Convergence Analysis of the Ensemble Kalman Filter for Inverse Problems: the Noisy Case

C. Schillings[a*] and A.M. Stuart[b]

[a] *Institute for Mathematics, University of Mannheim, A5, 6, 68131 Mannheim, Germany*;
[b] *Department of Computing and Mathematical Sciences, California Institute of Technology, CA 91125, USA*

We present an analysis of the ensemble Kalman filter for inverse problems based on the continuous time limit of the algorithm. The analysis of the dynamical behaviour of the ensemble allows to establish well-posedness and convergence results for a fixed ensemble size. We will build on the results presented in [17] and generalise them to the case of noisy observational data, in particular the influence of the noise on the convergence will be investigated, both theoretically and numerically.

Keywords: Bayesian Inverse Problems, Ensemble Kalman Filter, Parameter Identification

AMS Subject Classifications: 65N21, 62F15, 65N75

1. Introduction

The treatment and minimisation of uncertainties in inverse problems is indispensable to address the impact of limited knowledge in the unknown and noise in observational data. In the inverse setting, the Bayesian approach allows to incorporate uncertainties in the model and the observations and leads to a complete characterisation of the uncertainty in terms of the posterior distribution, i.e. the conditional distribution of the unknown parameters given the data. The approach is well-defined in the infinite-dimensional setting, see [3, 18]. However, for computationally intensive applications, the computation or approximation of the posterior is prohibitively expensive and is even with today's supercomputers, still intractable. The Ensemble Kalman Filter (EnKF) introduced in the 1990s by Evensen and coworkers [6] is widely and successfully used by practitioners in such cases. The low computational costs, the straightforward implementation and its non-intrusive nature make the method appealing in various areas of application, see e.g. [1, 2]. But, on the downside, the EnKF is underpinned by very limited theoretical understanding. In the data assimilation context, well-posedness results of the EnKF can be found in [12, 13, 19, 20] and a convergence analysis in the case of a fully observed system is presented in [4]. The analysis of the large ensemble size limit can be found in [8, 14]. For inverse problems, the connection to deterministic regularisation techniques and step-size strategies for nonlinear forward problems can be found in [9–11]. In this paper, we will build on the results presented in [17] and generalise them to the case of noisy observational data. The analysis of the EnKF

*Corresponding author. Email: c.schillings@uni-mannheim.de

is based on the continuous time scaling limits, which allow to study the properties of the EnKF for fixed ensemble size.

The inverse problem is defined as follows: Let $\mathcal{X}$ and $\mathcal{Y}$ denote separable Hilbert spaces. Furthermore, we denote by $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ the forward response operator mapping from the parameter space $\mathcal{X}$ to the data space $\mathcal{Y}$. The observations are assumed to be finite-dimensional, i.e. $\mathcal{Y} = \mathrm{R}^K$, $K \in \mathrm{R}$. The goal of computation is to recover the unknown parameters u from noisy observations y, where

$$y = Au + \eta.$$

The noise η in the observations is assumed to be normally distributed with $\eta \sim \mathcal{N}(0, \Gamma)$, $\Gamma \in \mathrm{R}^{K \times K}$ symmetric positive definite. In the Bayesian setting, the unknown parameter u is interpreted as a random variable or random field, distributed according to our prior knowledge μ_0. We assume that the noise η is independent of u. It is well known that the EnKF does not provide an approximation of the posterior measure in the nonlinear case, cp. e.g. [5]. We therefore present an analysis of the EnKF viewed as a minimisation method of the least-squares functional

$$\Phi(u; y) = \frac{1}{2} \|\Gamma^{-\frac{1}{2}}(y - A(u))\|_\Gamma^2 .$$

We restrict the analysis to the linear case, which allows us to completely understand the error propagation of the method for a fixed ensemble size. The results build the basis for the analysis of the nonlinear case, which will be subject of future work.

The paper is structured as follows: In Section 2, we introduce the EnKF for inverse problems and derive the continuous time limit of the algorithm. We study the properties of the method by analysing the dynamical behaviour of the ensemble and derive convergence results by considering the long-time behaviour. We present in Section 3 well-posedness results, the quantification of the ensemble collapse and convergence results for the noisy, linear setting. Numerical experiments illustrating the findings are presented in Section 4.

2. The Ensemble Kalman Filter for Inverse Problems and its Continuous time limit

The EnKF as originally introduced by Evensen and coworkers is a method for the filtering problem, i.e. it sequentially updates the probability distribution of the model state given the data, see e.g. [7, 15, 16] for more details. We follow the ideas in [10] and generalise the EnKF to inverse problems as follows. We introduce an artificial time $h = 1/N$ and use the observational data to update the unknown parameter at each time step nh with $n \in \mathrm{N}, n \leq N$.

We account for the repeated use of the observational data by amplifying the noise covariance by $1/h$, which leads to a consistent scaling of the misfit functional.

