j \|u^{(j)} - \overline{u}\|^2 + \|\mathcal{F}(u)\|_{\mathrm{HS}}^2 \right.$$

Integrating from 0 to $t \wedge \tau_{R,h}$ we finally obtain together with the bound from (B.1) for all $R > 1$ and $h \in (0,1)$ that $\mathbb{E}\varphi(\|\overline{u}(t \wedge \tau_{R,h})\|^2) \leq C$.

But as φ satisfies $\varphi(x+y) \leq \varphi(x) + y$ we obtain, again using (B.1),

$$\mathbb{E}\varphi(\frac{1}{J} \sum_j \|u^{(j)}(t \wedge \tau_{R,h})\|^2) \leq C,$$

which finishes the proof. ∎

Appendix C. Proofs of section 4. We first refer to the following useful auxiliary result which we will apply at several points.

LEMMA C.1 (see [8, Lemma A.2]). *Let S be a symmetric and nonnegative $d \times d$-matrix; then for all choices of $(z^{(k)})_{k=1,\dots,J}$ in $\mathbb{R}^d$ we have $\sum_{k,l=1}^{J} \langle z^{(k)}, z^{(l)} \rangle \langle z^{(k)}, S z^{(l)} \rangle \geq 0$.*

Proof of Lemma 4.1. First we define the operator

$$M^\varepsilon(u_n) := (hB(C(u_n) + \varepsilon I_p)B^T + I_K)^{-1},$$

for which it holds true that $\lim_{\varepsilon \to 0} M^\varepsilon(u_n) = M(u_n)$, since the mapping $\Sigma \mapsto \Sigma^{-1}$ is continuous over the set of invertible matrices. By

$$C(u_n)B^T M(u_n)\tilde{y} = \frac{1}{J} \sum_{k=1}^{J} \langle B(u_n^{(k)} - \bar{u}_n), M(u_n)\tilde{y} \rangle (u_n^{(k)} - \bar{u}_n),$$

it is sufficient to prove $\langle B(u_n^{(k)} - \bar{u}_n), M(u_n)\tilde{y} \rangle = 0$. We introduce $C^\varepsilon(u_n) := C(u_n) + \varepsilon I_p$ and apply the Woodbury matrix identity

$$\begin{aligned}
&\langle B(u_n^{(k)} - \bar{u}_n), M^\varepsilon(u_n)\tilde{y} \rangle \\
&= \langle B(u_n^{(k)} - \bar{u}_n), \left[I_K^{-1} - hI_K^{-1}B((C^\varepsilon(u_n))^{-1} + hB^T I_K^{-1} B)^{-1} B^T I_K^{-1} \right] \tilde{y} \rangle \\
&= \langle B(u_n^{(k)} - \bar{u}_n), \tilde{y} \rangle - \langle B(u_n^{(k)} - \bar{u}_n), hB((C^\varepsilon(u_n))^{-1} + hB^T B)^{-1} B^T \tilde{y} \rangle \\
&= 0 - \langle hB \left[((C^\varepsilon(u_n))^{-1} + hB^T B)^{-1} \right]^T B^T B(u_n^{(k)} - \bar{u}_n), \tilde{y} \rangle = 0,
\end{aligned}$$

where we have used that $\tilde{y} \in \mathrm{range}(B)^\perp$. Thus

$$\langle B(u_n^{(k)} - \bar{u}_n), M(u_n)\tilde{y} \rangle = \lim_{\varepsilon \to 0} \langle B(u_n^{(k)} - \bar{u}_n), M^\varepsilon(u_n)\tilde{y} \rangle = 0,$$

which concludes the proof. ∎

LEMMA C.2. *Let $u_0 = (u_0^{(j)})_{j \in \{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\Omega \to \mathbb{R}^p$ such that $\mathbb{E}[\|e_0^{(j)}\|^2] < \infty$. Then for all $n \in \mathbb{N}$ it holds true that*

$$\mathbb{E}\left[\frac{1}{J} \sum_{j=1}^{J} \|e_{n+1}^{(j)}\|^2 \right] \leq \mathbb{E}\left[\frac{1}{J} \sum_{j=1}^{J} \|e_n^{(j)}\|^2 \right].$$

Furthermore, there exists the constant $C = \mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}\|e_0^{(j)}\|^2]$ *independent of* h, J, *and* T *such that*

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^{J}\|e_n^{(j)}\|^2\right] \le C \quad \text{for all } n \in \mathbb{N}.$$

Proof. We can derive the evolution of the Euclidean norm by

$$\begin{aligned}\|e_{n+1}^{(j)}\|^2 = \|e_n^{(j)}\|^2 &- 2h\langle e_n^{(j)}, C(u_n)B^T M(u_n)Be_n^{(j)}\rangle \\ &+ 2\sqrt{h}\langle e_n^{(j)}, C(u_n)B^T M(u_n)(W_{n+1}^{(j)} - \bar{W}_{n+1})\rangle \\ &- 2h^{3/2}\langle C(u_n)B^T M(u_n)Be_n^{(j)}, C(u_n)B^T M(u_n)(W_{n+1}^{(j)} - \bar{W}_{n+1})\rangle \\ &+ h^2\|C(u_n)B^T M(u_n)Be_n^{(j)}\|^2 \\ &+ h\|C(u_n)B^T M(u_n)(W_{n+1}^{(j)} - \bar{W}_{n+1})\|^2.\end{aligned}$$

We first write, after plugging in the definition of $C(u_n)$ and inserting

$$M(u_n)(hBC(u_n)B^T + I) \ = I,$$

also abbreviating $M = M(u_n)$ and $C = C(u_n)$,

$$\begin{aligned}-2h\langle e_n^{(j)}, C(u_n)B^T MBe_n^{(j)}\rangle &= -2h\cdot\frac{1}{J}\sum_j \langle e_n^{(j)}, CB^T MBe_n^{(j)}\rangle \\ &= -2h\frac{1}{J}\sum_j \langle e_n^{(j)}, CB^T\cdot M[hBCB^T]\cdot MBe_n^{(j)}\rangle - 2h\frac{1}{J}\sum_j \langle e_n^{(j)}, CB^T\cdot M\cdot MBe_n^{(j)}\rangle \\ &= -2h^2\frac{1}{J}\sum_j \langle B^T MBCe_n^{(j)}, CB^T MBe_n^{(j)}\rangle - 2h\frac{1}{J}\sum_j \langle MBCe_n^{(j)}, MBe_n^{(j)}\rangle.\end{aligned}$$

