Nonlinear stability and ergodicity of ensemble based Kalman filters

Xin T Tong[1,2], **Andrew J Majda**[1,2] **and David Kelly**[1]

[1]Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA

[2]Center for Atmosphere Ocean Science, New York University, New York, NY 10012, USA

E-mail: `tong@cims.nyu.edu`

Abstract. The ensemble Kalman filter (EnKF) and ensemble square root filter (ESRF) are data assimilation methods used to combine high dimensional, nonlinear dynamical models with observed data. Despite their widespread usage in climate science and oil reservoir simulation, very little is known about the long-time behavior of these methods and why they are effective when applied with modest ensemble sizes in large dimensional turbulent dynamical systems. By following the basic principles of energy dissipation and controllability of filters, this paper establishes a simple, systematic and rigorous framework for the nonlinear analysis of EnKF and ESRF with arbitrary ensemble size, focusing on the dynamical properties of boundedness and geometric ergodicity. The time uniform boundedness guarantees that the filter estimate will not diverge to machine infinity in finite time, which is a potential threat for EnKF and ESQF known as the catastrophic filter divergence. Geometric ergodicity ensures in addition that the filter has a unique invariant measure and that initialization errors will dissipate exponentially in time. We establish these results by introducing a natural notion of observable energy dissipation. The time uniform bound is achieved through a simple Lyapunov function argument, this result applies to systems with complete observations and strong kinetic energy dissipation, but also to concrete examples with incomplete observations. With the Lyapunov function argument established, the geometric ergodicity is obtained by verifying the controllability of the filter processes; in particular, such analysis for ESQF relies on a careful multivariate perturbation analysis of the covariance eigen-structure.

1. Introduction

An important problem in scientific computing is the effective assimilation of observational data with high dimensional nonlinear forecast models. The classical filtering tools, such as the Kalman filter and extended Kalman filter, are poorly suited to these problems, due both to the nonlinearity of the models and the cost of computing covariance matrices for high dimensional state vectors. The ensemble Kalman filter (EnKF) and ensemble square root filter (ESRF) were designed to overcome these difficulties [1, 2, 3, 4]. The basic idea of these methods is to propagate an ensemble $\{V_n^{(1)}, \dots, V_n^{(K)}\}$ to describe the forecast distribution of the underlying system U_n, and then assimilating the new observation via a Kalman-type update using the ensemble mean and covariance. The state estimate remains useful even when the ensemble size is several orders of magnitude smaller than the state dimension, leading to a considerable benefit in computational cost. Due to their efficiency, EnKF and ESRF are broadly used, notably in ocean-atmosphere science [5, 4] and oil reservoir simulations [6].

Despite the ubiquitous application of EnKF and ESRF, little is known of their dynamical behavior beyond that provided by numerical experiments. Existing theoretical studies of EnKF and ESRF focus mainly on either error estimation of one single assimilation step [7, 8], or in the case of linear model dynamics, convergence to the classical Kalman filter as the ensemble size tends to infinity [9, 10]. The aim of this article is to address a more practical scenario, namely by looking at the long-time behavior of the ensemble when the ensemble size is fixed and where the underlying model is nonlinear. To be specific, we seek to address the following questions:

(i) Under what model conditions does the ensemble remain bounded on an infinite time horizon?

(ii) Are the filter processes ergodic and how quickly do they lose memory of initial conditions?

These two questions are of great practical importance. Boundedness of the ensemble prohibits the state estimate from diverging to infinity, thereby precluding the disastrous phenomena of catastrophic filter divergence [11, 12, 13, 4]. Ergodicity of the ensemble ensures that errors in the initialization of the filter will not affect the performance of the filter in the long run [14, 15, 16], geometric ergodicity further insures that the error will dissipate exponentially fast in time. To the best of our knowledge, the only article in a similar setting is [17], where the authors show well-posedness of EnKF with bounds which can grow exponentially in time and accuracy under variance inflation.

In the study of dynamical systems, boundedness can be demonstrated through the construction of a Lyapunov function $\mathcal{E}$, which is a positive function statisfying the dissipation criterion

$$\mathbb{E}_{n-1}\mathcal{E}(U_n) \leq (1-\beta)\mathcal{E}(U_{n-1}) + K \ . \tag{1.1}$$

Here $\mathbb{E}_{n-1}$ denotes the conditional expectation with respect to the information at time $n-1$, and $0 < \beta < 1$ is a constant. Using the discrete Grönwall inequality, we immediately find that

$$\mathbb{E}\mathcal{E}(U_n) \leq (1-\beta)^n \mathbb{E}\mathcal{E}(U_0) + K\beta^{-1},$$

which shows that, under expectation of the Lyapunov function $\mathcal{E}$, the state U_n is bounded uniformly in n. In geophysically relevant models, such as the Lorenz equations and Navier-Stokes equations, the corresponding $\mathcal{E}$ can be chosen as the kinetic energy $\mathcal{E}(\cdot) = |\cdot|^2$. In this scenario, the relation (1.1) is known as an energy principle.

A natural strategy for proving boundedness of EnKF and ESRF is to check whether the energy principles of the nonlinear system are inherited by the ensemble. In other words, if $\mathcal{E}$ is a Lyapunov function of the original system U_n, can we use $\mathcal{E}$ to construct a Lyapunov function for the ensemble processes $\{V_n^{(k)}\}_{k=1}^K$.

In Theorems 3.2 and 3.3 we will show that this construction is quite straight-forward, provided that the underlying model satisfies a so-called observable energy criterion. To be specific, suppose the model is observed linearly via $Z_n = HU_n + \zeta_n$, then the observable energy criterion states that the underlying model satisfies (1.1) with the choice $\mathcal{E}(u) = |Hu|^2$, that is

$$\mathbb{E}_{n-1}|HU_n|^2 \leq (1-\beta)|HU_{n-1}|^2 + K\ . \tag{1.2}$$

Under this assumption, it is shown in Theorems 3.2, 3.3 that the ensemble $\{V_n^{(k)}\}_{k=1}^K$ satisfies a related energy principle. Hence, if the model satisfies (1.2) with full rank H, then the ensemble $\{V_n^{(k)}\}$ must remain bounded on an infinite time horizon. When H is not of full rank, one still obtains an energy principle from (1.2), but can only conclude boundedness of the observable ensemble $\{HV^{(k)}\}_{k=1}^K$.

With a Lyapunov function established, the EnKF and ESRF are shown to be geometrically ergodic by Theorems 5.5, 5.6 and 5.8, assuming the nonlinear system is propagated with non-degenerate noise. The proofs are conceptually simple, as it suffices to check to the controllability of the filters, thanks to the classical work of [18, 19, 20]. The only technical challenge lies in the eigenvalue decomposition of matrices required in the assimilation step of ESRF. This can be resolved by a careful multivariate perturbation analysis of the underlying matrices.

With a short discussion in Section 4, we will demonstrate a few sufficient conditions that imply the observable energy criterion (1.2). When H is of full rank, the observable energy criterion holds provided that the dynamics have an energy principle with strong contraction parameter, depending on the condition number of H. When H is not of full rank, the observable energy criterion does not hold in general, but is verifiable in several concrete examples through direct calculation. This dichotomy of observational rank agrees with known numerical evidence, where the ensemble behaves stably when full rank observations are available, but can experience filter divergence when the observations are sparse [12, 4], or even reach machine infinity in finite time, which is known as catastrophic filter divergence [11, 4, 13]. Using the same philosophy in this paper, the authors have found a concrete dynamical system that satisfies the kinetic energy criterion but not the observable energy criterion, and whose EnKF ensemble (provably) experiences catastrophic filter divergence with large probability. The authors have also found a general adaptive covariance inflation scheme, which always insures that the ensemble remains bounded on an infinite time horizon, without hurting the accuracy of original filters. These results will be reported in two separate papers [21, 22].

The article is structured as follows. In Section 2.1 we formulate the EnKF and ESRF methods and also introduce the notion of Lyapunov functions and energy principles. In

Section 3 we establish a simple framework to verify energy principles for EnKF and ESRF using the observable energy. Section 4 discusses the applicability of this framework by studying a few sufficient conditions that guarantee the observable energy condition. In Section 5 we prove the geometric ergodicity of the filter processes assuming the stability results in Section 3 hold. In Section 6 we conclude this paper and discuss possible extensions.

2. Models setup and fundamental concepts

2.1. Model setup

In this paper, we assume the signal sequence $U_n \in \mathbb{R}^d$ is generated through a nonlinear mapping Ψ_h plus a mean zero noise ζ_n, and the observation is a linear one plus some mean zero noise in $\mathbb{R}^q$:

$$U_n = \Psi_h(U_{n-1}) + \zeta_n, \quad Z_n = HU_n + \xi_n. \tag{2.1}$$

Here $\{\xi_n\}$ is an i.i.d. noise sequence, and ζ_n is independent of $\zeta_1, \dots \zeta_{n-1}$ conditioned on the realization of U_{n-1}. In many cases, the model may be generated by the solution of a stochastic differential equation (SDE)

$$du_t = \psi(u_t)dt + \Sigma dW_t \tag{2.2}$$

by taking $U_n = u_{nh}$ for some fixed $h > 0$. A short discussion of this discrete time formulation is attached in Appendix B.

At this stage, we impose no restrictions on ζ_n except that it is mean zero with a conditional covariance depends on U_{n-1}:

$$\mathbb{E}(\zeta_n | U_{n-1}) = 0, \quad \mathbb{E}(\zeta_n \otimes \zeta_n | U_{n-1}) = R_h(U_{n-1}). \tag{2.3}$$

As for the observation part, we will assume in this paper that

$$\mathrm{rank}(H) = q \leq d, \quad \mathbb{E}(\xi_n | U_{n-1}) = 0, \quad \mathbb{E}(\xi_n \otimes \xi_n | U_{n-1}) = I_q.$$

The seemingly restrictive choice of observational noise covariance can be made without loss of generality. Indeed, any observational covariance can be reduced to the identity via a simple coordinate change on the filtering problem. To apply the results of this article to a filtering problem with non-trivial observational covariance, one must first apply the coordindate change and then check the assumptions in the new system of coordinates. Details are contained in the remark below.

Remark 2.1. *Suppose that ξ_n has a nonsingular covariance matrix Γ and $\Gamma^{-1/2}H$ has an SVD decomposition $\Gamma^{-1/2}H = \Phi\Lambda\Psi^T$, then we change the coordinate system and consider*

$$\widetilde{U}_n = \Psi^T U_n, \quad \tilde{\xi}_n = \Phi^T \Gamma^{-1/2} \xi_n, \quad \widetilde{Z}_n = \Phi^T \Gamma^{-1/2} Z_n = \Lambda \widetilde{U}_n + \tilde{\xi}_n. \tag{2.4}$$

Hence this change of coordinates also reduces the observation matrix to a diagonal matrix. If the observation dimension q is larger than the model dimension d, the last $d - q$ diagonal *entries of Λ are zero, so the last $d-q$ rows of $\widetilde{Z}_n$ are independent of the signal and useless for filtering purpose, which we can ignore and set $d=q$. Moreover $\tilde{\xi}_n$ will have covariance matrix I_q. Since all the transformations above are linear and bijective, filtering $\widetilde{U}_n$ with $\widetilde{Z}_n$ is equivalent to filtering U_n with Z_n. On the other hand, if the covariance Γ is singular, then certain linear subspace can be observed exactly, and may cause the filtering operation to be singular. We do not consider such pathological cases in this paper.*

2.2. Ensemble Kalman filter

In the standard Kalman filtering theory, the conditional distribution of the signal process U_n given the observation sequence $Z_1,\dots,Z_n$ is given by a Gaussian distribution. EnKF inherits this idea by using a group of ensembles $\{V_n^{(k)}\}_{k=1}^K$ to represent this Gaussian distribution, as the mean and covariance can be taken as the ensemble mean and covariance. The EnKF operates very much like a Kalman filter, except its forecast step requires a Monte Carlo simulation due to the nonlinearity of the system. In detail, the EnKF is an iteration of following two steps, with (for instance) $\widehat{V}_0^{(k)}$ being sampled from the equilibrium measure of U_n.