The EnKF then generates an ensemble of interacting particles, initialised by draws from the prior distribution, in the following way. The empirical mean $(\bar{u}_n)_{n=0}^N$ and covariance $(C(u_n))_{n=0}^N$ of the ensemble are updated in each iteration according to

$$\bar{u}_{n+1} = \bar{u}_n + K_n(y - A\bar{u}_n) \qquad C(u_{n+1}) = C(u_n) - K_n A C(u_n)\,, \tag{1}$$

where $K_n = C(u_n)A^*(AC(u_n)A^* + \frac{1}{h}\Gamma)^{-1}$ denotes the Kalman gain and $\bar{u}_n = \frac{1}{J}\sum_{j=1}^J u_n^{(j)}$, $C(u_n) = \frac{1}{J}\sum_{j=1}^J \big(u_n^{(j)} - \overline{u}_n\big) \otimes \big(u_n^{(j)} - \overline{u}_n\big)$. The iteration (1) does not define a unique linear transformation of each particle. We refer to [16] for more details on different transformations. Here, we will focus on the non-deterministic coupling, the EnKF with perturbed observations, leading to the iteration

$$u_{n+1}^{(j)} = u_n^{(j)} + C(u_n)A^*(AC(u_n)A^* + \frac{1}{h}\Gamma)^{-1}(y_{n+1}^{(j)} - Au_n^{(j)})$$

with $y_{n+1}^{(j)} = y + \eta_{n+1}^{(j)}$, $\eta_{n+1}^{(j)} \sim N(0, \frac{1}{h}\Gamma)$ for J particles, i.e. $j = 1, \dots, J$.

The analysis we present here relies on the continuous time limit of the EnKF. We therefore interpret the iterates $u_n^{(j)}$ as a discretization of a continuous function $u^{(j)}(nh)$. The limit is then formally given by

$$\mathrm{d}u^{(j)} = C(u)A^*\Gamma^{-1}A(u^\dagger + \eta - u^{(j)})\mathrm{d}t + C(u)A^*\Gamma^{-\frac{1}{2}}\mathrm{d}W^{(j)}\,,$$

where $W^{(1)}, \dots, W^{(J)}$ are pairwise independent cylindrical Wiener processes on $\mathcal{X}$, y denotes the noisy observational data and $C(u) = \frac{1}{J}\sum_{k=1}^J \big(u^{(k)} - \overline{u}\big) \otimes \big(u^{(k)} - \overline{u}\big)$. We will restrict the analysis to the case $\eta_{n+1}^{(j)} = 0$ for all $n \in \mathrm{N}$ and $j = 1, \dots, J$ leading to a limiting ordinary differential equation of the form

$$\frac{\mathrm{d}u^{(j)}}{\mathrm{d}t} = \frac{1}{J}\sum_{k=1}^J \big\langle A(u^{(k)} - \overline{u}), y - Au^{(j)}\big\rangle_\Gamma \big(u^{(k)} - \overline{u}\big), \quad j = 1, \cdots, J. \tag{2}$$

or equivalently

$$\frac{\mathrm{d}u^{(j)}}{\mathrm{d}t} = -C(u)D_u\Phi(u^{(j)}; y)$$

with potential $\Phi(u; y) = \frac{1}{2}\|\Gamma^{-\frac{1}{2}}(y - Au)\|_\Gamma^2$. Equation (2) reveals the well-known subspace property of the EnKF [10], since the vector field is in the linear span of the ensemble itself.

3. Convergence Analysis

In this section we generalise the results from [17] and allow for noise in the observational data; specifically we consider the case that the observational data $y^\dagger$ is polluted by additive noise $\eta^\dagger \in \mathrm{R}^K$ in the following way

$$y^\dagger = Au^\dagger + \eta^\dagger\,.$$

In subsection 3.1 we will demonstrate the undesirable effect of noise on the inversion methodology, and in subsection 3.2 we will suggest a stopping criterion to ameliorate the effect.

3.1. *Analysis of the EnKF With Noisy Data*

Following the notation introduced in [17], we inroduce the quantities

$$e^{(j)} = u^{(j)} - \bar{u}, \quad r^{(j)} = u^{(j)} - u^\dagger \quad j = 1, \ldots, J$$

$$E_{lj} = \langle Ae^{(l)}, Ae^{(j)} \rangle_\Gamma, \quad R_{lj} = \langle Ar^{(l)}, Ar^{(j)} \rangle_\Gamma, \quad F_{lj} = \langle Ar^{(l)}, Ae^{(j)} \rangle_\Gamma \quad l, j = 1, \ldots, J\,,$$

and the misfit $\vartheta^{(j)} = Au^{(j)} - y^\dagger, \mathrm{J} = 1, \ldots, J$. The quantity $e^{(j)}$ measures, for each particle j, the difference to the empirical mean (computed from the ensemble) and the quantity $r^{(j)}$ measures the difference from particle j to the truth. The matrix-valued quantities describe the interaction of these quantities mapped to the observation space. Due to the linearity of the forward operator, the misfit can be related to the residuals as follows

$$\vartheta^{(j)} = Ar^{(j)} - \eta^\dagger \quad j = 1, \ldots, J\,,$$

i.e. the misfit is a finite dimensional quantity in R^K. Furthermore, we define the matrix-valued quantity D by

$$D_{lj} = \langle \vartheta^{(l)}, Ae^{(j)} \rangle_\Gamma \qquad l, j = 1, \ldots, J\,.$$

THEOREM 3.1 *Let $y^\dagger$ denote the perturbed image of a truth $u^\dagger \in X$, i.e. $y^\dagger = Au^\dagger + \eta^\dagger$. Furthermore, an initial ensemble $u^{(j)}(0) \in X$ for $j = 1, \ldots, J$ is given, and we denote by X_0 the linear span of the $\{u^{(j)}(0)\}_{j=1}^J$. Then, equation (2) has a unique solution $u^{(j)}(\cdot) \in C([0,T); X_0)$ for $j = 1, \ldots, J$.*

Proof. Each particle $u^{(j)}$ satisfies

$$\frac{\mathrm{d}u^{(j)}}{\mathrm{d}t} = -\frac{1}{J}\sum_{k=1}^J F_{jk} e^{(k)} + \frac{1}{J}\sum_{k=1}^J \langle Ae^{(k)}, \eta^\dagger \rangle e^{(k)} \tag{3}$$

$$= -\frac{1}{J}\sum_{k=1}^J D_{jk} u^{(k)}\,. \tag{4}$$

The subspace property of the EnKF and the local Lipschitz continuity of the right-hand side of (4) ensures the local existence of a solution in $C([0,T); X_0)$ for $T > 0$. To establish global existence of solutions, we show the boundedness of the right-hand side of (4).