Defining $Z = B^T MB$ (this proof works for any self-adjoint matrix) it is easy to verify

$$\frac{1}{J}\sum_j \langle ZCe^{(j)}, CZe^{(j)}\rangle = \frac{1}{J}\sum_l \|CZe^{(l)}\|^2$$

and we can continue to write

$$\begin{aligned}-2h\cdot\frac{1}{J}\sum_j \langle e_n^{(j)}, CB^T MBe_n^{(j)}\rangle &= -2h^2\frac{1}{J}\sum_j \|CB^T MBe_n^{(j)}\|^2 \\ &\quad - 2h\frac{1}{J}\sum_{j,k}\langle e_n^{(k)}, e_n^{(j)}\rangle\langle MBe_n^{(k)}, MBe_n^{(j)}\rangle,\end{aligned}$$

where we have used the definition of $C(u_n)$ for the second term.

We define $S := MB$ and use $\mathbb{E}\langle a, W_i\rangle\langle b, W_j\rangle = \delta_{i,j}\langle a, b\rangle$ in order to derive

$$\begin{aligned}&\mathbb{E}\left[\left\|CB^T M(W_{n+1}^{(j)} - \bar{W}_{n+1})\right\|^2 \mid \mathcal{F}_n\right] \\ &= \mathbb{E}\left[\frac{1}{J^2}\sum_{k,l}\left\langle e^{(k)}\langle e^{(k)}, S^T(W_{n+1}^{(j)} - \bar{W}_{n+1})\rangle, e^{(l)}\langle e^{(l)}, S^T(W_{n+1}^{(j)} - \bar{W}_{n+1})\rangle\right\rangle \mid \mathcal{F}_n\right] \\ &= \frac{1}{J}\sum_{k,l}\langle e^{(k)}, e^{(l)}\rangle\left[\frac{J-1}{J^2}\langle Se^{(k)}, Se^{(l)}\rangle + \frac{(J-1)^2}{J^2}\langle Se^{(k)}, Se^{(l)}\rangle\right] \\ &= \frac{1}{J}\sum_{l,k}\langle e^{(l)}, e^{(k)}\rangle\langle MBe^{(l)}, MBe^{(k)}\rangle\cdot\frac{J-1}{J}.\end{aligned}$$

We take the expectation up to step n above to obtain

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{J}\sum_j \|e_{n+1}^{(j)}\|^2 - \|e_n^{(j)}\|^2 \mid \mathcal{F}_n\right] \\
&= -2h^2\frac{1}{J}\sum_j \|CB^TMBe_n^{(j)}\|^2 - 2h\frac{1}{J}\sum_{j,k}\langle e_n^{(k)}, e_n^{(j)}\rangle\langle MBe_n^{(k)}, MBe_n^{(j)}\rangle \\
&\quad + 0 + 0 + h^2\frac{1}{J}\sum_j \|CB^TMBe_n^{(j)}\|^2 + \frac{J-1}{J^2}\cdot h\sum_{j,k}\langle e_n^{(k)}, e_n^{(j)}\rangle\langle MBe_n^{(k)}, MBe_n^{(j)}\rangle \\
&= -h^2\frac{1}{J}\sum_j \|CB^TMBe_n^{(j)}\|^2 - \frac{J+1}{J}h\|CB^\top M\|_{\mathrm{HS}}^2 \quad \le 0,
\end{aligned}$$

where positivity of the last sum follows from lemma C.1 by setting $S = B^TM^2B$. The process $(\frac{1}{J}\sum_{j=1}^J \|e_n^{(j)}\|)_{n\in\mathbb{N}}$ is a supermartingale, and the assertion follows. ∎

Similarly, the next result states the bound of the particle deviation mapped by B.

COROLLARY C.3. *Let $u_0 = (u_0^{(j)})_{j\in\{1,\dots,J\}}$ be $\mathcal{F}_0$-measurable maps $\Omega \to \mathbb{R}^p$ such that $\mathbb{E}[\|Be_0^{(j)}\|^2] < \infty$. Then for all $n \in \mathbb{N}$ it holds true that*

$$\mathbb{E}\left[\frac{1}{J}\sum_{j=1}^J \|Be_{n+1}^{(j)}\|^2\right] \le \mathbb{E}\left[\frac{1}{J}\sum_{j=1}^J (\|Be_n^{(j)}\|^2\right].$$

Proof. The proof follows by similar computations as in the proof of Lemma C.2. ∎

For our last auxiliary result, we recall that the update of the residuals can be written as $r_{n+1}^{(j)} = r_n^{(j)} - hC(u_n)B^TM(u_n)Br_n^{(j)} + \sqrt{h}C(u_n)B^TM(u_n)W_{n+1}^{(j)}$ and provide the boundedness of the residuals in the observation space, which is formulated in the following lemma.