- Forecast step: from the posterior ensemble at time $n-1$, $\{V_{n-1}^{(k)}\}_{k=1}^K$, a forecast ensemble for time n is generated by
$$\widehat{V}_n^{(k)} = \Psi_h(V_{n-1}^{(k)}) + \zeta_n^{(k)},$$
where $\zeta_n^{(k)}$ are independent samples drawn from the same distribution as ζ_n. Then the prior distribution for time n is described by the ensemble mean and covariance:
$$\overline{\widehat{V}}_n := \frac{1}{K}\sum_{k=1}^K \widehat{V}_n^{(k)}, \quad \widehat{C}_n := \frac{1}{K-1}\sum_{k=1}^K (\widehat{V}_n^{(k)} - \overline{\widehat{V}}_n)\otimes(\widehat{V}_n^{(k)} - \overline{\widehat{V}}_n). \tag{2.5}$$
- Analysis step: upon receiving the new observation Z_n, random perturbations of it are generated by adding $\xi_n^{(k)}$:
$$Z_n^{(k)} = Z_n + \xi_n^{(k)},$$
where $\xi_n^{(k)}$ are independent samples drawn from the same distribution as ξ_n. Each ensemble member is then updated to
$$V_n^{(k)} = \widehat{V}_n^{(k)} - \widehat{C}_n H^T (I + H\widehat{C}_n H^T)^{-1}(H\widehat{V}_n^{(k)} - Z_n^{(k)})\ .$$

In summary, the EnKF is generated by the following dynamics

$$\begin{gathered}
V_n^{(k)} = \widehat{V}_n^{(k)} - \widehat{C}_n H^T (I + H\widehat{C}_n H^T)^{-1}(H\widehat{V}_n^{(k)} - Z_n^{(k)}),\\
\widehat{V}_n^{(k)} = \Psi_h(V_{n-1}^{(k)}) + \zeta_n^{(k)}, \quad \overline{\widehat{V}}_n := \frac{1}{K}\sum_{k=1}^K \widehat{V}_n^{(k)}, \quad Z_{n+1}^{(k)} = Z_n + \xi_n^{(k)},\\
\widehat{C}_n := \frac{1}{K-1}\sum_{k=1}^K (\widehat{V}_n^{(k)} - \overline{\widehat{V}}_n)\otimes(\widehat{V}_n^{(k)} - \overline{\widehat{V}}_n)\ .
\end{gathered} \tag{2.6}$$

From this formulation, it is clear that the augmented process $\{U_n, V_n^{(1)}, \dots, V_n^{(K)}\}$ is a Markov chain. In the following discussion, we will denote the natural filtration up to time n as $\mathcal{F}_n = \sigma\{U_m, V_m^{(1)}, \dots, V_m^{(K)}, m \le n\}$, and denote the conditional expectation with respect to $\mathcal{F}_n$ as $\mathbb{E}_n$.

2.3. Ensemble square root filters

One drawback of EnKF comes from its usage of artificial noise $\xi_n^{(k)}$, as this introduces unnecessary sampling errors, particularly when the ensemble size is small [23]. The motivation behind the artificial noise is to make the posterior ensemble covariance

$$C_n := \frac{1}{K-1}\sum_{k=1}^{K}(V_n^{(k)} - \overline{V}_n) \otimes (V_n^{(k)} - \overline{V}_n), \quad \overline{V}_n := \frac{1}{K}\sum_{k=1}^{K} V_n^{(k)},$$

satisfy the covariance update of the standard Kalman filter

$$C_n = \widehat{C}_n - \widehat{C}_n H^T (H^T \widehat{C}_n H + I)^{-1} H \widehat{C}_n, \tag{2.7}$$

when the left hand is averaged over $\xi_n^{(k)}$ [24, 25, 7] . ESRFs, including the ensemble transform Kalman filter (ETKF) and the ensemble adjustment Kalman filter (EAKF), aim to resolve this issue by manipulating the posterior spreads to ensure that (2.7) holds. Both ETKF and EAKF algorithms are described by the following update steps, with the only difference occurring in the assimilation step for the spread. As with EnKF, the initial ensemble $\{V_0^{(k)}\}_{k=1}^{K}$ is (for instance) sampled from the equilibrium distribution of U_n.

- Forecast step: identical to EnKF, the forecast ensembles at time n is generated from posterior ensembles at time $n-1$:

  $$\widehat{V}_n^{(k)} = \Psi_h(V_{n-1}^{(k)}) + \zeta_n^{(k)}.$$

  The forecast ensemble covariance $\widehat{C}_n$ is then computed using (2.5).
- Assimilation step for the mean: upon receiving the new observation Z_n, the posterior ensemble mean is updated through

  $$\overline{V}_n = \overline{\widehat{V}_n} - \widehat{C}_n H^T (I + H\widehat{C}_n H^T)^{-1}(H\overline{\widehat{V}}_n - Z_n), \quad \overline{\widehat{V}}_n = \frac{1}{K}\sum_{k=1}^{K}\widehat{V}_n^{(k)}. \tag{2.8}$$

- Assimilation step for the spread: The forecast ensemble spread is given by the $d \times K$ matrix

  $$\widehat{S}_n = [\widehat{V}_n^{(1)} - \overline{\widehat{V}}_n, \dots, \widehat{V}_n^{(K)} - \overline{\widehat{V}}_n] \,.$$

  To update the spread, first find a matrix $T_n \in \mathbb{R}^{d\times d}$ (for ETKF) or $A_n \in \mathbb{R}^{K\times K}$ (for EAKF) such that

  $$\frac{1}{K-1} T_n \widehat{S}_n \otimes T\widehat{S}_n = \frac{1}{K-1}\widehat{S}_n A_n \otimes \widehat{S}_n A_n = \widehat{C}_n - \widehat{C}_n H^T (H^T \widehat{C}_n H + I)^{-1} H \widehat{C}_n \,. \tag{2.9}$$

The posterior spread is updated to $S_n = T_n\widehat{S}_n$ (for ETKF) or $S_n = \widehat{S}_n A_n$ (EAKF), and the ensemble members are updated to

$$V_n^{(k)} = \overline{V}_n + S_n^{(k)} \ ,$$

where $S_n^{(k)}$ denotes the k-th column of the updated spread matrix S_n. By construction, the posterior covariance $C_n = (K-1)^{-1} S_n^T S_n$ satisfies (2.7).

At this stage it suffices to know that such A_n and T_n exist, their finer properties play no role in the discussion concerning stability. Their formulation will become important when we want to study ergodicity in Section 5, and a detailed formulation will be given there. Based on our description above, the augmented process $\{U_n, V_n^{(1)}, \dots, V_n^{(K)}\}$ is again a Markov chain. As above, we will denote the natural filtration up to time n as $\mathcal{F}_n = \sigma\{U_m, V_m^{(1)}, \dots, V_m^{(K)}, m \leq n\}$, and denote the conditional expectation with respect to $\mathcal{F}_n$ as $\mathbb{E}_n$.

2.4. Covariance inflation

When applying EnKF and ESRF, the forecast ensemble covariance $\widehat{C}_n$ often underestimates the uncertainty in the forecast model. An ad hoc solution is to use inflated or modified forecast ensemble covariance in the assimilation step. We will discuss three types of such methods in this paper:

- Additive inflation: replace $\widehat{C}_n$ with $\widehat{C}_n + \lambda I$ for a proper $\lambda > 0$;
- Uniform inflation: replace $\widehat{C}_n$ with $(1+\lambda)\widehat{C}_n$ for a proper $\lambda > 0$.

It should be noted that additive inflation is only used in EnKF and not in the square root filters since it is not clear how an additive inflation should be applied at the level of the matrix square root. There are other ad hoc ways of modifying ESRF methods with additive inflation, see page 147 of [4] for more details.

2.5. Energy principles and Lyapunov functions

Stability for nonlinear systems can be studied through energy principles. That is, certain types of energy are preserved or dissipated by the dynamics. In a stochastic setting, this idea is formalized using Lyapunov functions. In this paper, we say that $\mathcal{E}$ is a *Lyapunov function* for a Markov chain X_n if there exists positive constants $0 < \beta < 1$ and K such that

$$\mathbb{E}(\mathcal{E}(X_n)|X_{n-1}) \leq (1-\beta)\mathcal{E}(X_{n-1}) + K \tag{2.10}$$

for all $n \in \mathbb{Z}^+$. For a continuous time Markov process X_t with generator $\mathcal{L}$, the previous relation is replaced by

$$\mathcal{L}\mathcal{E}(x) \leq -\beta\mathcal{E}(x) + K \ , \tag{2.11}$$

where we only require $\beta, K > 0$. As a simple consequence of Grönwall's inequality, the existence of a Lyapunov function implies (respectively) that

$$\mathbb{E}\mathcal{E}(X_n) \leq (1-\beta)^n \mathbb{E}\mathcal{E}X_0 + K/\beta, \quad \text{or} \quad \mathbb{E}\mathcal{E}(X_t) \leq e^{-\beta t}\mathbb{E}\mathcal{E}X_0 + K/\beta. \tag{2.12}$$

In other words, $\mathbb{E}\mathcal{E}(X_n)$ (or respectively $\mathbb{E}\mathcal{E}(X_t)$) can be bounded uniformly in time. In this case, we say that X_n (or X_t) is *$\mathcal{E}$-bounded.*

When the sub-level sets of $\mathcal{E}$ are compact, we will call $\mathcal{E}$ a *strong Lyapunov function.* This additional requirement implies that an $\mathcal{E}$-bounded Markov chain revisits a large enough compact set arbitrarily many times and that the distribution X_n forms a tight sequence. Existence of strong Lyapunov functions will be crucial when proving geometric ergodicity.

In this paper, we will assume the kinetic energy of the process, $\mathcal{E}(\cdot) = |\cdot|^2$, is a strong Lyapunov function. Based on our formulation of the random sequence U_n, this is equivalent to the following.

Assumption 2.2 (Kinetic energy principle)**.** *There exist constants* $0 < \beta_h < 1, K_h > 0$, *such that*

$$|\Psi_h(u)|^2 + tr(R_h(u)) \leq (1-\beta_h)|u|^2 + K_h\ ,$$

for all $u \in \mathbb{R}^d$ *where* R_h *is the conditional covariance of the system noise defined in* (2.3)*.*

When the random sequence U_n is generated from discrete time solutions of an SDE u_t, this kinetic energy principle can be verified directly by computing the generator (2.11) or simply checking the drift of the SDE, see Appendix B for more details. Using this convenient argument, we can easily verify that the following examples all satisfy Assumption 2.2.

Example 2.3 (Stochastic turbulence models)**.** *When* U_n *is a time discretization of the SDE* (2.2)*, it suffices to require that for certain* $\beta, K > 0$

$$\mathcal{L}|u|^2 = \psi(u)\cdot u + \frac{1}{2}tr(\Sigma\Sigma^T) \leq -\beta|u|^2 + K, \tag{2.13}$$

since then Assumption 2.2 would hold with $\beta_h = 1 - e^{-\beta h}, K_h = Kh$*. Relation* (2.13) *holds for many stochastic turbulence models, which generally take the form*

$$dU_t = -DU_t dt + B(U_t)dt + f + \Sigma dW_t.$$

The linear operator D represents damping, so its symmetric part $\frac{1}{2}(D^T + D)$ *is positive semidefinite. The nonlinear interaction term B is energy preserving, with* $\langle U_t, B(U_t)\rangle = 0$*. Then it is easy to verify that* (2.13) *holds for* $\beta = \frac{1}{4}\lambda_{min}(D^T + D)$ *and* $K = |f|^2/\beta$*. For more information on these models and their application to turbulence, see [26, 4, 27].*

In the following, we present two well known turbulent systems that all satisfy relation (2.13). Hence, Assumption 2.2 holds for their discrete time formulation.