The following differential equation holds for the quantity $e^{(j)}$

$$\frac{\mathrm{d}e^{(j)}}{\mathrm{d}t} = -\frac{1}{J}\sum_{k=1}^J E_{jk} e^{(k)} = -\frac{1}{J}\sum_{k=1}^J E_{jk} r^{(k)}\,.$$

For the matrix-valued quantity E, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}E = -\frac{2}{J}E^2\,.$$

Thus, the dynamical behaviour of the quantities $e^{(j)}$ and $Ae^{(j)}$ is not influenced by the noise in the data. Therefore, the results presented in [17] for the noise free

case still hold true, i.e. for the orthogonal matrix X defined through the eigendecomposition of $E(0)$ it holds true that

$$E(t) = X\Lambda(t)X^\top \tag{5}$$

with $\Lambda(t) = \operatorname{diag}\{\lambda^{(1)}(t), \dots, \lambda^{(J)}(t)\}$, $\Lambda(0) = \operatorname{diag}\{\lambda_0^{(1)}, \dots, \lambda_0^{(J)}\}$ and

$$\lambda^{(j)}(t) = \Big(\frac{2}{J}t + \frac{1}{\lambda_0^{(j)}}\Big)^{-1}, \tag{6}$$

if $\lambda_0^{(j)} \neq 0$, otherwise $\lambda^{(j)}(t) = 0$.

The misfit $\vartheta^{(j)}$ satisfies

$$\frac{\mathrm{d}\vartheta^{(j)}}{\mathrm{d}t} = -\frac{1}{J}\sum_{k=1}^{J} D_{jk} A e^{(k)}$$

and the dynamical behaviour of the corresponding matrix-valued quantity D is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}D = -\frac{2}{J}DE\,.$$

The boundedness of $D(t)$ follows from the boundedness of the misfit $\vartheta^{(j)}$, which can be derived from

$$\frac{1}{2}\frac{\mathrm{d}\vartheta^{(j)}}{\mathrm{d}t} = -\frac{1}{J}\sum_{k=1}^{J} D_{jk}D_{jk}\,.$$

Hence, the misfit $\vartheta^{(j)}$ is bounded uniformly in time. By the Cauchy-Schwartz inequality, the bound on D follows with

$$D_{ij}^2 = \langle \vartheta^{(i)}, Ae^{(j)}\rangle_\Gamma^2 \leq |\vartheta^{(i)}|_\Gamma^2 \cdot |Ae^{(j)}|_\Gamma^2 \leq C|Ae^{(j)}|_\Gamma^2$$

for a constant $C > 0$ independent of T, i.e. $D_{ij} \to 0$ at least as fast as $\frac{1}{\sqrt{t}}$ as $t \to \infty$, in particular, D is uniformly bounded in time. Global existence for $u^{(j)}$ (and $e^{(j)}$, $r^{(j)}$) follows. ∎

The proof of Theorem 3.1 reveals that the behaviour of the quantity $e^{(j)}$, which is an indicator of the ensemble collapse, is not affected by the noise. Hence, [17, Theorem 3] can be directly generalised to the perturbed case.

COROLLARY 3.2 *Let $y^\dagger$ denote the perturbed image of a truth $u^\dagger \in X$, i.e. $y^\dagger = Au^\dagger + \eta^\dagger$. Furthermore, assume that an initial ensemble $u^{(j)}(0) \in X$ for $j = 1, \dots, J$ is given. Then, the matrix valued quantity $E(t)$ converges to 0 for $t \to \infty$ with an algebraic rate of convergence: $\|E(t)\| = \mathcal{O}(Jt^{-1})$.*

The ensemble collapse is a further form of regularisation as the solution not only remains in the linear span of the initial ensemble, but actually asymptotically lives in the span of a single element. The preceding result shows that the ensemble collapse, namely the fact that all particles converge to their common mean, does not depend on the realisation of the noise. We now discuss the convergence properties of

the EnKF in the perturbed case. The analysis presented in [17, Theorem 4] indicates that we can transfer the convergence result straightforwardly to the mismatch $\vartheta^{(j)}$. However, the convergence of the residuals $r^{(j)}$ depends on the realisation of the noise.