PROPOSITION C.4. *For all $n \in \mathbb{N}$ it holds true that*

$$\frac{1}{J}\sum_{j=1}^J \mathbb{E}[\|Br_{n+1}^{(j)}\|^2 + \|Be_{n+1}^{(j)}\|^2] \le \frac{1}{J}\sum_{j=1}^J \mathbb{E}[(\|Br_n^{(j)}\|^2 + \|Be_n^{(j)}\|^2)].$$

Proof. The update of the mapped residuals is given by

$$Br_{n+1}^{(j)} = Br_n^{(j)} - hBC(u_n)B^TM(u_n)Br_n^{(j)} + \sqrt{h}BC(u_n)B^TM(u_n)W_{n+1}^{(j)}.$$

Using $M(u_n)(hBC(u_n)B^\top + I) = I$ and abbreviating again $M = M(u_n)$ and $C = C(u_n)$, we obtain

$$\begin{aligned}
&\mathbb{E}[\|Br_{n+1}^{(j)}\|^2 - \|Br_n^{(j)}\|^2 \mid \mathcal{F}_n] \\
&\qquad = -2h\langle Br_n^{(j)}, BCB^TMBr_n^{(j)}\rangle + 0 + 0 + h^2\|BCB^TMBr_n^{(j)}\|^2 \\
&\qquad\quad + h\|BCB^TMW_{n+1}^{(j)}\|^2 \\
&\qquad = -2h\langle Br_n^{(j)}, M(hBCB^\top + I)BCB^TMBr_n^{(j)}\rangle \\
&\qquad\quad + h^2\|BCB^TMBr_n^{(j)}\|^2 + h\mathbb{E}\left[\|BCB^TMW_{n+1}^{(j)}\|^2 \mid \mathcal{F}_n\right]
\end{aligned}$$

$$
\begin{aligned}
&= -2h^2\langle Br_n^{(j)}, MBCB^\top BCB^T MBr_n^{(j)}\rangle - 2h\langle Br_n^{(j)}, MBCB^T MBr_n^{(j)}\rangle \\
&\quad + h^2\|BCB^T MBr_n^{(j)}\|^2 + h\mathbb{E}\left[\|BCB^T MW_{n+1}^{(j)}\|^2 \mid \mathcal{F}_n\right] \\
&= -2h^2\langle BCB^\top MBr_n^{(j)}, BCB^T MBr_n^{(j)}\rangle \\
&\quad - 2h\langle C^{1/2}B^\top MBr_n^{(j)}, C^{1/2}B^T MBr_n^{(j)}\rangle \\
&\quad + h^2\|BCB^T MBr_n^{(j)}\|^2 + h\mathbb{E}\left[\|BCB^T MW_{n+1}^{(j)}\|^2 \mid \mathcal{F}_n\right] \\
&= -2h^2\|BCB^\top MBr_n^{(j)}\|^2 - 2h\|C^{1/2}B^\top MBr_n^{(j)}\|^2 \\
&\quad + h^2\|BCB^T MBr_n^{(j)}\|^2 + h\mathbb{E}\left[\|BCB^T MW_{n+1}^{(j)}\|^2 \mid \mathcal{F}_n\right].
\end{aligned}
$$

We note that

$$
\mathbb{E}\left[\|BCB^T MW_{n+1}^{(j)}\|^2 \mid \mathcal{F}_n\right] = h\frac{1}{J^2}\sum_{l=1}^{J}\|C^{1/2}B^\top MBe_n^{(l)}\|^2.
$$

Similarly as in the proof of Lemma C.2 we obtain

$$
\begin{aligned}
\frac{1}{J}\sum_{j=1}^{J}\mathbb{E}[\|Be_{n+1}^{(j)}\|^2 - \|Be_n^{(j)}\|^2 \mid \mathcal{F}_n] &= -h^2\frac{1}{J}\sum_{j=1}^{J}\|CB^T MBe_n^{(j)}\|^2 \\
&\quad - h\frac{J+1}{J^2}\sum_{j=1}^{J}\|C^{1/2}B^\top MBe_n^{(j)}\|^2.
\end{aligned}
$$

We conclude with

$$
\begin{aligned}
&\mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}(\|Br_{n+1}^{(j)}\|^2 + \|Be_{n+1}^{(j)}\|^2) - \frac{1}{J}\sum_{j=1}^{J}(\|Br_n^{(j)}\|^2 + \|Be_n^{(j)}\|^2) \mid \mathcal{F}_n] \\
&= -h^2\frac{1}{J}\sum_{j=1}^{J}\|BCB^\top MBr_n^{(j)}\|^2 - 2h\frac{1}{J}\sum_{j=1}^{J}\|C^{1/2}B^\top MBr_n^{(j)}\|^2 \\
&\quad - h^2\frac{1}{J}\sum_{j=1}^{J}\|CB^T MBe_n^{(j)}\|^2 - h\frac{1}{J^2}\sum_{j=1}^{J}\|C^{1/2}B^\top MBe_n^{(j)}\|^2 \quad \le 0. \qquad \text{(C.1)}
\end{aligned}
$$

∎

While proving the above two auxiliary results, we have derived explicit update formulas for $\frac{1}{J}\sum_{j=1}^{J}\mathbb{E}[\|e_n^{(j)}\|^2]$ and $\frac{1}{J}\sum_{j=1}^{J}\mathbb{E}[\|Br_n^{(j)}\|^2] + \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}[\|Be_n^{(j)}\|^2]$. Using these explicit update formulas, we are further able to bound the following summations.

COROLLARY C.5. *For all $n \in \mathbb{N}$ it holds true that*

$$
\frac{J+1}{J}\sum_{k=0}^{n-1}h\mathbb{E}[\|C(u_k)B^\top M(u_k)\|_{\mathrm{HS}}^2] \le \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\left[\|e_0^{(j)}\|^2\right],
$$

and

$$
\sum_{k=0}^{n-1}h\frac{1}{J}\sum_{j=1}^{J}\mathbb{E}[\|C(u_k)^{1/2}B^\top M(u_k)Br_k^{(j)}\|^2] \le \frac{1}{2J}\sum_{j=1}^{J}\mathbb{E}[\|Br_0^{(j)}\|^2 + \|Be_0^{(j)}\|^2].
$$

Proof. From the proof of Lemma C.2 we know that

$$0 \leq \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\left[\|e_n^{(j)}\|^2\right] = \mathbb{E}\left[\|e_0^{(j)}\|^2\right] - \sum_{k=0}^{n-1} h^2 \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}[\|CB^T MBe_k^{(j)}\|^2] - \sum_{k=0}^{n-1}\frac{J+1}{J}h\mathbb{E}[\|CB^\top M\|_{\mathrm{HS}}^2],$$

and it implies that for all $n \in \mathbb{N}$ we have that

$$\sum_{k=0}^{n-1}\frac{J+1}{J}h\|CB^\top M\|_{\mathrm{HS}}^2 \leq \mathbb{E}\left[\|e_0^{(j)}\|^2\right].$$