Example 2.4 (Lorenz 96)**.** *Let $U_t = (u_{1,t}, \dots, u_{N,t})$ be an $N \geq 4$ (usually $N = 40$) dimensional system, with its dynamics given by*

$$\dot{u}_{i,t} = -u_{i-2,t}u_{i-1,t} + u_{i-1,t}u_{i+1,t} - u_{i,t} + F,$$

with the periodic boundary condition $u_{t,k} = u_{t,k-N} = u_{t,k+N}$ for all k and where F is a constant forcing. One can easily show that

$$\dot{|U_t|^2} = -2|U_t|^2 + 2F\sum_{i=1}^{N} u_{i,t} \leq -|U_t|^2 + NF^2.$$

Example 2.5 (Truncated stochastic Navier-Stokes system)**.** *The incompressible stochastic Navier-Stokes equation on a two dimensional torus can be described through the vorticity field*

$$dv_t = \nu\Delta v_t dt - B(\mathcal{K}v_t, v_t)dt + \sum_{k\in\mathbb{Z}^2} \sigma_k e_k dW_{k,t}.$$

Here e_k is the Fourier basis for square integrable functions on the torus, so that $e_k(x) = e^{ik\cdot x}$, $\mathcal{K}$ is the linear Biot-Savart integral operator, mapping e_k to $e_k ik^{\perp}/|k|^2$, B is the advection effect $B(u,v) := (u\cdot\nabla)v$ and $(W_{k,t})_{k\in\mathbb{Z}^2}$ is a sequence of independent Wiener processes. It is well known that for this process the L^2-norm is dissipative in time.

For practical numerical implementation, one needs to truncate the infinite dimensional object v_t. For example, one can ignore the high Fourier modes and assume the truncated v_t has the following Fourier decomposition:

$$\tilde{v}_t = \sum_{k\in I} v_{k,t}e_k + v^*_{k,t}e_{-k}.$$

Here $*$ denote complex conjugacy, and

$$I = \{k = (k_1, k_2) \in \mathbb{Z}^2 : |k| \leq N, k \neq \vec{0}, arg(k_1 + k_2 i) \in [0, \pi)\}.$$

Note by formulation, $\tilde{v}_t$ is real valued, and the dynamics of $\tilde{v}_{k,t}$ can be specified by the dynamics of $v_{k,t}$ as follow:

$$dv_{k,t} = -\nu|k|^2 v_{k,t}dt - \mathbf{P}_k B(\mathcal{K}\tilde{v}_t, \tilde{v}_t)dt + \sigma_k dW_{k,t}.$$

where $\mathbf{P}_k : v \mapsto \langle v, e_k\rangle$ evaluates the k-th Fourier coefficient. Then the full dynamics $U_t = (v_{k,t})_{k\in I}$ follows a energy principle:

$$\mathcal{L}|U_t|^2 = \sum_{k\in I}\mathcal{L}|v_{k,t}|^2 = -2\nu\sum_{k\in I}|k|^2|v_{k,t}|^2 + \sum_{k\in I}\sigma_k^2 \leq -2\nu|U_t|^2 + \sum_{k\in I}\sigma_k^2.$$

Here we used the identity $\langle v, B(\mathcal{K}v, v)\rangle = 0$, so

$$\sum_{k\in I} v_{k,t}\mathbf{P}_k B(\mathcal{K}\tilde{v}_t, \tilde{v}_t) = \langle\sum_{k\in I} v_{k,t}e_k, B(\mathcal{K}\tilde{v}_t, \tilde{v}_t)\rangle = \frac{1}{2}\langle\tilde{v}_t, B(\mathcal{K}\tilde{v}_t, \tilde{v}_t)\rangle = 0.$$

In some other nonlinear models, the stability can be demonstrated only after first applying a linear coordinate change.

Example 2.6 (Lorenz 63)**.** *Let $U_t = (x_t, y_t, z_t)$ be a three dimensional system following an ordinary differential equation (ODE):*

$$\frac{d}{dt}x_t = \sigma(y_t - x_t), \quad \frac{d}{dt}y_t = x_t(r - z_t) - y_t, \quad \frac{d}{dt}z_t = x_t y_t - b z_t.$$

Then we can define

$$\mathcal{E}(U_t) = r x_t^2 + \sigma y_t^2 + \sigma(z_t - 2r)^2,$$

so using Young's inequality and $\beta := \min\{2, 2\sigma, b\}$,

$$\begin{aligned}\frac{d}{dt}\mathcal{E}(U_t) &= -2\sigma(r x_t^2 + y_t^2 + b(z_t - r)^2) + 2b\sigma r^2\\ &\leq -2\sigma(r x_t^2 + y_t^2 + \tfrac{1}{2}b(z_t - 2r)^2) + 4b\sigma r^2\\ &\leq -\beta\mathcal{E}(U_t) + 4b\sigma r^2.\end{aligned}$$

One should note here that $|U_t|^2$ does not satisfy relation (2.13) *for all choices of parameters.*

Section 4 will have a detailed discussion of this type of Lyapunov function, where it will be shown that the Lyapunov dissipation relation (2.10) can be preserved through constant shifts for quadratic functions, but not through linear transformations in general.

3. Bounding the observable energy

In the analysis of ensemble Kalman filters, one natural strategy is obtaining a control over the configuration of the ensemble members. However, such control is very difficult in general, as nonlinear dynamics are known be chaotic and turbulent. In this section, we will discuss one type of condition which circumvents this problem. Generally speaking, this condition requires that the energy of the observable part, $\mathcal{E}(\cdot) = |H\cdot|^2$, be a Lyapunov function for the model U_n. In another words, we assume there is a $\beta_h \in (0,1)$ and $K_h > 0$ such that

$$\mathbb{E}_{n-1}|HU_n|^2 \leq (1-\beta_h)|HU_{n-1}|^2 + K_h, \quad \text{a.s.}$$

This condition can be formulated in terms of the propagation equation (2.1).

Assumption 3.1 (Observable energy criterion)**.** *There exists a $\beta_h \in (0,1)$ and a K_h, such that*

$$|H\Psi_h(u)|^2 + tr(HR_h(u)H^T) \leq (1-\beta_h)|Hu|^2 + K_h\ ,$$

for all $u \in \mathbb{R}^d$. Here R_h again is the conditional covariance of the system noise in (2.3).

The objective of the current section is to show that this property is inherited by the ensemble and square root Kalman filters, so the observable energy of the ensembles, $\sum_k |HV_n^{(k)}|^2$ has uniformly bounded expectation in time.

Under the assumption of full observations $H = I_d$, as made in [7, 17], it is clear that Assumption 3.1 is equivalent to Assumption 2.2, hence the observable energy assumption is quite natural. When $\mathrm{rank}(H) = q = d$, the observable energy is actually equivalent to the standard kinetic energy, as

$$|v| = |(H^T H)^{-1} H^T \cdot Hv| \leq |(H^T H)^{-1} H^T||Hv|, \quad |Hv| \leq |H||v|.$$

However, Assumption 2.2 does not in general imply Assumption 3.1. To achieve this implication, one requires that the dissipation in Assumption 2.2 be 'strong enough'. Section 4 will provide a detailed discussion of when and how Assumption 3.1 can be verified.

3.1. Boundedness of the observable energy for EnKF

The advantage of the observable energy $|HU_n|^2$ over the kinetic energy $|U_n|^2$ is that it is preserved in the assimilation step (2.6). To see this, left multiply the assimilation equation by H and rearrange to obtain

$$HV_n^{(k)} = (I + H\widehat{C}_n H^T)^{-1} H\widehat{V}_n^{(k)} + H\widehat{C}_n H^T (I + H\widehat{C}_n H^T)^{-1} Z_n^{(k)}.$$

The first term on the right can be bounded in terms of $HV_{n-1}^{(k)}$ using Assumption 3.1 and the second term can similarly be bounded in terms of $|HU_n|^2$ and an additive constant. This simple observation is the crux of the following result.

Theorem 3.2. *Assume that the signal process U_n satisfies the observable energy criterion, Assumption 3.1, and let $\{V_n^{(k)}\}_{k=1}^K$ be the EnKF ensemble process. Then*

(i) There exist constants $D, M > 0$ such that

$$\mathbb{E}_{n-1}(|HV_n^{(k)}|^2 + M|HU_n|^2) \leq (1 - \tfrac{1}{2}\beta_h)(|HV_{n-1}^{(k)}|^2 + M|HU_{n-1}|^2) + D \tag{3.1}$$

for each $k = 1 \dots K$ and uniformly in $n \geq 1$. In particular, the function

$$\mathcal{E}(U, V^{(1)}, \dots, V^{(K)}) = \sum_{k=1}^{K} |HV_n^{(k)}|^2 + KM|HU_n|^2$$

is a Lyapunov function for the Markov chain $(U_n, V_n^{(1)}, \dots, V_n^{(K)})$ and hence the signal-ensemble process is $\mathcal{E}$-bounded.

(ii) When $\mathrm{rank}(H) = q = d$, $\mathcal{E}$ is a strong Lyapunov function and

$$\mathbb{E}|U_n|^2 + \sum_{k=1}^{K} \mathbb{E}|V_n^{(k)}|^2$$

is bounded uniformly in $n \geq 0$. The precise upper bound can be read off directly from (2.12)

(iii) *Finally, all the claims above hold for any positive semi-definite choice of $\widehat{C}_n$, in particular any covariance inflation scheme from Section 2.4 satisfies the same relation.*

Proof. Left multiply the first equation of (2.6) with H,

$$\begin{aligned} HV_n^{(k)} &= H\widehat{V}_n^{(k)} - H\widehat{C}_nH^T(I+H\widehat{C}_nH^T)^{-1}(H\widehat{V}_n^{(k)} - Z_n^{(k)}) \\ &= (I+H\widehat{C}_nH^T)^{-1}H\widehat{V}_n^{(k)} + H\widehat{C}_nH^T(I+H\widehat{C}_nH^T)^{-1}Z_n^{(k)}. \end{aligned} \tag{3.2}$$

Then based on elementary Lemma Appendix A.1 and Young's inequality, Lemma Appendix A.2

$$|HV_n^{(k)}|^2 \leq (1+\tfrac{1}{2}\beta_h)|H\widehat{V}_n^{(k)}|^2 + (1+2\beta_h^{-1})|\widehat{Z}_n^{(k)}|^2.$$

Using Assumption 3.1 and the conditional independence of $\zeta_n^{(k)}$

$$\mathbb{E}_{n-1}(|H\widehat{V}_n^{(k)}|^2) = \mathbb{E}_{n-1}(|H\Psi_h(V_{n-1}^{(k)}) + H\zeta_{n-1}^{(k)}|^2) \leq (1-\beta_h)|HV_{n-1}^{(k)}|^2 + K_h\ .$$

Furthermore, by Young's inequality

$$\mathbb{E}_{n-1}(|\widehat{Z}_n^{(k)}|^2) = \mathbb{E}_{n-1}(|HU_n + \xi_n + \xi_n^{(k)}|^2) \leq 2\mathbb{E}_{n-1}(|HU_n|^2) + 4q \leq 2|HU_{n-1}|^2 + 2(K_h+2q).$$

Combining these inequalities and using $(1-\beta_h)(1+\frac{1}{2}\beta_h) < (1-\frac{1}{2}\beta_h)$ we have

$$\mathbb{E}_{n-1}|HV_n^{(k)}|^2 \leq (1-\tfrac{1}{2}\beta_h)|HV_{n-1}^{(k)}|^2 + (2+4\beta_h^{-1})|HU_{n-1}|^2 + (1+\tfrac{1}{2}\beta_h)K_h + 2(1+2\beta_h^{-1})(K_h+2q).$$

On the other hand, by Assumption 3.1, for any $M>0$ we have

$$M\mathbb{E}_{n-1}|HU_n|^2 \leq M(1-\beta_h)|HU_{n-1}|^2 + MK_h.$$

Hence, by fixing M such that $\frac{1}{2}\beta_h M > (2+4\beta_h^{-1})$ and adding the previous two inequalities, we see that we can always find a constant D such that

$$\mathbb{E}_{n-1}(|HV_n^{(k)}|^2 + M|HU_n|^2) \leq (1-\tfrac{1}{2}\beta_h)(|HV_{n-1}^{(k)}|^2 + M|HU_{n-1}|^2) + D.$$

This completes the proof of the first claim. The second claim is simply summing the result of the first claim over all k. And when H is of rank d, the observable energy $|Hv|^2$ is equivalent to the square energy $|v|^2$

$$|v| = |(H^TH)^{-1}H^T \cdot Hv| \leq |(H^TH)^{-1}H^T||Hv|.$$

Finally, notice that we have not used any properties of $\widehat{C}_n$ other than it is positive semi-definite. □

3.2. Boundedness of the observable energy for ESRF

The boundedness of ESRF ensembles is not too different from EnKF, since the assimilation step for mean (2.8) is similar to the assimilation step of EnKF, while the posterior ensemble spread in the observable space can be bounded a.s.