THEOREM 3.3 *Let $y^\dagger$ denote the perturbed image of a truth $u^\dagger \in X$, i.e. $y^\dagger = Au^\dagger + \eta^\dagger$ and the forward operator A is one-to-one. Let $\mathcal{Y}^\parallel$ denote the linear span of the $\{Ae^{(j)}(0)\}_{j=1}^J$ and let $\mathcal{Y}^\perp$ denote the orthogonal complement of $\mathcal{Y}^\parallel$ in Y and assume that the initial ensemble members are chosen so that $\mathcal{Y}^\parallel$ has the maximal dimension $\min\{J-1, \dim(Y)\}$. Then $\vartheta^{(j)}(t)$ may be decomposed uniquely as $\vartheta_\parallel^{(j)}(t) + \vartheta_\perp^{(j)}(t)$ with $\vartheta_\parallel^{(j)} \in \mathcal{Y}^\parallel$ and $\vartheta_\perp^{(j)} \in \mathcal{Y}^\perp$, where $\vartheta_\parallel^{(j)}(t) \to 0$ as $t \to \infty$ and $\vartheta_\perp^{(j)}(t) = \vartheta_\perp^{(j)}(0) = \vartheta_\perp^{(1)}$.*

Furthermore, if $\langle \eta^\dagger, Ae^{(k)} \rangle \leq \langle Ar^{(k)}, Ae^{(k)} \rangle$, the residual is monotonically decreasing. The rate of convergence of the component of the residual mapped forward to the observational space, which belongs to $\mathcal{Y}^\parallel$, can be arbitrarily slow, i.e. depending on the realisation of the noise, the rate of convergence can be arbitrarily close to 0.

Proof. The first part of the theorem follows with the same arguments as used for the proof of [17, Theorem 4].

The norm of the residuals satisfies the following differential equation

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|Ar^{(j)}\|_\Gamma^2 = -\frac{1}{J}\sum_{k=1}^J F_{jk}^2 + \frac{1}{J}\sum_{k=1}^J \langle Ar^{(k)}, Ae^{(k)} \rangle_\Gamma \langle \eta^\dagger, Ae^{(k)} \rangle_\Gamma \ .$$

Provided that $\langle \eta^\dagger, Ae^{(k)} \rangle \leq \langle Ar^{(k)}, Ae^{(k)} \rangle_\Gamma$ for $k = 1, \ldots, J$, i.e. $\|\eta^\dagger\|_\Gamma \cos(\theta_1) \leq \|Ar^{(j)}\|_\Gamma \cos(\theta_2)$ with θ_1 and θ_2 denoting the angle between $\eta^\dagger$ and $Ae^{(k)}$, and between $Ar^{(k)}$ and $Ae^{(k)}$, respectively, the residuals mapped to the image space of the forward operator are monotonically decreasing. Expanding the quantities $Ar^{(k)}$ and $\eta^\dagger$ in $\mathcal{Y}^\parallel$ and the orthogonal complement $\mathcal{Y}^\perp$

$$Ar^{(j)}(t) = \sum_{k=1}^J \alpha_k Ae^{(k)}(t) + Ar_\perp^{(1)}$$

$$\eta^\dagger = \sum_{k=1}^J \eta_k Ae^{(k)}(t) + A\eta_\perp^{(1)} \ ,$$

cp. [17, Lemma 8] yields

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|Ar^{(j)}\|_\Gamma^2 = -\frac{1}{J}\sum_{k=1}^J\sum_{l=1}^J E_{lk}\alpha_k E_{kl}\alpha_l + \frac{1}{J}\sum_{k=1}^J\sum_{l=1}^J E_{lk}\alpha_k E_{kl}\eta_l \ .$$

If the coefficients of the noise are of the size of α_k, the right hand side becomes 0 and the claim follows. ∎

Note that the proof of Theorem 3.3 suggests an a posteriori stopping criterion for the EnKF. We will discuss in the following section the discrepancy principle as a suitable criterion. Furthermore, assume that the noise is orthogonal to the space

spanned by the linear ensemble, then the theorem shows the convergence of the residuals in the image space.

3.2. *Stopping Criteria for the EnKF*

The Bayesian approach suggests an integration of the limit (2) up to time $T = 1$. This can be interpreted as an a priori regularisation strategy motivated by the probabilistic viewpoint. However, this stopping rule does not take into account the actual realisation of the noise nor the additional regularisation effect due to the ensemble collapse. The numerical experiments will show that this strategy often leads to an too early stopping of the algorithm.

The proof of Theorem 3.3 suggests an a posteriori stopping criterion for the method. We will investigate the discrepancy principle as a suitable criterion. Furthermore, we note that if the noise is orthogonal to the space spanned by the linear ensemble, then Theorem 3.3 shows the convergence of the residuals in the image space.

Motivated by the deterministic regularisation methods, we consider the discrepancy principle. Based on the noise model, we can choose a noise level $\delta > 0$ with $\|\eta^\dagger\| = \|y^\dagger - Au^\dagger\| < \delta$ (with given probability P_0). Note that the noise in the observations is assumed to be normally distributed, i.e. realisations of the noise cannot be bounded from above and below. Therefore, realisations of the noise are smaller than δ with probability $P_0 = P(\|\eta\| \leq \delta)$. The idea of this stopping rule is that, due to noisy data, the information in the observations cannot be distinguished from the noise for a residual in the order of δ, which means, asking for a residual with discrepancy smaller than δ leads to fitting of the unknown parameters to the noise. Therefore, the iterations of the EnKF will be stopped, when

$$\|A\bar{u}(t) - y^\dagger\| \leq \tau\delta$$

with an appropriately chosen parameter $\tau > 1$ and $\bar{u}(t)$ denotes the empirical mean of the ensemble at artificial time t.

In the numerical experiments, we will observe that a suitable stopping criterion is indispensable in order to avoid the overfitting effect to the noisy data and thus, to compute reliable and stable estimates of the unknown parameters.