The other bound follow similarly by using the update formula

$$\begin{aligned}
0 &\leq \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}[\|Br_n^{(j)}\|^2 + \|Be_n^{(j)}\|^2 \mid \mathcal{F}_n] \\
&= \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}[\|Br_0^{(j)}\|^2 + \|Be_0^{(j)}\|^2] \sum_{k=0}^{n-1} h^2 \frac{1}{J}\sum_{j=1}^{J}\|BCB^\top MBr_k^{(j)}\|^2 \\
&\quad - 2\sum_{k=0}^{n-1} h\frac{1}{J}\sum_{j=1}^{J}\mathbb{E}[\|C^{1/2}B^\top MBr_k^{(j)}\|^2] - \sum_{k=0}^{n-1} h^2 \frac{1}{J}\sum_{j=1}^{J}\mathbb{E}[\|CB^T MBe_k^{(j)}\|^2] \\
&\quad - \sum_{k=0}^{n-1} h\frac{1}{J^2}\sum_{j=1}^{J}\mathbb{E}[\|C^{1/2}B^\top MBe_k^{(j)}\|^2].
\end{aligned}$$

∎

We emphasize that it is not true that the quantity $\frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\|r_n^{(j)}\|^2$ is decreasing. This can be seen directly in the continuous and deterministic setting: Here it can be proven that $\frac{1}{J}\sum_{j=1}^{J}\|Br^{(j)}(t)\|^2$ is decreasing, but $\frac{1}{J}\sum_{j=1}^{J}\|r^{(j)}(t)\|^2$ does not have this property.

First, the mapping via B only keeps track of the data-informed parameter dimensions, i.e., those orthogonal to the kernel of A. And secondly, even invertibility of B still does not imply monotonicity of $\|\overline{u}(t) - \hat{u}\|$ as the mapping B can warp the coordinate system in such a way that this property is lost, with $\hat{u}$ defined as in (4.1). This can be seen in an elementary example unrelated to the EKI: Consider the curve $x(t) = (\cos(t), \sin(t))$ for which $V(t) := \|x(t)\|^2$ is constant, i.e., monotonously decreasing. On the other hand, with $A = \mathrm{diag}(2,1)$ and $\Gamma = E_{2\times 2}$, the mapping $\tilde{V}(t) = \|Ax(t)\|^2$ is not monotonous.

As a concrete example for the nonmonotonicity of the mean and the residual, we can consider the forward operator $A = \mathrm{diag}(100, 1)$, observation noise covariance $\Gamma = I_{2\times 2}$, observation $y = (0,0)^T$, and an initial ensemble with mean $\overline{u}_0 = (100, 100)^T$ and empirical covariance $C(u(0)) = \left(\begin{smallmatrix} 25 & -24 \\ -24 & 25 \end{smallmatrix}\right)$, whose eigenvectors are $(-1,1)^T$ and $(1,1)^T$ with eigenvalues 49 and 1, respectively.

Figure 1 shows the initial ensemble and the trajectories of the ensemble and its sample mean in the parameter space. Clearly, the sample mean and the whole ensemble move away from their final limit $(0,0)^T$ for quite some time until they finally "change direction" and converge towards their limit. The initial shearing of

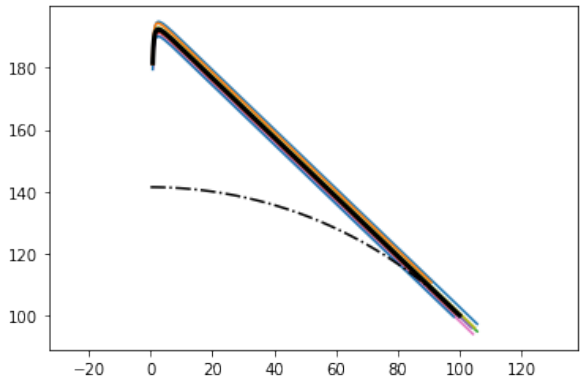

FIG. 1. *Trajectories of the EKI (starting at the lower right corner; black curve is the mean $\overline{u}(t)$) for $t \in [0,1]$. Dotted sphere is the Euclidean sphere through $\overline{u}_0$, demonstrating non-monotonicity of the mean.*

the ensemble combined with the strong weighting of the horizontal direction, which is encoded in the forward operator, leads to an initial movement of the ensemble along its principal axis to the top left, increasing the value of $\|\overline{r}(t)\|^2$.

In other words, the Euclidean norm is not the natural norm with respect to which we should view the dynamics of the ensemble, and either we need to settle for *nonmonotonous convergence* of the residuals $\|\overline{u}(t) - \hat{u}\|$ in parameter space, or we need to pick a more problem-adapted norm. In the deterministic setting, the latter can be done by diagonalizing $C(u(0))B^TB$: It can be shown that this yields a basis of eigenvectors which diagonalize $C(u(t))C(u(0))^{-1}$ for all times; see [9]. In the stochastic setting, this favorable property is lost.