Theorem 3.3. *Assume that the signal process U_n satisfies the observable energy criterion, Assumption 3.1, and let $\{V_n^{(k)}\}_{k=1}^K$ denote either the EAKF or ETKF ensemble. Then*

(i) The observable posterior covariance $HC_nH^T \preceq I_d$ a.s., where

$$C_n := \frac{1}{K}\sum_{k=1}^K (V_n^{(k)} - \overline{V}_n) \otimes (V_n^{(k)} - \overline{V}_n).$$

(ii) There exist constants $D, M > 0$ such that the ensemble mean $\overline{V}_n = \frac{1}{K}\sum_{k=1}^K V_n^{(k)}$ satisfies

$$\mathbb{E}_{n-1}(|H\overline{V}_n|^2 + M|U_n|^2) \leq (1 - \tfrac{1}{2}\beta_h)(|H\overline{V}_{n-1}|^2 + M|HU_{n-1}|^2) + D\ , \tag{3.3}$$

for all integers $n \geq 2$. In particular, the function

$$\mathcal{E}(U, V^{(1)}, \dots, V^{(K)}) = \sum_{k=1}^K |HV_n^{(k)}|^2 + KM|HU_n|^2.$$

is a Lyapunov function for the signal-ensemble process $(U_n, V_n^{(1)}, \dots, V_n^{(K)})$ and hence the process is $\mathcal{E}$-bounded.

(iii) When $\operatorname{rank}(H) = d$, $\mathcal{E}$ is a strong Lyapunov function and

$$\mathbb{E}|U_n|^2 + \sum_{k=1}^K \mathbb{E}|V_n^{(k)}|^2$$

is bounded uniformly in $n \geq 0$. The precise bound can be read off directly from (2.12).

(iv) Again the claims above hold for any choice of positive semi-definite covariance matrix $\widehat{C}_n$, in particular the uniform covariance inflation scheme in Section 2.4.

Proof. From the definition of C_n in both ESRF methods, we have that

$$\begin{aligned} HC_nH^T &= H\widehat{C}_nH^T - H\widehat{C}_nH^T(H^T\widehat{C}_nH + I)^{-1}H\widehat{C}_nH^T \\ &= H\widehat{C}_nH^T(H^T\widehat{C}_nH + I)^{-1}(I + H\widehat{C}_nH) - H\widehat{C}_nH^T(H^T\widehat{C}_nH + I)^{-1}H\widehat{C}_nH^T \\ &= H\widehat{C}_nH^T(H\widehat{C}_nH^T + I)^{-1}. \end{aligned}$$

By Lemma Appendix A.1, we clearly have $0 \preceq HC_nH^T \preceq I_d$. As a consequence,

$$\sum_{k=1}^K |HV_n^{(k)}|^2 = \operatorname{tr}(HC_nH^T) + K|H\overline{V}_n|^2 \leq K|H\overline{V}_n|^2 + d. \tag{3.4}$$

The inequality (3.3) follows almost identically to the proof of (3.1). Indeed, the ensemble mean assimilation step (2.8) implies that

$$\begin{aligned} H\overline{V}_n &= H\overline{\widehat{V}_n} - H\widehat{C}_nH^T(I + H\widehat{C}_nH^T)^{-1}(H\overline{\widehat{V}}_n - Z_n) \\ &= (I + H\widehat{C}_nH^T)^{-1}H\overline{\widehat{V}_n} + H\widehat{C}_nH^T(I + H\widehat{C}_nH^T)^{-1}Z_n. \end{aligned}$$

The only difference between this and the proof of (3.1) is that we need to bound $\mathbb{E}_{n-1}|H\overline{\widehat{V}_n}|^2$, but by Jensen's inequality and (3.4) we have

$$\mathbb{E}_{n-1}|H\overline{\widehat{V}_n}|^2 \le \frac{1}{K}\sum_{k-1}^{K}\mathbb{E}_{n-1}|H\widehat{V}_n^{(k)}|^2 \le \frac{1-\beta_h}{K}\sum_{k=1}^{K}|HV_{n-1}^{(k)}|^2 + K_h \le (1-\beta_h)|H\overline{V}_{n-1}|^2 + d + K_h.$$

Therefore the argument after (3.2) applies to the process $\overline{V}_n$ verbatim.

To show that $\mathcal{E}$ is a Lyapunov function, it suffices to apply Young's inequality

$$|HV_n^{(k)}|^2 \le (1 + \tfrac{1}{4}\beta_h)|H\overline{V}_n|^2 + (1 + 4\beta_h^{-1})|H\overline{V}_n^{(k)} - H\overline{V}_n|^2,$$

and also see that

$$\sum_{k=1}^{K}|HV_n^{(k)} - H\overline{V}_n|^2 = \mathrm{tr}(S_n^TH^THS_n) = \mathrm{tr}(HS_nS_n^TH^T) = \mathrm{tr}(HC_nH^T) \le \mathrm{tr}(I_q) \le q,$$

where S_n is the posterior spread matrix, which is given by $T_n\widehat{S}_n$ or $\widehat{S}_nA_n$ as in (2.9). Therefore

$$\begin{aligned} \mathbb{E}_{n-1}\mathcal{E}(U_n, V_n^{(1)}, \dots, V_n^{(K)}) &\le (1 + \tfrac{1}{4}\beta_h)K\mathbb{E}_{n-1}(|H\overline{V}_n|^2 + M|HU_n|^2) + (1 + 4\beta_h^{-1})q \\ &\le (1 - \tfrac{1}{4}\beta_h)K(|H\overline{V}_{n-1}|^2 + M|HU_{n-1}|^2) + KD + (1 + 4\beta_h^{-1})q \\ &\le (1 - \tfrac{1}{4}\beta_h)(\sum_{k=1}^{K}|HV_{n-1}^{(k)}|^2 + KM|HU_{n-1}|^2) + KD + (1 + 4\beta_h^{-1})q \\ &= (1 - \tfrac{1}{4}\beta_h)\mathcal{E}(U_{n-1}, V_{n-1}^{(1)}, \dots, V_{n-1}^{(K)}) + KD + (1 + 4\beta_h^{-1})q. \end{aligned}$$

Here we applied Jensen's inequality inequality in the penultimate step. The proofs for the second two claims are identical to Theorem 3.2. □

4. Validity of the observable energy criterion

In Section 3, we have established a series of uniform boundeness results based on the observable energy criterion, Assumption 3.1. This criterion is different from the usual energy principle for dynamical systems, Assumption 2.2, although they share similar formulations. In this section, we will demonstrate a few sufficient conditions that lead to Assumption 2.2 when the observation is of full rank, and discuss a few concrete examples where the observation is rank deficient and Assumption 2.2 still holds.

4.1. Full rank observation

As stated earlier, even when H is of full rank, Assumption 2.2 does not necessarily imply Assumption 3.1. Indeed, even though the norms $|\cdot|^2$ and $|H\cdot|^2$ are equivalent, the constants of proportionality may preclude the dissipation relation of Assumption 3.1. In a separate work [21], the authors have constructed a concrete nonlinear system which satisfies the kinetic energy principle, Assumption 2.2, but not the observable energy principle, Assumption 3.1, and which exhibits catastrophic filter divergence [11, 13] with large probability. This indicates the importance of verifying the observable energy criterion over the typical energy criterion.

If the condition number of the matrix H is small enough, a kinetic energy principle implies an observable energy principle. To be specific, define the condition number

$$\mathcal{C}_H := \max\{|Hu||v| : |u| = 1, |Hv| = 1\},$$

which must be finite since H is of full rank. Then we have the following.

Theorem 4.1. *If the kinetic energy principle, Assumption 2.2, holds, then*

$$|H\Psi_h(u)|^2 + tr(HR_h(u)H^T) \leq (1-\beta_h)\mathcal{C}_H^2|Hu|^2 + |H|^2K_h$$

where $\mathcal{C}_H$ is the condition number of matrix H. In particular, if $(1-\beta_h)\mathcal{C}_H^2 < 1$, then the observable energy criterion, Assumption 3.1, also holds; therefore the average square norms of the ensemble members for EnKF and ESQF are bounded uniformly in time as a consequence of Theorems 3.2 and 3.3.

Proof. By Assumption 2.2

$$\begin{aligned}|H\Psi_h(u)|^2 + \mathrm{tr}(HR_h(u)H^T) &\leq |H|^2[|\Psi_h(u)|^2 + \mathrm{tr}(R_h(u))]\\ &\leq (1-\beta_h)|H|^2|u|^2 + |H|^2K_h,\\ &\leq (1-\beta_h)\mathcal{C}_H^2|Hu|^2 + |H|^2K_h.\end{aligned}$$

□

Another situation in which a dissipation criterion implies Assumption 3.1 is when the dissipation is of a higher polynomial order. This is an immediate consequence of Theorem 4.1.

Corollary 4.2. *Suppose that the stochastic system U_n is generated through the solution of the SDE*

$$du_t = \psi(u_t)dt + \Sigma dW_t,$$

as $U_n = u_{nh}$, while for some $\delta, K > 0$

$$\langle u, \psi(u)\rangle \leq -\delta|u|^{1+\delta} + K.$$

Then the kinetic energy principle, Assumption 2.2, holds with any fixed $\beta_h \in (0,1)$. By Theorem 4.1, if rank$(H) = d$, then the expected square norms of the ensemble members for EnKF and ESQF are bounded uniformly in time.

Proof. Since

$$\mathcal{L}|u_t|^2 = 2\langle u_t, \psi(u_t)\rangle + \mathrm{tr}(\Sigma\Sigma^T) \leq -2\delta|u|^{1+\delta} + 2K + \mathrm{tr}(\Sigma\Sigma^T).$$

By Hölder's inequality, for any $\alpha > 0$, there is a constant D such that

$$\mathcal{L}|u_t|^2 \leq -\alpha|u_t|^2 + D.$$

By Grönwall's inequality and Dynkin's formula we obtain that

$$\mathbb{E}|U_1|^2 = \mathbb{E}|u_h|^2 \leq e^{-\alpha h}|u_0|^2 + D\alpha^{-1} = e^{-\alpha h}|U_0|^2 + D\alpha^{-1}.$$

So letting $\alpha = -\ln(1-\beta_h)h^{-1}$ we conclude our proof. □

The higher order dissipation described in Corollary 4.2 has been used in many stochastic turbulence models to obtain better stability, like the canonical scalar model with cubic nonlinearity [28, 29] and the conceptual dynamical model for turbulence in [30]. Moreover, this corollary indicates that in the full rank case EnKF can be stabilized if we are willing to filter with a model error that stabilizes the system. For example instead of running the EnKF with vector field ψ of Lorenz 63 or 96, we can run EnKF with the altered system $\tilde{\psi} = \psi - \lambda|u|u$, with any strictly positive λ, then by the above results the filter will have bounded observable energy on an infinite time horizon.