4. Numerical Experiments

The forward model is described by the one dimensional elliptic equation

$$-\frac{\mathrm{d}^2 p}{\mathrm{d}x^2} + p = u \quad \text{in } D := (0, \pi)\,, \; p = 0 \quad \text{in } \partial D\,.$$

The solution operator of the model is a mapping $G : L^2 \to H^2(D) \cap H_0^1(D)$. The solution is observed at $K = 2^4 - 1$ equispaced observation points at $x_k = \frac{k}{2^4}, k = 1, \ldots, 2^4 - 1$, which defines the observation operator $O : H^2(I) \cap H_0^1 \to \mathrm{R}^k$, i.e. the operator A is a mapping from L^2 to R^K defined by the composition of the solution operator and the observation operator. We use a finite element method with continuous, piecewise linear ansatz functions on a uniform mesh with meshwidth $h = 2^{-8}$ to solve the forward problem (the spatial discretization leads to a discretization of u, i.e. $u \in \mathrm{R}^{2^8-1}$).

Then, the inverse problem consists of recovering the unknown data u from noisy observations

$$y^\dagger = p + \eta = Au^\dagger + \eta \,. \tag{7}$$

The measurement noise is chosen to be normally distributed, $\eta \sim \mathcal{N}(0, \gamma I)$, $\gamma = 0.01^2 \in \mathrm{R}$, $I \in \mathrm{R}^{K\times K}$. Furthermore, the prior is $\mu_0 = N(0, C_0)$ with covariance operator $C_0 = 10(-\Delta)^{-1}$. Here, we consider the Laplacian Δ with domain $H^2(D) \cap H_0^1(D)$. The initial ensemble is based on the eigendecomposition of the covariance operator C_0, i.e. $u^{(j)}(0) = \sqrt{\lambda_j}\zeta_j z_j$ with $\zeta_j \sim \mathcal{N}(0,1)$ for $j = 1, \ldots, J$ and $\{\lambda_j, z_j\}_{j\in\mathrm{N}}$ denoting eigenvalues and eigenfunctions of C_0.

To illustrate and numerically verify the results presented in this paper, we investigate the dynamical behaviour of the quantities e, r and the misfit ϑ. The theoretical results presented hold true for each particle, we therefore consider in the following a rather small ensemble of size $J = 5$. For the sake of presentation, the empirical mean (and minimum and maximum deviations) of the ensemble is shown.

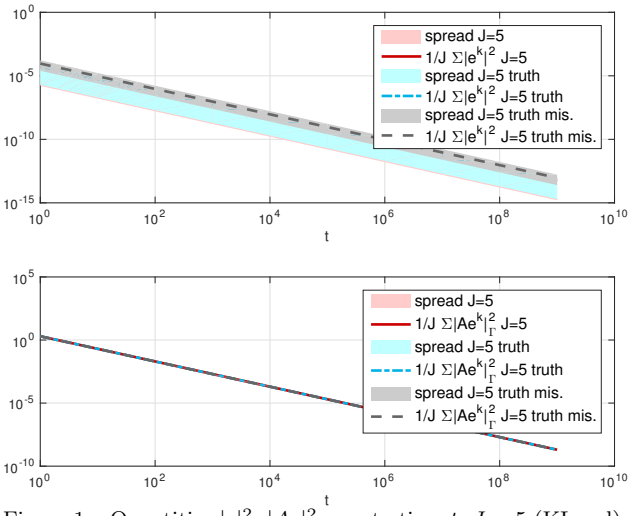

Figure 1. Quantities $|e|_2^2$, $|Ae|_\Gamma^2$ w.r. to time t, $J = 5$ (KL red), ($u^\dagger$ adaptive blue), ($\tilde{u}$ adaptive gray), $\beta = 10$, $K = 2^4 - 1$, initial ensemble chosen based on KL expansion of $C_0 = \beta(-\Delta)^{-1}$, $\eta \sim \mathcal{N}(0, 0.01^2\,\mathrm{id})$.

To investigate the convergence results further, we compare the performance of three ensembles (all of size $J = 5$): the first one (shown in red) is based on the first five terms in the Karhunen-Loève (KL) expansion of the covariance operator C_0, the second one (shown in blue) is chosen such that the contribution of $Ar_\perp(t)$ in Theorem 3.3 is minimised (i.e. $Ar^{(1)} = \sum_{k=1}^J \alpha_k Ae^{(k)}$ for some coefficients $\alpha_k \in \mathrm{R}$. Given $u^{(2)}, \ldots, u^{(J)}$ and coefficients $\alpha_1, \ldots, \alpha_J$, we define $u^{(1)} = (1 - \alpha_1 + \sum_{k=1}^J \alpha_k/J)^{-1}(u^\dagger - \alpha_1/J\sum_{j=2}^J u^{(j)} + \sum_{k=2}^J \alpha_k u^{(k)} - \alpha_k/J\sum_{j=2}^J u^{(j)}))$, the third ensemble (shown in grey) is chosen such that the contribution of $\vartheta(t)_\perp$ in Theorem 3.3 is minimised (i.e. $\vartheta^{(1)} = \sum_{k=1}^J \alpha_k Ae^{(k)}$ for some coefficients $\alpha_k \in \mathrm{R}$. Given $u^{(2)}, \ldots, u^{(J)}$ and coefficients $\alpha_1, \ldots, \alpha_J$, we define $u^{(1)} = (1 - \alpha_1 + \sum_{k=1}^J \alpha_k/J)^{-1}(\tilde{u} - \alpha_1/J\sum_{j=2}^J u^{(j)} + \sum_{k=2}^J \alpha_k u^{(k)} - \alpha_k/J\sum_{j=2}^J u^{(j)})$, where $\tilde{u}$ is the minimiser of the underdetermined least-squares problem).