Proof of Lemma 4.3. Let $p = 1$, and write

$$\begin{aligned}\|r_{n+1}\|_{L_p} &:= \mathbb{E}[\|r_{n+1}^{(j)}\|^p]^{1/p} \\ &= \|r_n^{(j)} - hC(u_n)B^\top M(u_n)Br_n^{(j)} + \sqrt{h}C(u_n)B^\top M(u_n)W_{n+1}^{(j)}\|_{L_p} \\ &\le \|r_n^{(j)}\|_{L_1} + \|C(u_n)^{1/2}\|_{L_2}\|hC(u_n)^{1/2}B^\top M(u_n)Br_n^{(j)}\|_{L_2} \\ &\quad + \|\sqrt{h}C(u_n)B^\top M(u_n)W_{n+1}^{(j)}\|_{L_1}.\end{aligned}$$

First, note that we can write

$$(C(u_n))^{1/2} = (1/J \cdot (e_n^{(1)}, e_n^{(2)}, \dots, e_n^{(J)})(e_n^{(1)}, e_n^{(2)}, \dots, e_n^{(J)})^\top)^{1/2}$$

and hence, it holds true that

$$\|C(u_n)^{1/2}\|_{L_2} \le \left(\frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\|e_n^{(j)}\|^2\right)^{1/2} \le \left(\frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\|e_0^{(j)}\|^2\right)^{1/2} =: C_1.$$

Furthermore, we can bound

$$\begin{aligned}\|r_n^{(j)}\|_{L_1} \le \|r_0^{(j)}\|_{L_1} &+ C_1\sum_{k=0}^{n}\|hC(u_k)^{1/2}B^\top M(u_k)Br_k^{(j)}\|_{L_2} \\ &+ \sum_{k=0}^{n}\|\sqrt{h}C(u_n)B^\top M(u_n)W_{n+1}^{(j)}\|_{L_1}.\end{aligned}$$

From Corollary C.5 we have that for all $n \geq 1$

$$2\sum_{k=0}^{n} h\mathbb{E}[\|C(u_k)^{1/2}B^\top M(u_k)Br_k^{(j)}\|^2] \leq \mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}\|Br_0^{(j)}\|^2 + \|Be_0^{(j)}\|^2],$$

and it follows by Jensen's inequality that

$$\begin{aligned}\sum_{k=0}^{n}\|hC(u_k)^{1/2}B^\top M(u_k)Br_k^{(j)}\|_{L_2} &\leq \sum_{k=0}^{N\cdot T}\|hC(u_k)^{1/2}B^\top M(u_k)Br_k^{(j)}\|_{L_2}\\ &\leq T\cdot\left(\sum_{k=0}^{N\cdot T}\frac{1}{T}h\mathbb{E}[\|C(u_k)^{1/2}B^\top M(u_k)Br_k^{(j)}\|^2]\right)^{1/2}\\ &\leq \sqrt{T}\cdot\mathbb{E}\left[1/2\frac{1}{J}\sum_{j=1}^{J}\|Br_0^{(j)}\|^2 + \|Be_0^{(j)}\|^2\right]^{1/2}\end{aligned}$$

providing a uniform bound in h for all n. Similarly, we obtain from Corollary C.5 that

$$\begin{aligned}\frac{J+1}{J}\sum_{k=0}^{n}\mathbb{E}[\|\sqrt{h}C(u_k)B^\top M(u_k)W_{k+1}^{(j)}\|^2] &= \frac{J+1}{J}\sum_{k=0}^{n}\mathbb{E}[h\|C(u_k)B^\top M(u_k)\|_{\mathrm{HS}}^2]\\ &\leq \mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}\|e_0^{(j)}\|^2],\end{aligned}$$

and applying again Jensen's inequality gives

$$\begin{aligned}\sum_{k=0}^{n}\|\sqrt{h}C(u_n)B^\top M(u_n)W_{n+1}^{(j)}\|_{L_1} &\leq \sum_{k=0}^{N\cdot T}\mathbb{E}[\|\sqrt{h}C(u_n)B^\top M(u_n)W_{n+1}^{(j)}\|^2]^{1/2}\\ &\leq T\cdot\left(\sum_{k=0}^{N\cdot T}\frac{1}{T}h\mathbb{E}[\|C(u_k)B^\top M(u_k)\|_{\mathrm{HS}}^2]\right)^{1/2}\\ &\leq \sqrt{T}\cdot\left(\frac{J}{J+1}\frac{1}{J}\sum_{j=1}^{J}\mathbb{E}[\|e_0^{(j)}\|^2]\right)^{1/2}.\end{aligned}$$

We conclude the proof by

$$\begin{aligned}\|r_n^{(j)}\|_{L_1} \leq \|r_0^{(j)}\|_{L_1}&\\ &+\sqrt{T}\cdot\left(\frac{1}{J}\sum_{j=1}^{J}\mathbb{E}\|e_0^{(j)}\|^2\right)^{1/2}\cdot\left(\frac{1}{2J}\sum_{j=1}^{J}\mathbb{E}[\|Br_0^{(j)}\|^2+\|Be_0^{(j)}\|^2]\right)^{1/2}\\ &+\sqrt{T}\cdot\left(\frac{J}{J+1}\frac{1}{J}\sum_{j=1}^{J}\mathbb{E}[\|e_0^{(j)}\|^2]\right)^{1/2}.\end{aligned}$$

∎

Proof of Lemma 4.6. We first note that $\mathbb{E}[\|r_n^{(j)}\|^2] = \mathbb{E}[\|Pr_n^{(j)}\|^2] + \mathbb{E}[\|(I-P)r_n^{(j)}\|^2]$, and we consider both terms separately.

Step 1: Bounding $\mathbb{E}[\|Pr_n^{(j)}\|^2]$.
We observe that

$$\|Pr_n^{(j)}\|^2 = \|B^\top (BB^\top)^- Br_n^{(j)}\|^2 \le \|B^\top (BB^\top)^-\|_{\mathrm{HS}}^2 \|Br_n^{(j)}\|^2.$$

Application of Proposition C.4 gives the uniform bound in n and h, i.e., $\|Pr_n^{(j)}\|^2 \le c_1$ for some $c_1 > 0$ independent of n and h.