We now address the situation where an energy principle holds for a linearly translated version of the typical energy. That is

$$\mathbb{E}_{n-1}|HU_n - u^*|^2 \leq \beta|HU_{n-1} - u^*|^2 + K$$

for some fixed $u^* \in \mathbb{R}^q$. Then by Young's inequality Lemma Appendix A.2

$$\begin{aligned}\mathbb{E}_{n-1}|HU_n|^2 &\leq (1+\epsilon)\mathbb{E}_{n-1}|HU_n - u^*|^2 + (1+\epsilon^{-1})|u^*|^2\\ &\leq (1+\epsilon)\{\beta|HU_{n-1} - u^*|^2 + K\} + (1+\epsilon^{-1})|u^*|^2\\ &\leq (1+\epsilon)^2\beta|HU_{n-1}|^2 + (1+\epsilon^{-1})^2(K + 2|u^*|^2). \qquad (4.1)\end{aligned}$$

Therefore our framework can be applied to the Lorenz 63 system, Example 2.6, as long as $H/\mathrm{diag}\{r, \sigma, \sigma\}$ satisfies the condition number requirement of Theorem 4.1.

4.2. Observation with rank deficiency

In the case where H is rank deficient we do not expect the observable energy criterion to hold for broad classes of models. This constraint is in a sense not surprising, since EnKF and ESQF with sparse observations can potentially diverge to machine infinity, in a well documented phenomena known as catastrophic filter divergence [11, 4, 13]. Nevertheless, under certain scenarios one can verify Assumption 3.1 through explicit calculation. We will now verify the observable energy criterion for two concrete examples.

Example 4.3. *Consider a stable linear dynamical system given by $U_n = AU_{n-1} + \zeta_n$, where A produces a contraction,*

$$|Au| \le (1-\beta)|u| \quad \textit{for all } u,$$

with a $\beta \in (0,1)$. For simplicity, we assume that ζ_n are i.i.d. random variables with mean 0 and covariance R. Suppose that H commutes with A, $AH = HA$, then

$$\begin{aligned}\mathbb{E}(|HU_n|^2|U_{n-1}=u) &= |HAu|^2 + \mathbb{E}|H\zeta_n|^2\\ &= |AHu|^2 + tr(HRH^T)\\ &\le (1-\beta)^2|Hu|^2 + tr(HRH^T).\end{aligned}$$

Hence Assumption 3.1 holds.

Example 4.4. *Recall the Lorenz 63 model, Example 2.6, where U_n is given by $u_{nh} = (x_{nh}, y_{nh}, z_{nh})$, where the dynamics of u_t is specified by*

$$\frac{d}{dt}x_t = \sigma(y_t - x_t), \quad \frac{d}{dt}y_t = x_t(r - z_t) - y_t, \quad \frac{d}{dt}z_t = x_ty_t - bz_t.$$

Suppose that we have direct noisy observations of the latter two coordinates, that is $H = diag\{0,1,1\}$. We can define the linearly translated observable energy as $\mathcal{E}_H(u_t) = y_t^2 + (z_t - r)^2$. Direct computation yields

$$\frac{d}{dt}\mathcal{E}_H(u_t) = -2y_t^2 - 2bz_t^2 + 2brz_t \le -\gamma\mathcal{E}_H(u_t) + r^2,$$

where $\gamma = \min\{2,b\}$. Therefore by Grönwall's inequality,

$$\mathcal{E}_H(u_{t+h}) \le e^{-\gamma h}\mathcal{E}_H(u_t) + hr^2.$$

Following our manipulation for linearly translated energy, (4.1), we can conclude that Assumption 3.1 holds by taking a sufficiently small $\epsilon \le 1$ in the following application of Lemma Appendix A.2

$$\begin{aligned}|HU_n|^2 = y_{nh}^2 + z_{nh}^2 &\le (1+\epsilon)\mathcal{E}_H(u_{nh}) + (1+\epsilon^{-1})r^2\\ &\le (1+\epsilon)e^{-\gamma t}\mathcal{E}_H(u_{(n-1)h}) + (1+\epsilon)hr^2 + (1+\epsilon^{-1})r^2\\ &\le (1+\epsilon)^2e^{-\gamma t}(y_{(n-1)h}^2 + z_{(n-1)h}^2) + 3(1+\epsilon^{-1})r^2 + (1+\epsilon)hr^2\\ &= (1+\epsilon)^2e^{-\gamma t}|HU_{n-1}|^2 + 3(1+\epsilon^{-1})r^2 + (1+\epsilon)hr^2.\end{aligned}$$

5. Geometric ergodicity of the ensemble based Kalman filters

The objective of this section is to verify geometric ergodicity for the signal-ensemble process. In particular, if P denotes the Markov transition kernel for the signal ensemble process, then we will show that there exists a constant $\gamma \in (0,1)$ such that

$$\|P^n\mu - P^n\nu\|_{TV} \le C_{\mu,\nu}\gamma^n\ ,$$

where μ, ν are two arbitrary initial probability distributions, $C_{\mu,\nu}$ is a time uniform constant that depends on μ, ν, and $\| \cdot \|_{TV}$ denotes the total variation norm. Furthermore the nonlinear filter has a unique invariant attracting state, the analogue of the asymptotic filter for linear Kalman filters [31, 4]. Hence, geometric ergodicity implies that discrepancies in the initialization of the ensemble filters, which is usually inevitable in practice, will dissipate exponentially with time.

To prove the ergodicity of the signal-ensemble process, we invoke a standard result of Markov chain theory [18]. Here we use a simple adaptation of the form given in [19, Theorem 2.3].

Theorem 5.1. *Let X_n be a Markov chain in a space E such that*

(i) There is a strong Lyapunov function $\mathcal{E} : E \mapsto \mathbb{R}^+$ for the Markov process X_n

(ii) For any fixed $M > 0$, the compact set $C = \{x : \mathcal{E}(x) \leq M\}$ satisfies the minorization assumption. That is, there is a probability measure ν with $\nu(C) = 1$, and a $\eta > 0$ such that for any given set A

$$\mathbb{P}(X_n \in A | X_{n-1} = x) \geq \eta \nu(A)$$

for all $x \in C$.

Then there is a unique invariant measure π and constants $r \in (0, 1), \kappa > 0$ such that

$$\|\mathbb{P}^\mu(X_n \in \ \cdot) - \pi\|_{TV} \leq \kappa r^n \bigg(1 + \int \mathcal{E}(x)\mu(dx)\bigg).$$

In the following we discuss the conditions we require in order to apply Theorem 5.1. In the previous sections, we established the existence of Lyapunov functions for signal-ensemble processes. In particular, when H is of full rank, Theorems 3.2, 3.3 and 4.1 provide simple criteria which guarantee that the filtering process satisfies the following assumption

Assumption 5.2 (Existence of a strong Lyapunov function)**.** *There is a function $\mathcal{E} : \mathbb{R}^d \times \mathbb{R}^{d\times K} \to \mathbb{R}^+$ with compact sublevel sets and constants $K_h, \beta_h > 0$ such that*

$$\mathbb{E}_{n-1}\mathcal{E}(U_n, V_n^{(1)}, \dots, V_n^{(K)}) \leq (1 - \beta_h)\mathcal{E}(U_{n-1}, V_{n-1}^{(1)}, \dots, V_{n-1}^{(K)}) + K_h \ .$$

Although Assumption 5.2 may be difficult to hold in general scenarios, the authors have found an adaptive covariance inflation scheme that always guarantees Assumption 5.2, which will be reported in a separate paper [22]. This assumption provides the first hypothesis of Theorem 5.1.

In order to verify the minorization condition of Theorem 5.1, we need to assume there is a density for the noise ζ_n appearing in the time discrete model (2.1) .

Assumption 5.3 (Nondegenerate system noise)**.** *For any $M_1, M_2 > 0$, there is a constant $\alpha > 0$ such that*

$$\mathbb{P}(\zeta_n \in \cdot \, | U_{n-1} = u) \geq \alpha \lambda_{M_2}(\cdot)$$

for all $|u| \leq M_1$, where $\lambda_{M_2}(dx)$ is the Lebesgue measure of $\mathbb{R}^d$ restricted to $\{u : |u| \leq M_2\}$.

Assumption 5.3 holds for many practical examples. When U_n is produced by time discretization of an SDE (2.2), it suffices to require Σ being nonsingular, please see Appendix B for a detailed discussion. In other situations where U_n is produced genuinely as a random sequence, ζ_n is usually a sequence of random Gaussian variables, which Assumption 5.3 also satisfies.

5.1. Controllability

In this subsection, we establish a framework which verifies the minorization condition using Assumption 5.3 and controllability of the Kalman update map. The signal-ensemble process $X_n := (U_n, V_n^{(1)}, \dots, V_n^{(K)})$ is a Markov chain taking values in $\mathcal{X} = \mathbb{R}^d \times \mathbb{R}^{d\times K}$. For all three ensemble filters, the evolution of X_n is described by the composition of two maps. The first is a random map from $\mathcal{X}$ to a signal-forecast-observation space $\mathcal{Y}$, described by a Markov kernel $\Phi : \mathcal{X} \times \mathcal{B}(\mathcal{Y}) \to [0, 1]$. The second is a deterministic map $\Gamma : \mathcal{Y} \to \mathcal{X}$, which combines the forecast with the observed data to produce the updated posterior ensemble. The details of these maps, as well as the definition of the intermediate space $\mathcal{Y}$, differs between EnKF, ETKF and EAKF.

For EnKF, the intermediate space is $\mathcal{Y} := \mathbb{R}^d \times \mathbb{R}^{d\times K} \times \mathbb{R}^{q\times K}$ and the random mapping is

$$(U_{n-1}, V_{n-1}^{(1)}, \dots, V_{n-1}^{(K)}) \mapsto Y_n := (U_n, \widehat{V}_n^{(1)}, \dots, \widehat{V}_n^{(K)}, Z_n^{(1)}, \dots, Z_n^{(K)}) \,.$$

The deterministic map Γ is given by

$$\Gamma(U_n, \widehat{V}_n^{(1)}, \dots, \widehat{V}_n^{(K)}, Z_n^{(1)}, \dots, Z_n^{(K)}) = (U_n, \Gamma^{(1)}, \dots, \Gamma^{(K)})$$

where

$$\begin{aligned} \Gamma^{(k)} &= \widehat{V}^{(k)} - \widehat{C} H^T (I + H\widehat{C}H^T)^{-1}(H\widehat{V}^{(k)} - Z^{(k)}) \\ \widehat{C} &= \frac{1}{K-1}\sum_{k=1}^{K} (\widehat{V}^{(k)} - \overline{\widehat{V}}) \otimes (\widehat{V}^{(k)} - \overline{\widehat{V}}) \quad \overline{\widehat{V}} = \frac{1}{K}\sum_{k=1}^{K} \widehat{V}^{(k)} \,. \end{aligned} \tag{5.1}$$

The corresponding formulas for ETKF and EAKF will be given in Sections 5.3 and 5.4 respectively.

Given this formuation, it suffices to show that the push-forward kernel $\Gamma^*\Phi(x, \cdot) = \Phi(x, \Gamma^{-1}(\cdot))$ satisfies the minorization condition. It is easy to see that, given the assumptions on the noise, the kernel $\Phi(x, \cdot)$ has a density with respect to Lebesgue measure, so we simply need to show that the pushforward inherits the density property from Φ. To achieve this, we use the following simple fact.