In practice, the second strategy is not implementable, since the truth is used to construct the ensemble. However, the performance of the second strategy gives useful insight into the convergence behaviour of the EnKF.

The ensemble collapse is not affected by the choice of the initial ensemble. We

observe the predicted algebraic rate of convergence to the empirical mean, cp Figure 1.

The convergence behaviour of the residuals and the misfit, both projected to the subspace spanned by the initial ensemble and the complement are shown in the following two figures 2 and 3.

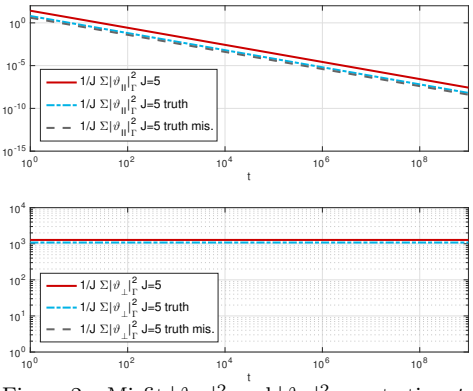

Figure 2. Misfit $|\vartheta_{II}|^2_\Gamma$ and $|\vartheta_\perp|^2_\Gamma$ w.r. to time t, $J = 5$ (KL red), ($u^\dagger$ adaptive blue), ($\tilde{u}$ adaptive gray), $\beta = 10$, $K = 2^4 - 1$, initial ensemble chosen based on KL expansion of $C_0 = \beta(-\Delta)^{-1}$, $\eta \sim \mathcal{N}(0, 0.01^2\,\mathrm{id})$.

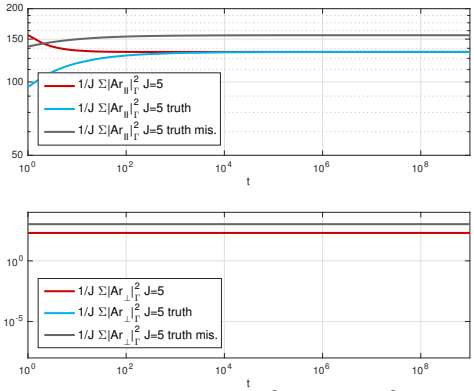

Figure 3. Residuals $|Ar_{II}|^2_\Gamma$ and $|Ar_\perp|^2_\Gamma$ w.r. to time t, $J = 5$ (KL red), ($u^\dagger$ adaptive blue), ($\tilde{u}$ adaptive gray), $\beta = 10$, $K = 2^4 - 1$, initial ensemble chosen based on KL expansion of $C_0 = \beta(-\Delta)^{-1}$, $\eta \sim \mathcal{N}(0, 0.01^2\,\mathrm{id})$.

The algebraic rate of the misfit is clearly confirmed. Furthermore, the convergence behaviour of the residuals for the KL based ensemble (shown in red in Figure 2) illustrates the arbitrarily slow convergence predicted by the theory, i.e. we observe a convergence rate deteriorating to 0. For the other two ensembles, we even observe an increase in the residual, since the angle conditions are not satisfied. The comparison of the resulting estimates with the truth reveals the strong overfitting effect of the third ensemble, cp Figure 4. This behaviour is expected due to the construction of the ensemble, which implies an amplification of the noise in the data.

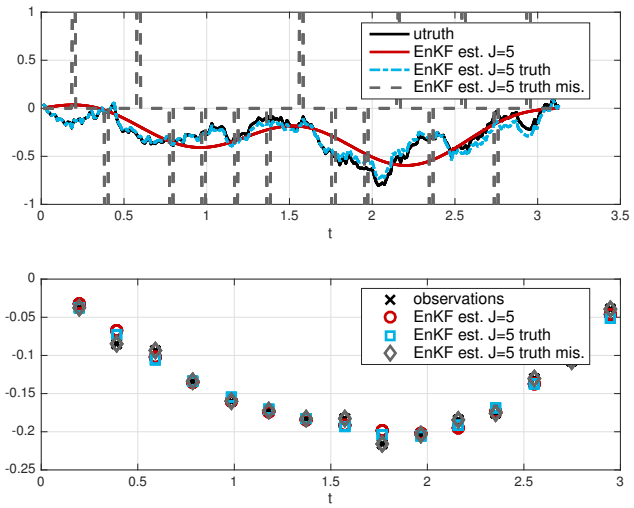

Figure 4. Comparison of the EnKF estimate with the truth and the observations, $J = 5$ (KL red), ($u^\dagger$ adaptive blue), ($\tilde{u}$ adaptive gray), $\beta = 10$, $K = 2^4 - 1$, initial ensemble chosen based on KL expansion of $C_0 = \beta(-\Delta)^{-1}$, $\eta \sim \mathcal{N}(0, 0.01^2\,\mathrm{id})$.

To illustrate the effect of the angle condition and the resulting degradation of the convergence order of the residuals, we repeat the experiments with noise in the data, which is orthogonal to the subspace spanned by the initial ensemble. The theoretical results suggest an algebraic rate of convergence, which can be confirmed by the results presented in Figure 5.