Step 2: Bounding $\mathbb{E}[\|(I-P)r_n^{(j)}\|^2]$.
For the update of $(I-P)r_n^{(j)}$ we have that

$$\begin{aligned}(I-P)r_{n+1}^{(j)} &= (I-P)r_n^{(j)} - h(I-P)C(u_n)B^\top M(u_n)Br_n^{(j)} \\ &\quad + \sqrt{h}(I-P)C(u_n)B^\top M(u_n)W_{n+1}^{(j)} \\ &= (I-P)r_n^{(j)} \\ &\quad + \frac{1}{J}\sum_{k=1}^{J}\langle -hM(u_n)Br_n^{(j)} + \sqrt{h}M(u_n)W_{n+1}^{(j)}, Be_n^{(k)}\rangle (I-P)e_n^{(k)}.\end{aligned}$$

Similarly, we have that

$$\begin{aligned}(I-P)e_{n+1}^{(j)} &= (I-P)e_n^{(j)} \\ &\quad + \frac{1}{J}\sum_{k=1}^{J}\langle -hM(u_n)Be_n^{(j)} + \sqrt{h}M(u_n)W_{n+1}^{(j)}, Be_n^{(k)}\rangle (I-P)e_n^{(k)} \\ &= (I-P)e_0^{(j)}.\end{aligned}$$

Hence, we imply that $(I-P)e_n^{(k)} = 0$ for all k, i.e., $e_n^{(k)}$ is in the range of P, and it follows that

$$(I-P)r_{n+1}^{(j)} = (I-P)r_n^{(j)} = (I-P)r_0^{(j)}.$$

Finally, we conclude with

$$\|(I-P)r_{n+1}^{(j)}\|^2 = \|(I-P)r_0^{(j)}\|^2 \le c_2.$$

∎

Proof of Lemma 4.9. We again decompose $\tilde{y} = \hat{y} + y'$, where $\hat{y} \in \mathrm{range}(\tilde{B})$ and $y' \in \mathrm{range}(\tilde{B})^\perp$. By Lemma 4.1 there exists $\hat{u}$ (not necessarily unique), such that we can write the update for $r_n^{(j)} = u_n^{(j)} - \hat{u}$ by

$$r_{n+1}^{(j)} = r_n^{(j)} - hC(u_n)\tilde{B}^T M(u_n)\tilde{B}r_n^{(j)} + \sqrt{h}C(u_n)\tilde{B}^T M(u_n)W_{n+1}^{(j)}.$$

By Proposition C.4 it follows that

$$\begin{aligned}\sup_{n\in\{1,\dots,N\}} &\frac{1}{J}\sum_{j=1}^{J}\mathbb{E}[\|\tilde{B}r_{n+1}^{(j)}\|_{\mathbb{R}^K\times\mathcal{X}}^2 + \|\tilde{B}e_{n+1}^{(j)}\|_{\mathbb{R}^K\times\mathcal{X}}^2] \\ &\le \mathbb{E}[\frac{1}{J}\sum_{j=1}^{J}(\|\tilde{B}r_n^{(j)}\|_{\mathbb{R}^K\times\mathcal{X}}^2 + \|\tilde{B}e_n^{(j)}\|_{\mathbb{R}^K\times\mathcal{X}}^2)].\end{aligned}$$

The definition of $\tilde{B}$ implies that

$$\|\tilde{B}r_{n+1}^{(j)}\|_{\mathbb{R}^K\times\mathcal{X}}^2 = \|B(u_{n+1}^{(j)} - \hat{u})\|_{\mathbb{R}^K}^2 + \|(u_{n+1}^{(j)} - \hat{u})\|_{\mathcal{X}}^2,$$

and hence, we conclude with

$$\sup_{n\in\{1,\dots,N\}} \mathbb{E}[\|u_n^{(j)}\|^2] \leq C$$

for all $j \in \{1,\dots,J\}$, where $C > 0$ is independent from h. ∎

REFERENCES

[1] D. Armbruster, M. Herty, and G. Visconti, *A Stabilization of a Continuous Limit of the Ensemble Kalman Filter*, preprint, arXiv: 2020, https://arxiv.org/abs/2006.15390.

[2] Y. Ba, J. de Wiljes, D. S. Oliver, and S. Reich, *Randomized maximum likelihood based posterior sampling*, Comput. Geosci., 26 (2022), pp. 217–239, https://doi.org/10.1007/s10596-021-10100-y.

[3] J. M. Bardsley, A. Solonen, H. Haario, and M. Laine, *Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems*, SIAM J. Sci. Comput., 36 (2014), pp. A1895–A1910, https://doi.org/10.1137/140964023.

[4] M. Benning and M. Burger, *Modern regularization methods for inverse problems*, Acta Numer., 27 (2018), pp. 1–111, https://doi.org/10.1017/S0962492918000016.

[5] K. Bergemann and S. Reich, *A localization technique for ensemble Kalman filters*, Q. J. R. Meteorol. Soc., 136 (2010), pp. 701–707, https://doi.org/10.1002/qj.591.

[6] K. Bergemann and S. Reich, *A mollified ensemble Kalman filter*, Q. J. R. Meteorol. Soc., 136 (2010), pp. 1636–1643, https://doi.org/10.1002/qj.672.

[7] D. Blömker, C. Schillings, and P. Wacker, *A strongly convergent numerical scheme from ensemble Kalman inversion*, SIAM J. Numer. Anal., 56 (2018), pp. 2537–2562, https://doi.org/10.1137/17M1132367.

[8] D. Blömker, C. Schillings, P. Wacker, and S. Weissmann, *Well posedness and convergence analysis of the ensemble Kalman inversion*, Inverse Probl., 35 (2019), 085007, https://doi.org/10.1088/1361-6420/ab149c.

[9] L. Bungert and P. Wacker, *Complete Deterministic Dynamics and Spectral Decomposition of the linear Ensemble Kalman Inversion*, preprint, arXiv:2104.13281 [math.NA], 2021, https://arxiv.org/abs/2104.13281.

[10] N. K. Chada, A. Jasra, and F. Yu, *Multilevel Ensemble Kalman-Bucy Filters*, preprint, arXiv:2011.04342 [math.NA], 2021, https://arxiv.org/abs/2011.04342.

[11] N. K. Chada, C. Schillings, and S. Weissmann, *On the incorporation of box-constraints for ensemble Kalman inversion*, Found. Data Sci., 1 (2019), pp. 433–456, https://doi.org/10.3934/fods.2019018.