Lemma 5.4. *Let Φ be a Markov transition kernel from $\mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^m$ with a Lebesgue density $p(x, y) = p(x, (y_1, y_2))$ and let $\Gamma : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$. Given a compact set C, suppose that there is a point $y^* = (y_1^*, y_2^*) \in \mathbb{R}^n \times \mathbb{R}^m$ and $\beta > 0$ such that*

(i) For all $x \in C$, the density function $p(x, y) > \beta$ for y around y^* ,

(ii) Γ *is* C^1 *in a neighborhood of* y^* *and* $\det(\mathcal{D}_{y_1}\Gamma)|_{y^*} > 0$ *.*

Then there is a $\delta > 0$ *and a neighborhood* O_1 *of* $\Gamma(y^*)$ *such that for all* $x \in C$

$$\Gamma^*\Phi(x, \cdot) \geq \delta\lambda_{O_1}(\cdot)$$

where λ_{O_1} *is the Lebesgue measure restricted to the set* O_1*. In other words, the minorization condition holds for the transition kernel* $\Gamma^*\Phi$*.*

Proof. By continuity and compactness, we can find a small neighborhood O of y^*, a neighborhood O' of $(\Gamma(y^*), y_2^*)$, and a small positive number $\epsilon > 0$ such that for all $x \in C, (y_1, y_2) \in O$

$$p(x, (y_1, y_2)) > \epsilon, \quad \epsilon^{-1} > |\det(\mathcal{D}_{y_1}\Gamma(y_1, y_2))| > \epsilon \; .$$

such that Γ is a C^1 diffeomorphism from O to O'. Let $D_0 = \{(y_1, y_2) : \Gamma(y_1, y_2) \in A, y_2 \in B, (y_1, y_2) \in O\}$ then, using the change of variables $(y_1, y_2) \mapsto (\Gamma(y_1, y_2), y_2)$, we have by the change of variables formula

$$\begin{aligned} \Phi(x, D_0) &:= \int \mathbb{1}_A(\Gamma(y_1, y_2))\mathbb{1}_B(y_1, y_2)p(x, y_1, y_2)dy_1dy_2 \\ &= \int \mathbb{1}_{(A\times B)}(z, y_2)\mathbb{1}_{O'}(z, y_2)q(x, (z, y_2))dzdy_2 \end{aligned}$$

where

$$q(x, (z, y_2)) = p(x, y_1, y_2)\left|\det\begin{bmatrix} \mathcal{D}_{y_1}z & \mathcal{D}_{y_2}z \\ \mathcal{D}_{y_1}y_2 & \mathcal{D}_{y_2}y_2 \end{bmatrix}\right|^{-1} = p(x, y_1, y_2)|\det(\mathcal{D}_{y_1}\Gamma(y_1, y_2))|^{-1}.$$

with $z = \Gamma(y_1, y_2)$. By construction, q is strictly above ϵ^2 for $x \in C$ and $(z, y_2) \in O'$. Pick neighborhood O_1 of $z^* = \Gamma(y^*)$ and O_2 of y_2 such that $O_1 \times O_2 \subset O'$. Then, since the set $\Gamma^{-1}(A) \cap O$ is of the form D_0 (with $B = \mathbb{R}^m$) we can apply the above change of variables formula to obtain

$$\begin{aligned} \Phi(x, \Gamma^{-1}(A)) \geq \Phi(x, \Gamma^{-1}(A) \cap O) &= \int \mathbb{1}_A(z)\mathbb{1}_{O'}(z, y_2)q(x, z, y_2)dzdy_2 \\ &\geq \int_{O_1\times O_2} \mathbb{1}_A(z)q(x, z, y_2)dzdy_2 \\ &\geq \epsilon^2\lambda(O_1)\lambda(O_2 \cap A) \; . \end{aligned}$$

Then taking $\delta = \epsilon^2\lambda(O_2)$ satisfies our requirement. □

5.2. Ergodicity for the EnKF

In the application of Lemma 5.4 to the signal-ensemble process, we will use the variables

$$x = (U_{n-1}, V_{n-1}^{(1)}, \ldots, V_{n-1}^{(K)})$$

$$y_1 = (U_n, \widehat{V}_n^{(1)}, \dots, \widehat{V}_n^{(K)})$$
$$y_2 = (Z_n, Z_n^{(1)}, \dots, Z_n^{(K)})$$

The choice of the intermediate point (y_1^*, y_2^*) can be quite delicate. Although the non degeneracy of the Jacobian should in principal be verifiable for general Γ, a well chosen intermediate point can simplify the computation significantly.

With Theorem 5.1 and Lemma 5.4, the verification of EnKF becomes rather straight forward.

Theorem 5.5. *If the unfiltered signal process U_n has an kinetic energy principle with nondegenerate system noise, and the EnKF signal-ensemble process has a strong Lyapunov function, in other words Assumptions 2.2, 5.3 and 5.2 hold, then the EnKF signal-ensemble process is geometrically ergodic in total variation distance.*

Proof. Fix any $M_1 > 0$, we apply Lemma 5.4 to the compact set

$$C = \left\{ (u, v^{(1)}, \dots, v^{(K)}) : |u|^2 + \sum_{k=1}^{K} |v^{(k)}|^2 \leq M_1 \right\}.$$

Pick the intermediate point y^* with all its components at the origin. It is easy to see that the first condition of Lemma 5.4 holds. Indeed the random variable $(\zeta_n, \zeta_n^{(k)}, \xi_n, \xi_n^{(k)})$ satisfies the density condition by assumption and (y_1, y_2) is obtained from this random variable via an onto linear transformation, hence (y_1, y_2) inherits the density condition.

It is also elementary to verify the differentiability and nondegeneracy of Γ at y^*, where Γ is defined by (5.1). Indeed, using the formula for gradients of inverse matrices $\mathcal{D}L^{-1} = -L^{-1}\mathcal{D}LL^{-1}$, it is clear that Γ is a polynomial combination of several continuously differentiable functions and in particular must be C^1 near y^*. To prove non-degeneracy, notice that both $\widehat{C}$ and $\mathcal{D}_{y_1}\widehat{C}$ vanish at y^*. Using this fact, a simple calculation yields

$$\mathcal{D}_{y_1}\Gamma|_{y^*} = I \ ,$$

which proves non-degeneracy and hence the Theorem follows from Lemma 5.4 and Theorem 5.1.

□

5.3. Ergodicity of ETKF

For both ETKF and EAKF, the intermediate space is slightly different since there are no longer perturbed observations. In particular we have $\mathcal{Y} := \mathbb{R}^d \times \mathbb{R}^{d\times K} \times \mathbb{R}^q$ and the Markov kernel $\Phi : \mathcal{X} \times \mathcal{B}(\mathcal{Y}) \to [0,1]$ is described by

$$(U_{n-1}, V_{n-1}^{(1)}, \dots, V_{n-1}^{(K)}) \mapsto (U_n, \widehat{V}_n^{(1)}, \dots, \widehat{V}_n^{(K)}, Z_n) \ .$$

The deterministic step is given by the map $\Gamma(U, V, Z) = (U, \Gamma^{(1)}, \dots, \Gamma^{(K)})$ where

$$\Gamma^{(k)} = \overline{\widehat{V}} - \widehat{C}H^T(I + H\widehat{C}H^T)^{-1}(H\overline{\widehat{V}} - Z) + S^{(k)} \tag{5.2}$$

where $\bar{V} = \frac{1}{K}\sum_{k=1}^{K} \widehat{V}^{(K)}$ and $\widehat{C} = \frac{1}{K-1}\sum_{k=1}^{K}(\widehat{V}^{(K)} - \bar{V}) \otimes (\widehat{V}^{(K)} - \bar{V})$ and $S^{(k)}$ is the k-th column of the updated spread matrix $S = \widehat{S}T(\widehat{S})$ where $\widehat{S}$ is the forecast spread matrix $\widehat{S} = (\widehat{V}^{(1)} - \overline{\widehat{V}}, \ldots, \widehat{V}^{(K)} - \overline{\widehat{V}})$ and $T(\widehat{S})$ is the transform matrix. We have not yet defined the transform matrix $T(\widehat{S})$, other than to require that it yields the covariance condition (2.9). One reasonable choice for the transform matrix, which we will adopt, is the matrix square root

$$T(\widehat{S}) = \big(I_K + (K-1)^{-1}\widehat{S}^T H^T H \widehat{S}\big)^{-1/2} = \big(I_K - (K-1)^{-1}\widehat{S}^T H^T (I + H\widehat{C}H^T)^{-1} H \widehat{S}^T\big)^{1/2} . \tag{5.3}$$

Note that the square root is well defined and unique since the argument is symmetric and positive semi-definite. We will now apply Lemma 5.4 with

$$x = (U_{n-1}, V_{n-1}^{(1)}, \ldots, V_{n-1}^{(K)}) \quad y_1 = (U_n, \widehat{V}_n^{(1)}, \ldots, \widehat{V}_n^{(K)}) \quad y_2 = Z_n \ .$$

and the intermediate point $y^* = (y_1^*, y_2^*) = (0, 0)$.

Theorem 5.6. *If the unfiltered signal process U_n has an kinetic energy principle with nondegenerate system noise, and the ETKF signal-ensemble process has a strong Lyapunov function, in other words Assumptions 2.2, 5.3 and 5.2 hold, then the ETKF signal-ensemble process is geometrically ergodic.*

Proof. Fix any number $M_1 > 0$ and define the compact set

$$C = \Big\{(u, v^{(1)}, \ldots, v^{(k)}) : |u|^2 + \sum_{k=1}^{K} |v^{(k)}|^2 \leq M_1\Big\}.$$

As with Theorem 5.5, we pick the intermediate point y^* to be the origin. Showing that Φ satisfies the first condition of Lemma 5.4 is identical to Theorem 5.5, hence it suffices to show differentiability and non-degeneracy.

By Lemma Appendix D.1, the transform matrix $T(\widehat{S})$ is continuously differentiable and hence it follows trivially that Γ is C^1 near y^*. For the non-degeneracy condition, we have that

$$\mathcal{D}_{y_1}\Gamma^{(k)} = \mathcal{D}_{y_1}\overline{\widehat{V}} - \mathcal{D}_{y_1}\big(\widehat{C}H^T(I + H\widehat{C}H^T)^{-1}(H\overline{\widehat{V}} - Z)\big) + \mathcal{D}_{y_1}S^{(k)} \ . \tag{5.4}$$

Precisely as in Theorem 5.5, the second term on the right hand side vanishes at y^*. For the third term, we have by Leibniz rule

$$\mathcal{D}_{y_1}S = \mathcal{D}_{y_1}\widehat{S}T(\widehat{S}) + \widehat{S}\mathcal{D}_{y_1}T(\widehat{S}) \ .$$

But by the definition of $T(\widehat{S})$, it is clear that $\mathcal{D}_{y_1}(T(\widehat{S})T(\widehat{S}))$ vanishes as y^* and that $T(\widehat{S}) = I$ at y^*. Hence we have

$$0 = \mathcal{D}_{y_1}(T(\widehat{S})T(\widehat{S}))|_{y^*} = 2\mathcal{D}_{y_1}T(\widehat{S})|_{y^*} \ ,$$

thus $\mathcal{D}_{y_1}S|_{y^*} = \mathcal{D}_{y_1}\widehat{S}|_{y^*}$. Returning to (5.4), we see that

$$\mathcal{D}_{y_1}\Gamma^{(k)}|_{y^*} = \mathcal{D}_{y_1}\overline{\widehat{V}}|_{y^*} + \mathcal{D}_{y_1}\widehat{S}^{(k)}|_{y^*} = \mathcal{D}_{y_1}\widehat{V}^{(k)}|_{y^*} \ .$$

It follows that $\mathcal{D}_{y_1}\Gamma|_{y^*} = I$ and as in Theorem 5.5, this completes the proof.

□

Remark 5.7. *ETKF has different formulations to* (5.3)*, where* $T(\widehat{S})$ *is obtained by multiplying the original formula by a rotation matrix on the right. The same principles apply to these ETKF, as long as the rotation matrix is differentiable around the intermediate point. However, the origin may not be good choice of intermediate point, since the rotation map can be singular to perturbations around the origin. In this case, one must use the methods in the following section for EAKF, where we deal with the same issue.*

5.4. Ergodicity of EAKF

In this section, we apply the same strategy as EnKF and ETKF to obtain geometric ergodicity for EAKF. For EAKF, the intermediate space $\mathcal{Y}$ and the Markov kernel Φ are identical to those of ETKF. The deterministic map Γ is still defined by (5.2), but now the spread matrix S is defined by $S = A(\widehat{S})\widehat{S}$ where $A(\widehat{S})$ is an adjustment matrix. As with ETKF, the adjustment matrix can be any matrix that ensures the covariance condition (2.9). In the next section, we will discuss how to construct such an adjustment matrix.