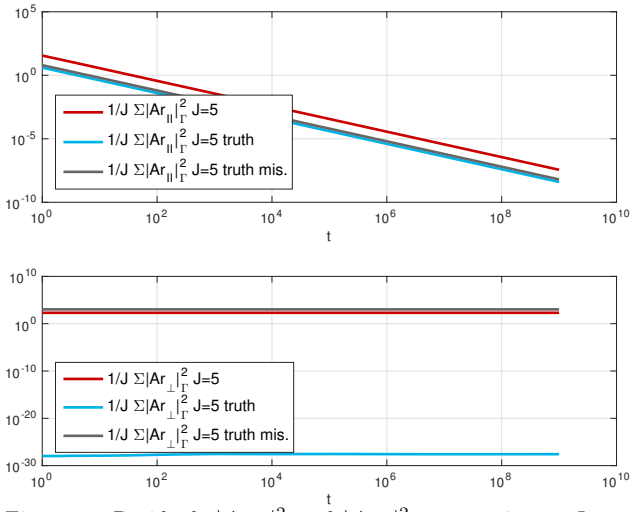

Figure 5. Residuals $|Ar_{II}|^2_\Gamma$ and $|Ar_\perp|^2_\Gamma$ w.r. to time t, $J = 5$ (KL red), ($u^\dagger$ adaptive blue), ($\tilde{u}$ adaptive gray), $\beta = 10$, $K = 2^4 - 1$, initial ensemble chosen based on KL expansion of $C_0 = \beta(-\Delta)^{-1}$, $\eta \sim \mathcal{N}(0, 0.01^2\,\mathrm{id})$, observational noise orthogonal to the subspace spanned by the initial ensembles.

The result on the ensemble collapse Corollary 3.2 indicates that the regularisation effect of the method strongly depends on the number of particles in the ensemble. The Bayesian stopping rule, which can be interpreted as an a priori stopping rule, does not reflect this behaviour. We will show in the following that the discrepancy principle leads to suitable stopping strategy, in particular, it has the potential to substantially improve the accuracy of the EnKF estimate. To do so, we repeat the experiments with 10 randomly chosen ensembles (based on the KL expansion of the prior covariance operator) of size $J = 5$ and $J = 50$. The noise in the data is randomly chosen from $\mathcal{N}(0, \gamma I)$ with $\gamma = 0.01^2 \in \mathrm{R}^2$. Motivated by the previous discussion on the discrepancy principle, we implement a stopping rule of the form $\|A\bar{u}(t) - y^\dagger\|_\Gamma \leq 1.2K$, where K denotes the number of observations. Figures 6 and 8 show the comparison of the estimates based on the discrepancy principle with the truth.

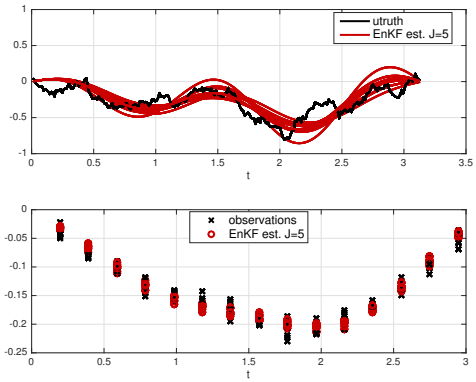

Figure 6. Comparison of the EnKF estimate with the truth and the observations with discrepancy stopping rule, $J = 5$ based on KL expansion of $C_0 = \beta(-\Delta)^{-1}$ (red), $\beta = 10$, $K = 2^4 - 1$, 10 randomly initialised ensembles, 10 randomly perturbed observation.

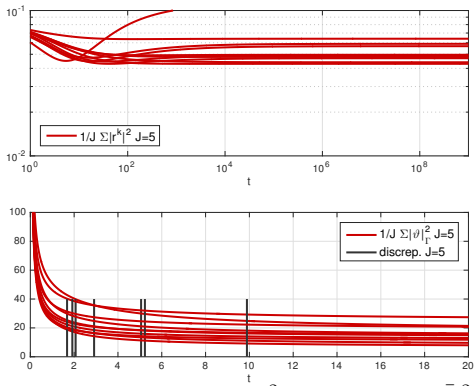

Figure 7. Residuals $|A\bar{r}|^2_\Gamma$ (above) and $|\bar{\vartheta}|^2_\Gamma$ with discrepancy stopping rule (below) w.r. to time t, $J = 5$ based on KL expansion of $C_0 = \beta(-\Delta)^{-1}$ (red), $\beta = 10$, $K = 2^4 - 1$, 10 randomly initialised ensembles, 10 randomly perturbed observation.

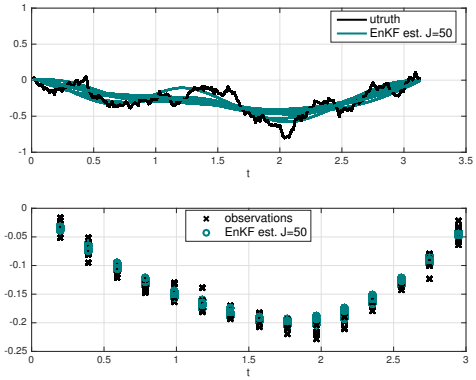

Figure 8. Comparison of the EnKF estimate (with stopping rule) with the truth and the observations with discrepancy stopping rule, $J = 50$ based on KL expansion of $C_0 = \beta(-\Delta)^{-1}$ (blue), $\beta = 10$, $K = 2^4 - 1$, 10 randomly initialised ensembles, 10 randomly perturbed observation.