[12] N. K. Chada, A. M. Stuart, and X. T. Tong, *Tikhonov regularization within ensemble Kalman inversion*, SIAM J. Numer. Anal., 58 (2020), pp. 1263–1294, https://doi.org/10.1137/19M1242331.

[13] N. K. Chada and X. T. Tong, *Convergence Acceleration of Ensemble Kalman Inversion in Nonlinear Settings*, preprint, arXiv:1911.02424 [math.NA], 2019, https://arxiv.org/abs/1911.02424.

[14] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock, *An Introduction to Total Variation for Image Analysis*, De Gruyter, Berlin, 2010, https://doi.org/10.1515/9783110226157.263.

[15] Y. Chen and D. S. Oliver, *Ensemble randomized maximum likelihood method as an iterative ensemble smoother*, Math. Geosci., 44 (2012), pp. 1–26, https://doi.org/10.1007/s11004-011-9376-z.

[16] A. Chernov, H. Hoel, K. Law, F. Nobile, and R. Tempone, *Multilevel Ensemble Kalman Filtering for Spatially Extended Models*, preprint, arXiv:1608.08558 [math.NA], 2016, https://arxiv.org/abs/1608.08558.

[17] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, UK, 2014.

[18] J. de Wiljes, S. Reich, and W. Stannat, *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 (2018), pp. 1152–1181, https://doi.org/10.1137/17M1119056.

[19] P. Del Moral and J. Tugaut, *On the stability and the uniform propagation of chaos properties of ensemble Kalman Bucy filters*, Ann. Appl. Probab., 28 (2018), pp. 790–850, https://doi.org/10.1214/17-AAP1317.

[20] Z. Ding and Q. Li, *Ensemble Kalman inversion: Mean-field limit and convergence analysis*, Stat. Comput., 31 (2021), 9, https://doi.org/10.1007/s11222-020-09976-0.

[21] Z. Ding and Q. Li, *Ensemble Kalman sampler: Mean-field limit and convergence analysis*, SIAM J. Math. Anal., 53 (2021), pp. 1546–1578, https://doi.org/10.1137/20M1339507.

[22] Z. Ding, Q. Li, and J. Lu, *Ensemble Kalman inversion for nonlinear problems: Weights, consistency, and variance bounds*, Found. Data Sci., 3 (2021), pp. 371–411, https://doi.org/10.3934/fods.2020018.

[23] A. A. Emerick and A. C. Reynolds, *Ensemble smoother with multiple data assimilation*, Comput. Geosci., 55 (2013), pp. 3–15.

[24] H. Engl, M. Hanke, and G. Neubauer, *Regularization of Inverse Problems*, Math. Appl., Springer, Cham, 1996, https://books.google.de/books?id=DF7R.

[25] H. W. Engl, K. Kunisch, and A. Neubauer, *Convergence rates for Tikhonov regularisation of non-linear ill-posed problems*, Inverse Probl., 5 (1989), pp. 523–540, https://doi.org/10.1088/0266-5611/5/4/007.

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

[27] G. Evensen, *Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics*, J. Geophys. Res. Oceans, 99 (1994), pp. 10143–10162, https://doi.org/10.1029/94JC00572.

[28] G. Evensen, *The ensemble Kalman filter: Theoretical formulation and practical implementation*, Ocean Dynam., 53 (2003), pp. 343–367, https://doi.org/10.1007/s10236-003-0036-9.

[29] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart, *Interacting Langevin diffusions: Gradient structure and ensemble Kalman sampler*, SIAM J. Appl. Dyn. Syst., 19 (2020), pp. 412–441, https://doi.org/10.1137/19M1251655.

[30] A. Garbuno-Inigo, N. Nüsken, and S. Reich, *Affine invariant interacting Langevin dynamics for Bayesian inference*, SIAM J. Appl. Dyn. Syst., 19 (2020), pp. 1633–1658, https://doi.org/10.1137/19M1304891.

[31] P. A. Guth, C. Schillings, and S. Weissmann, *Ensemble Kalman Filter for Neural Network Based One-shot Inversion*, preprint, arXiv:2005.02039 [math.NA], 2020, https://arxiv.org/abs/2005.02039.

[32] M. Herty and G. Visconti, *Kinetic methods for inverse problems*, Kinet. Relat. Models, 12 (2019), pp. 1109–1130, https://doi.org/10.3934/krm.2019042.

[33] D. J. Higham, X. Mao, and A. M. Stuart, *Strong convergence of Euler-type methods for nonlinear stochastic differential equations*, SIAM J. Numer. Anal., 40 (2002), pp. 1041–1063, https://doi.org/10.1137/S0036142901389530.

[34] H. Hoel, K. Law, and R. Tempone, *Multilevel ensemble Kalman filtering*, SIAM J. Numer. Anal., 54 (2016), pp. 1813–1839, https://doi.org/10.1137/15M100955X.

[35] H. Hoel, G. Shaimerdenova, and R. Tempone, *Multilevel ensemble Kalman filtering based on a sample average of independent EnKF estimators*, Found. Data Sci., 2 (2020), pp. 351–390, https://doi.org/10.3934/fods.2020017.

[36] M. Hutzenthaler and A. Jentzen, *Numerical Approximations of Stochastic Differential Equations with Non-Globally Lipschitz Continuous Coefficients*, American Mathematical Society, Providence, RI, 2015.

[37] M. Hutzenthaler, A. Jentzen, and P. E. Kloeden, *Strong and weak divergence in finite time of Euler's method for stochastic differential equations with non-globally Lipschitz continuous coefficients*, Proc. A, 467 (2011), pp. 1563–1576, https://doi.org/10.1098/rspa.2010.0348.

[38] M. Hutzenthaler, A. Jentzen, and P. E. Kloeden, *Strong convergence of an explicit numerical method for sdes with nonglobally Lipschitz continuous coefficients*, Annal. Appl. Probab., 22 (2012), pp. 1611–1641 http://www.jstor.org/stable/41713370.