5.4.1. Detailed formulation of EAKF We adopt the formulation of EAKF from [3, 4], described by the following steps.

(i) Compute the SVD decomposition of $\widehat{S}$, denoted by $\widehat{S} = Q\Lambda R$. When there is rank deficiency, i.e. Λ is not square or not invertible, we can further decompose Q and R into parts corresponding to null and complementary subspaces:

$$\widehat{S} = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} R_1 \\ R_2 \end{bmatrix} = Q_1\Lambda_1 R_1 \tag{5.5}$$

where Λ_1 is a square diagonal invertible $k \times k$ matrix with $k \le \min(d, K)$. Without loss of generality, we assume Λ_1 has its diagonal entries descending.

(ii) Let $G^T D G$ be the eigenvalue decomposition of $(K-1)^{-1}\Lambda^T Q^T H^T H Q \Lambda$ where D is positive semi-definite with decreasing diagonal entries. As in the first step, in the rank deficient case we can write

$$G_1^T D_1 G_1 = (K-1)^{-1}\Lambda_1 Q_1^T H^T H Q_1 \Lambda_1$$

where G_1, D_1 are $k \times k$ matrices.

(iii) In the general (rank deficient) case, the assimilation matrix is given by

$$A(\widehat{S}) = Q_1\Lambda_1 G_1^T (I + D_1)^{-1/2}\Lambda_1^{-1} Q_1^T \tag{5.6}$$

or equivalently updating the spread matrix to be

$$S = Q_1\Lambda_1 G_1^T (I + D_1)^{-1/2} R_1.$$

In the full rank case this reduces to the more well known formulation

$$A(\widehat{S}) = Q\Lambda G^T(I+D)^{-1/2}\Lambda^{\dagger}Q^T \ . \tag{5.7}$$

In existing EAKF literature, the rank deficient case is rarely examined except as a footnote in [32]. However, rank deficiency is unavoidable when the ensemble size K is less than the model dimension d. As a matter of fact, in these degenerate scenarios, the choice of the eigen-basis is very subtle, and simply following the classical formulation (5.7) with different choice of G could end up with inaccurate posterior covariance. The assimilation matrix formulation (5.6) we used here guarantees that the posterior covariance follows the Kalman update relation (2.7). In practice, direct application of (5.7) in most mathematical programs with default settings will generate the same update rule. A more detailed discussion of these issues are presented in Appendix C.

Since the above EAKF formulation relies on a singular value decomposition, the choice of singular vectors and hence the map Γ has a non-unique definition. To show geometric ergodicity, we rely on differentiability properties of this map Γ. Hence we must rely on a rigid definition of the map Γ involving a fixed choice of singular value decomposition. From here on, any function Γ that fits into the above formulation will be called an *EAKF update map*. The following theorem, which is the main result of the section states that there exists a choice of EAKF update map that renders the EAKF process geometrically ergodic. Due to the complexity in defining Γ, statements concerning the ergodicity of an arbitrary EAKF algorithm seem out of reach.

Theorem 5.8. *There exists an EAKF update mapping Γ such that if the signal-ensemble process generated by it has a strong Lyapunov function, and the unfiltered signal process U_n has an kinetic energy principle with nondegenerate system noise, in other words Assumptions 2.2, 5.2 and 5.3 hold, then the EAKF signal-ensemble process is geometrically ergodic in total variation distance.*

5.4.2. The choice of intermediate point Unlike the ETKF case, we cannot pick the intermediate point to be the origin. This is because at the origin, the spectrum of $\widehat{S}^TH^TH\widehat{S}$ clusters at 0 and a perturbation of $\widehat{S}$ may split the eigenvalues into branches while leaving the eigenvector basis matrices G and R to be singular [33]. We will instead choose an intermediate point that has simple nonzero spectrum.

The intermediate point $y^* = (y_1^*, y_2^*)$ will be defined using a matrix M_0 constructed below. Precisely, we choose

$$y_1^* = (U^*, \widehat{V}^{*(1)}, \dots, \widehat{V}^{*(K)}) = (0, M_0^{(1)}, \dots, M_0^{(K)}) \quad y_2^* = Z^* = 0 \ , \tag{5.8}$$

where $M_0^{(j)}$ denotes the j-th column of M_0. By construction of M_0, this choice will satisfy $\overline{\widehat{V}^*} = 0$ and hence $\widehat{S}^* = M_0$ and moreover will attain the highest possible rank for such a matrix. As we shall see, these properties ensure that the EAKF map is (locally) well behaved under perturbations.

In the sequel, we assume that the observation matrix H is diagonal with rank $q \le d$. Using the change of coordinates described in Remark 2.1 as well as the fact that Assumptions 5.2, 5.3 and the statement of geometric ergodicity hold equivalently in both coordinate systems, this can be achieved without loss of generality.

Lemma 5.9. *If M is a $d \times K$ matrix such that $M\vec{1} = 0$ then the rank of M is at most $r := \min\{K-1, d\}$. Moreover, if we define the $d \times K$ matrix M_0 by*

$$M_0 := \begin{cases} \begin{bmatrix} \vec{1} & \Xi_0 & 0 \end{bmatrix} & \text{if } d \le K-1 \\ \\ \begin{bmatrix} \vec{1} & \Xi_0 \\ 0 & 0 \end{bmatrix} & \text{if } d \ge K-1 \end{cases}$$

where $\vec{1}$ is the $d \times 1$ vector of 1's and Ξ_0 is the $r \times r$ matrix

$$\Xi_0 = \begin{bmatrix} 1 & 1 & \dots & 1 & -r \\ \vdots & \vdots & \vdots & \iddots & 0 \\ 1 & 1 & -3 & \dots & 0 \\ 1 & -2 & 0 & \dots & 0 \\ -1 & 0 & 0 & \dots & 0 \end{bmatrix}$$

then M_0 has the following properties

(i) $M_0\vec{1} = 0$, $\operatorname{rank}(M_0) = r$ *and all nonzero eigenvalues of M_0 are simple.*

(ii) M_0 has the SVD $M_0 = Q_0\Lambda_0R_0$ where Q_0 is the $d \times d$ identity matrix and the last row of R_0 is $\vec{1}^T$.

(iii) When H is a diagonal matrix with descending diagonal entries, $(K-1)^{-1}\Lambda_0Q_0^TH^THQ_0\Lambda_0$ has eigen-decomposition $G_0^TD_0G_0$, where D_0 has descending diagonal entries and is of rank $r_H := \min(q, K-1)$, and G_0 is the $r \times r$ identity matrix.

Proof. It is elementary to see that $\operatorname{rank}(M) \le d$. Also, since M has the same rank as M^TM, while M^TM has $\vec{1}$ as a null right vector, so $\operatorname{rank}(M) \le K-1$. For the claims for M_0, direct verification will be sufficient, where the orthogonality between different rows of M_0 easily leads to

$$M_0M_0^T = \operatorname{diag}\{r(r+1), \dots, 2 \times 3, 1 \times 2, 0 \dots\}$$

Clearly their spectrums are as requested. The claims for G_0 and D_0 easily follow from direct verification. □

5.4.3. Differentiability of EAKF near the intermediate point The intermediate point y^* was chosen to ensure stability of the eigenvalues of $\widehat{S}^T\widehat{S}$ and $\Lambda_1Q_1^TH^TH\Lambda_1Q_1$ as a function of y, in a neighborhood of the intermediate point. We will now demonstrate this.

Lemma 5.10. *We can construct an EAKF update map Γ which is continuously differentiable in a neighborhood of the intermediate point y^*.*

Proof. Recall that Γ is defined by $\Gamma = (U, V^{(1)}, \ldots, V^{(K)})$ where

$$V^{(k)} = \bar{V} + S^{(k)}$$

for each $k = 1, \ldots, K$ where $S^{(k)}$ is the k-th column of S and

$$S = Q_1 \Lambda_1 G_1^T (I + D_1)^{-1/2} R_1, \quad \overline{V} = \overline{\widehat{V}} - \widehat{C} H^T (I + H^T \widehat{C} H)^{-1} [H\overline{\widehat{V}} - Z] \,.$$

The only unspecified parts of the map are the matrices Q, Λ, R and G (and thus Q_1, Λ_1, R_1 and G_1), hence these must be constructed. It suffices to show that all terms appearing in the above map, including the constructed terms G_1, R_1, are continuously differentiable as a function of $y = (U, \widehat{V}^{(1)}, \ldots, \widehat{V}^{(K)}, Z)$ in some neighborhood of y^*.

Clearly $\overline{V}$, $\widehat{S}$ and $\widehat{S}^T \widehat{S}$ depend smoothly on y, differentiability of $\overline{V}$ is shown in the proof of Theorem 5.5. The latter two can be directly seen from the definitions of the mean and spread update maps in (2.8) and (5.7).

We claim that there is a C^1 extension of Q, R as $Q(y), R(y)$ in a neighborhood of $y = y^*$ such that

(i) $Q(y)$ and $R(y)$ are the eigenbasis matrices in the decomposition of $\widehat{S}(y)\widehat{S}^T(y)$ and $\widehat{S}^T(y)\widehat{S}(y)$ respectively.

(ii) There is a diagonal matrix $\Lambda(y)$ such that $\widehat{S}(y) = Q(y)\Lambda(y)R^T(y)$.

To verify the first claim, note that all r nonzero eigenvalues of $\widehat{S}^T(y)\widehat{S}(y)$ are by construction simple at y^*, hence they are smooth with respect to y by II.2.2 of [33], and in particular stay positive, distinct and maintain the same descending order for y near y^*. Moreover, because the rank of $\widehat{S}^T(y)\widehat{S}(y)$ is at most r, 0 will be an eigenvalue of multiplicity $K - r$ for y near y^*. By Lemma Appendix D.2, there is a transformation matrix $U(y)$ that transforms the simple nonzero eigenvectors and null space of $\widehat{S}^T(y^*)\widehat{S}(y^*)$ to the ones of $\widehat{S}^T(y)\widehat{S}(y)$. So $R(y) := R_0 U(y)^T$ provides an eigenbasis matrix for $\widehat{S}(y)^T\widehat{S}(y)$ for y near y^*. Likewise we can also find a $Q(y)$ as the eigenbasis matrix for $\widehat{S}(y)\widehat{S}^T(y)$ for y near y^*.

We will now check the second point of the claim. If there is another SVD, $\widehat{S}(y) = \widetilde{Q}\widetilde{\Lambda}\widetilde{R}$ at y close to y^* with $\widetilde{\Lambda}$ having descending eigenvalues, then $\widetilde{R}$ has its first r rows being the eigenvectors of $\widehat{S}(y)^T\widehat{S}(y)$ associated with descending nonzero eigenvalues, and the remaining $K - r$ rows correspond to the basis of the null space. Therefore, $R(y)$ has the first r rows being either identical to the ones of $\widetilde{R}$ or their additive inverse. Likewise we have the same conclusion for $Q(y)$. Then because the choice of eigenvectors for the null space are unimportant, which can be told from the fact that Q_2, R_2 play no role in (5.5), we have

$$\Lambda(y) := Q(y)^T \widehat{S}(y) R(y)^T$$

is diagonal, and each component is either the same as $\widetilde{\Lambda}$ or the additive inverse. Yet $\Lambda(y)$ is continuous with respect to y, so for there exists a neighborhood of y^* so that they remain nonnegative, this indicates $\Lambda(y) = \widetilde{\Lambda}$ and completes the proof of the two claims.

Similarly, we can find a C^1 function $G_1(y)$ defined in a neighborhood of y^* such that $G_1(y^*) = I_r$ and such that $G_1(y)$ is an eigen-basis matrix for $\Lambda_1(y)Q_1(y)^T H^T H Q_1(y)\Lambda_1(y)$ with corresponding diagonal matrix $D_1(y)$, as in part (ii) of the EAKF formulation.

Finally, the function $(I + D_1)^{1/2}$ is always C^1 in y. To see this, note that

$$D_1 = (K-1)^{-1} G_1 \Lambda_1 Q_1^T H^T H Q_1 \Lambda_1 G_1^T,$$

which is C^1 in y. But since D_1 is diagonal with entries nonnegative, it follows that $(I+D_1)^{-1/2}$ must also be C^1 in y. Hence all terms involved in Γ are C^1 in a neighborhood of y^* and the proof is complete.