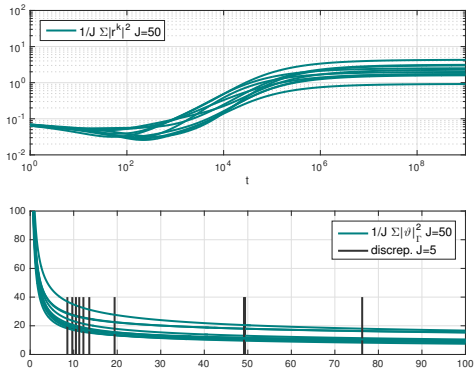

Figure 9. Residuals $|A\bar{r}|^2_\Gamma$ (above) and $|\bar{\vartheta}|^2_\Gamma$ with discrepancy stopping rule (below) w.r. to time t, $J = 50$ based on KL expansion of $C_0 = \beta(-\Delta)^{-1}$ (blue), $\beta = 10$, $K = 2^4 - 1$, 10 randomly initialised ensembles, 10 randomly perturbed observation.

We observe that the overfitting effect is much more pronounced for the larger ensemble of size 50, cp. the empirical residuals in Figure 7 and Figure 9. The KL expansion of the first 50 terms includes more fine-scale (oscillatory) details, which can be fitted to the noise in the observational data and therefore cause the overfitting effect. The smaller ensemble based on the first 5 terms of the KL expansion avoids the overfitting effect due to the smaller ensemble size leading to a faster ensemble collapse, but also due to the smoothness of the first KL terms, i.e. the subspace property preserves the smoothness of the KL terms. Furthermore, we note that the discrepancy principle leads in all experiments to a stopping time larger than 1 (Bayesian stopping rule), which leads for all experiments to a further improvement in the EnKF estimate. Due to the delayed ensemble collapse, the stopping times for the larger ensemble are on average greater than the ones for the smaller ensemble. The experiments suggest that an a posteriori stopping rule can significantly improve the performance of the EnKF. This observation is consistent with previous works on stopping rules for the EnKF, cp. [9].

5. Conclusions

The presented analysis of the ensemble Kalman filter for inverse problems shows that the well-posedness results and the quantification of the ensemble collapse derived in [17] can be straightforwardly generalised to the noisy case. However, the convergence behaviour of the ensemble is strongly affected by the noise in the observational data, no convergence rate of the residuals can be proven, i.e. the convergence rate can be arbitrarily slow. The numerical experiments confirm the theory. In addition, the numerical experiments demonstrate the importance of an appropriate stopping rule in the presence of noise in order to avoid the well-known overfitting effect. It is also shown that the ensemble itself has a regularisation effect, caused by the ensemble collapse as well as by the chosen initialisation of the ensemble in terms of the KL expansion. Even though the presented results are confined to the linear case, they provide useful insights into the performance of the filter in the presence of noise and can also enhance our understanding of the nonlinear case.

Acknowledgments Both authors are grateful to the EPSRC Programme Grant EQUIP for funding of this research. AMS is also grateful to DARPA and to ONR for funding parts of this research.

References

[1] Kay Bergemann and Sebastian Reich. A localization technique for ensemble Kalman filters. *Quarterly Journal of the Royal Meteorological Society*, 136(648):701–707, 2010.

[2] Kay Bergemann and Sebastian Reich. A mollified ensemble Kalman filter. *Quarterly Journal of the Royal Meteorological Society*, 136(651):1636–1643, 2010.

[3] M. Dashti and A.M. Stuart. The Bayesian approach to inverse problems. *arXiv preprint arXiv:1302.6989*, 2014.

[4] 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. *ArXiv e-prints*, December 2016.

[5] O.G. Ernst, B. Sprungk, and H. Starkloff. Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems. *arXiv preprint arXiv:1504.03529*, 2015.

[6] G. Evensen. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53(4):343–367, 2003.

[7] G. Evensen. *Data Assimilation: The Ensemble Kalman Filter*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[8] S. Gratton, J. Mandel, et al. On the convergence of a non-linear ensemble Kalman smoother. *arXiv preprint arXiv:1411.4608*, 2014.

[9] M.A. Iglesias. Iterative regularization for ensemble data assimilation in reservoir models. *Computational Geosciences*, pages 1–36, 2014.

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

[11] Marco A Iglesias. A regularizing iterative ensemble Kalman method for pde-constrained inverse problems. *arXiv preprint arXiv:1505.03876*, 2015.

[12] David Kelly, Andrew J. Majda, and Xin T. Tong. Concrete ensemble Kalman filters with rigorous catastrophic filter divergence. *Proceedings of the National Academy of Sciences*, 112(34):10589–10594, 2015.

[13] D.T.B. Kelly, K.J.H. Law, and A.M. Stuart. Well-posedness and accuracy of the

ensemble Kalman filter in discrete and continuous time. *Nonlinearity*, 27(10):2579, 2014.

[14] E. Kwiatkowski and J. Mandel. Convergence of the square root ensemble Kalman filter in the large ensemble limit. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):1–17, 2015.

[15] K.J.H. Law, A.M. Stuart, and K.C. Zygalakis. *Data Assimilation: A Mathematical Introduction*. Springer, 2015.

[16] S. Reich and C. Cotter. *Probabilistic Forecasting and Bayesian Data Assimilation*. Cambridge University Press, 2015.

[17] C. Schillings and Stuart A. Analysis of the Ensemble Kalman Filter for Inverse Problems. *SIAM Numerical Analysis (accepted)*, 2017.

[18] A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451559, May 2010.

[19] Xin T Tong, Andrew J Majda, and David Kelly. Nonlinear stability of the ensemble Kalman filter with adaptive covariance inflation. *arXiv:1507.08319*, 2015.

[20] Xin T Tong, Andrew J Majda, and David Kelly. Nonlinear stability and ergodicity of ensemble based Kalman filters. *Nonlinearity*, 29(2):657, 2016.