[39] M. Iglesias and Y. Yang, *Adaptive Regularisation for Ensemble Kalman Inversion*, preprint, arXiv:2006.14980 [math.NA], 2020, https://arxiv.org/abs/2006.14980.

[40] M. A. Iglesias, *Iterative regularization for ensemble data assimilation in reservoir models*, Comput. Geosci., 19 (2015), pp. 177–212, https://doi.org/10.1007/s10596-014-9456-5.

[41] M. A. Iglesias, *A regularizing iterative ensemble Kalman method for PDE-constrained inverse problems*, Inverse Probl., 32 (2016), 025002, http://stacks.iop.org/0266-5611/32/i=2/a=025002.

[42] M. A. Iglesias, K. Law, and A. M. Stuart, *Ensemble Kalman methods for inverse problems*, Inverse Probl., 29 (2013), 045001, http://stacks.iop.org/0266-5611/29/i=4/a=045001.
[43] D. Kelly, K. Law, and A. M. Stuart, *Well-posedness and accuracy of the ensemble Kalman filter in discrete and continuous time*, Nonlinearity, 27 (2014), p. 2579, http://stacks.iop.org/0951-7715/27/i=10/a=2579.
[44] D. Kelly, A. J. Majda, and X. T. Tong, *Nonlinear stability and ergodicity of ensemble based Kalman filters*, Nonlinearity, 29 (2016), p. 657, http://stacks.iop.org/0951-7715/29/i=2/a=657.
[45] R. Z. Khasminskii, *Stochastic Stability of Differential Equations*, Monographs and Textbooks on Mechanics of Solids and Fluids, Sijthoff & Noordhoff, Alphen aan den Rijn, The Netherlands, 1980.
[46] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer, Cham, 1992.
[47] N. B. Kovachki and A. M. Stuart, *Ensemble Kalman inversion: A derivative-free technique for machine learning tasks*, Inverse Probl., 35 (2019), 095005, https://doi.org/10.1088/1361-6420/ab1c3a.
[48] E. Kwiatkowski and J. Mandel, *Convergence of the square root ensemble Kalman filter in the large ensemble limit*, SIAM/ASA J. Uncertain. Quantif., 3 (2015), pp. 1–17, https://doi.org/10.1137/140965363.
[49] T. Lange, *Derivation of ensemble Kalman-Bucy filters with unbounded nonlinear coefficients*, Nonlinearity, 35 (2021), p. 1061.
[50] T. Lange and W. Stannat, *Mean field limit of ensemble square root filters - discrete and continuous time*, Found. Data Sci., 3 (2021), pp. 563–588, https://doi.org/10.3934/fods.2021003.
[51] T. Lange and W. Stannat, *On the Continuous Time Limit of Ensemble Square Root Filters*, preprint, arXiv:1910.12493 [math.PR], 2021, https://arxiv.org/abs/1910.12493.
[52] T. Lange and W. Stannat, *On the continuous time limit of the ensemble Kalman filter*, Math. Comput., 90 (2021), pp. 233–265, https://doi.org/10.1090/mcom/3588.
[53] K. Law, H. Tembine, and R. Tempone, *Deterministic mean-field ensemble Kalman filtering*, SIAM J. Sci. Comput., 38 (2016), pp. A1251–A1279, https://doi.org/10.1137/140984415.
[54] F. Le Gland, V. Monbet, and V.-D. Tran, *Large Sample Asymptotics for the Ensemble Kalman Filter*, Research Report RR-7014, INRIA, 2009, https://hal.inria.fr/inria-00409060.
[55] G. J. Lord, C. E. Powell, and T. Shardlow, *An Introduction to Computational Stochastic PDEs*, Cambridge University Press, Cambridge, UK, 2014.
[56] A. J. Majda and X. T. Tong, *Performance of ensemble Kalman filters in large dimensions*, Comm. Pure Appl. Math., 71 (2018), pp. 892–937, https://doi.org/10.1002/cpa.21722.
[57] X. Mao, *Stochastic Differential Equations and Applications*, Elsevier, New York, 2007.
[58] V. A. Morozov, *On the solution of functional equations by the method of regularization*, Dokl. Akad. Nauk SSSR, 167 (1966), pp. 510–512.
[59] F. Parzer and O. Scherzer, *On Convergence Rates of Adaptive Ensemble Kalman Inversion for Linear Ill-Posed Problems*, preprint, arXiv:2104.10895 [math.NA], 2021, https://arxiv.org/abs/2104.10895.
[60] S. Reich, *A dynamical systems framework for intermittent data assimilation*, BIT, 51 (2011), pp. 235–249, https://doi.org/10.1007/s10543-010-0302-4.
[61] S. Reich and S. Weissmann, *Fokker-Planck particle systems for Bayesian inference: Computational approaches*, SIAM/ASA J. Uncertain. Quantif., 9 (2021), pp. 446–482, https://doi.org/10.1137/19M1303162.
[62] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268, https://doi.org/10.1016/0167-2789(92)90242-F.
[63] C. Schillings and A. M. Stuart, *Analysis of the ensemble Kalman filter for inverse problems*, SIAM J. Numer. Anal., 55 (2017), pp. 1264–1290, https://doi.org/10.1137/16M105959X.
[64] C. Schillings and A. M. Stuart, *Convergence analysis of ensemble Kalman inversion: The linear, noisy case*, Appl. Anal., 97 (2018), pp. 107–123, https://doi.org/10.1080/00036811.2017.1386784.
[65] X. Tong, A. Majda, and D. Kelly, *Nonlinear stability of the ensemble Kalman filter with adaptive covariance inflation*, Commun. Math. Sci., 14 (2016), pp. 1283–1313, https://doi.org/10.4310/CMS.2016.v14.n5.a5.
[66] X. T. Tong, *Performance analysis of local ensemble Kalman filter*, J. Nonlinear Sci., 28 (2018), pp. 1397–1442, https://doi.org/10.1007/s00332-018-9453-2.