□

5.4.4. Controllability of EAKF at the intermediate point

Lemma 5.11. *The EAKF update map Γ constructed in Lemma 5.10 has its Jacobian at y^* being non-degenerate.*

Proof. Before proceeding, we recall some notation. Let $\widehat{S} = \widehat{S}(y)$ be the spread matrix constructed from y, similarly let $Q_1, \Lambda_1, R_1, G_1, D, D_1$ be the matrix valued functions of y that were constructed for the purposes of the EAKF map in Lemma 5.10. In the proof below, we will frequently use the subscript 0 to denote a matrix valued function being evaluated at the intermediate point $y = y^*$.

Recall from Lemma 5.9 that the SVD of $\widehat{S}(y^*) = M_0$ is given by

$$\widehat{S}(y^*) = I_d \Lambda_0 R_0 = I_d \begin{bmatrix} \Lambda_{0,1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} R_{0,1} \\ R_{0,2} \end{bmatrix}.$$

Since the last row of $R_{0,2}$ is $\vec{1}$ by Lemma 5.9, we can also write the SVD decomposition as

$$I_d \Lambda_0 R_0 = I_d \begin{bmatrix} \Lambda_S & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} R_S \\ \vec{1}^T \end{bmatrix},$$

where R_S is the first $K-1$ rows of R_0, and Λ_S is the upper $(K-1)\times(K-1)$ sub-block of Λ_0. We are only interested in the derivatives of Γ in the directions $y_1 = [U, \widehat{V}^{(1)}, \dots, \widehat{V}^{(K)}]$, these directions clearly form a $d(K+1)$ dimensional subspace. We can always write this subspace as

$$\mathcal{M} := \{[u, \overline{v} \otimes \vec{1}_K + B R_S] : u, \overline{v} \in \mathbb{R}^d, B \in \mathbb{R}^{d\times(K-1)}\}.$$

Indeed, it suffices to see that any spread matrix $\widehat{S}$ can be decomposed $\widehat{S} = BR_S$ for some suitable B.

To prove the lemma, suppose that there exists a direction $M \in \mathcal{M}$ such that the derivative of $[U, V^{(1)}, \dots, V^{(K)}]$ in the direction M vanishes at y^*. We will show that M must be zero. Write $M = [u, \overline{v} \otimes \vec{1}_K + BR_S]$. If the derivative vanishes at y^*, we must also have

$$\mathcal{D}_M U = \mathcal{D}_M \overline{V} = 0, \quad \mathcal{D}_M S = 0.$$

at y^*. Through the following steps, we will show that M must be zero. In steps $1-3$ we show $u = \bar{v} = 0$ and in steps $4) - 9)$ we show that $B = 0$.

Step 1) $u = 0$: This is a direct consequence of $\mathcal{D}_M U = u$.

Step 2) If $\widehat{C} = \widehat{S}\widehat{S}^T$ then $\mathcal{D}_M \widehat{C} = 0$ at y^*: Note that since $\mathcal{D}_M S = 0$ at y^*,

$$\mathcal{D}_M[\widehat{C} - \widehat{C}H^T(H^T\widehat{C}H + I)^{-1}H\widehat{C}] = \mathcal{D}_M(SS^T) = 0,$$

where $\widehat{C} = \widehat{S}\widehat{S}^T$ is clearly differentiable. Using the formula of the derivative of an inverse, we find the previous equation is equivalent to

$$\begin{aligned} 0 =& \mathcal{D}_M\widehat{C} - \mathcal{D}_M\widehat{C}H^T(H^T\widehat{C}H + I)^{-1}H\widehat{C} - \widehat{C}H^T(H^T\widehat{C}H + I)^{-1}H\mathcal{D}_M\widehat{C} \\ &+ \widehat{C}H^T(H^T\widehat{C}H + I)^{-1}H^T\mathcal{D}_M\widehat{C}H(H^T\widehat{C}H + I)^{-1}H\widehat{C}. \end{aligned} \tag{5.9}$$

Now let c_{ij} denote the entries of $\mathcal{D}_M\widehat{C}$ and let λ_i, h_i denote the diagonal entries of the diagonal matrices Λ_0 and H respectively. Since the right hand side of (5.9) has all matrices being diagonal except $\mathcal{D}_M\widehat{C}$, we can compute it explicitly, and find the ij-th entry is

$$(\lambda_i^2 h_i^2 + 1)^{-1}(\lambda_j^2 h_j^2 + 1)^{-1} c_{ij}.$$

Therefore (5.9) implies that $\mathcal{D}_M\widehat{C} = 0$ at y^*.

Step 3) $\overline{v} = \vec{0}$: Since $\mathcal{D}_M\overline{V} = 0$ and $\overline{V}$ is given by

$$\overline{V} = \overline{\widehat{V}} - \widehat{C}H^T(I + H^T\widehat{C}H)^{-1}[H\overline{\widehat{V}} - Z]$$

so using the fact that $\mathcal{D}_M\widehat{C} = 0$ at y^*,

$$\mathcal{D}_M\overline{V} = \mathcal{D}_M\overline{\widehat{V}} - \widehat{C}H^T(I + H^T\widehat{C}H)^{-1}[H\mathcal{D}_M\overline{\widehat{V}}]$$

which at y^* can be computed explicitly because the matrices are diagonal:

$$(\mathcal{D}_M\overline{V})_i = (1 + h_i^2\lambda_i^2)^{-1/2}(\mathcal{D}_M\overline{\widehat{V}})_i \quad \Rightarrow \quad 0 = \mathcal{D}_M\overline{\widehat{V}}_i = \overline{v}.$$

Step 4) $\lambda_j B_{ij} + \lambda_i B_{ji} = 0$ and the diagonal terms of B are zero: From here and after, the range of index (i, j) is in $\{1, \dots, d\} \times \{1, \dots, K-1\}$. Our claim comes from the fact that $\widehat{C} = \widehat{S}\widehat{S}^T$ and $\mathcal{D}_M\widehat{S} = BR_S$ so at $y = y^*$ we have

$$0 = \mathcal{D}_M\widehat{C} = (\mathcal{D}_M\widehat{S})\widehat{S}^T(y^*) + \widehat{S}(y^*)(\mathcal{D}_M\widehat{S})^T = BR_0\widehat{S}^T(y^*) + \widehat{S}(y^*)R_0^T B^T = B\Lambda_0^T + \Lambda_0 B^T.$$

Then notice that B is $d \times (K-1)$ dimensional, so its diagonal terms correspond to the first $r = \min\{d, K-1\}$ nonzero entries of λ_i, therefore they are zero.

Step 5) $\mathcal{D}_M Q_1 = 0, \mathcal{D}_M \Lambda_1 = 0$ at y^*: Because Q_1 is formed by eigenvectors of matrix $\widehat{C}$ associated with the nonzero simple eigenvalues, which are the entries of Λ_1, by Lemma Appendix D.2 and formula (D.3) we have our claim because $\mathcal{D}_M \widehat{C} = 0$.

Step 6) $[\mathcal{D}_M(R_1)R_S^T]_{ij} = (\lambda_i B_{ij} + \lambda_j B_{ji})/(\lambda_i^2 - \lambda_j^2)$ at y^* for $i \neq j$: by Lemma Appendix D.2, if we denote r_i to be the i-th row of $R_{0,1}$, while Ψ_i being $e_i^T r_i$, then because $R_{0,2}^T R_{0,2}$ is the projection to the null space of $\widehat{S}^T \widehat{S}$ at y^*,

$$\mathcal{D}_M r_i = r_i \mathcal{D}_M(\widehat{S}^T \widehat{S}) \left[\lambda_i^{-2} R_{0,2}^T R_{0,2} + \sum_{k \neq i}^{r} (\lambda_i^2 - \lambda_k^2)^{-1} \Psi_k^T \Psi_k \right].$$

So for $1 \leq j \leq r$, because $\Psi_k^T \Psi_k r_j^T = 1_{k=j} r_j^T$, $R_{0,2}^T R_{0,2} r_j^T = 0$, we find that

$$[\mathcal{D}_M(R_1)R_S^T]_{ij} = \mathcal{D}_M(r_i) r_j^T = (\lambda_i^2 - \lambda_j^2)^{-1} r_i \mathcal{D}_M(\widehat{S}^T \widehat{S}) r_j^T.$$

For $r \leq j \leq K-1$, because $\Psi_k^T \Psi_k r_j^T = 0$, $R_{0,2}^T R_{0,2} r_j^T = r_j^T$, we find that

$$[\mathcal{D}_M(R_1)R_S^T]_{ij} = \mathcal{D}_M(r_i) r_j^T = \lambda_i^{-2} r_i \mathcal{D}_M(\widehat{S}^T \widehat{S}) r_j^T.$$

Then using $r_i R_S^T = e_i$, the row with zero component except being 1 on i-th coordinate,

$$r_i \mathcal{D}_M(\widehat{S}^T \widehat{S}) r_j^T = r_i (R_S^T B^T \Lambda_0 R_S + R_S^T \Lambda_0^T B R_S) r_j^T = \lambda_j e_i B^T e_j^T + \lambda_i e_i B e_j^T,$$

which concludes our claim.

Step 7) $\mathcal{D}_M G_1 = 0, \mathcal{D}_M D_1 = 0$ at y^*: Because $G_1^T D_1 G_1$ is the eigenvalue decomposition of $(K-1)^{-1} \Lambda_1 G_1^T H^T H G_1 \Lambda_1$, due to Step 5), the derivative in the direction M must vanish at y^*, so by Lemma Appendix D.2 we have our claim.

Step 8) $\mathcal{D}_M D = 0$ and $\mathcal{D}_M D_1 = 0$ at y^*: recall that $G^T D G$ is the eigen-decomposition of $\widehat{S} H^T H \widehat{S}$ and D_1 is the upper $r \times r$ block of D. Notice that D has its entries being the eigenvalues of $\widehat{S}^T H^T H \widehat{S}$, which is the same as $H^T \widehat{S} \widehat{S}^T H = H^T \widehat{C} H$, but from Step 2) $\mathcal{D}_M \widehat{C} = 0$ at y^*, so we have our claim.

Step 9) $B = 0$: By step 4), we only need to verify non-diagonal entries. From $\mathcal{D}_M S = 0$, $S = Q_1 \Lambda_1 G_1^T (I + D_1)^{-1/2} R_1$ and results from Steps 4), 6), 7), we have

$$0 = Q_1 \Lambda_1 G_1^T (I + D_1)^{-1/2} \mathcal{D}_M R_1$$

which by $Q_1 \Lambda_1 = \Lambda_1$ which is invertible at y^* and $G_1 = I_r$ at y^* leads to

$$0 = (I + D_1)^{-1/2} (\mathcal{D}_M R_1) R_S^T$$

at y^*. Because $(I + D_1)^{-1/2}$ is a diagonal matrix with entries $(1 + \lambda_i^2 h_i^2)^{-1/2}$, using the results from Step 6), the (i,j)-th entry of the right hand side is

$$\frac{\lambda_i B_{ij} + \lambda_j B_{ji}}{\sqrt{1 + \lambda_i^2 h_i^2 (\lambda_i^2 - \lambda_j^2)}},$$

so $\lambda_i B_{ij} + \lambda_j B_{ji} = 0$ for $i \neq j$. We claim that $\lambda_i \neq \lambda_j$ for $i \neq j$, this is because $\lambda_i = \lambda_j$ only when they are both zeros, yet either $i \leq d = r$ or $j \leq K - 1 = r$, and hence λ_i or λ_j is not zero. Recall that from Step 4) we have that $\lambda_j B_{ij} + \lambda_i B_{ji} = 0$, combining it with our results that $\lambda_i B_{ij} + \lambda_j B_{ji} = 0$ and $\lambda_i \neq \lambda_j$, this can only be possible when $B_{ij} = 0$.

Therefore, the Jacobian of mapping Γ can not be degenerate at y^*. □

We now have the ingredients required to prove our Theorem 5.8

Proof of Theorem 5.8. In light of Lemmas 5.10 and 5.11, the result follows identically to the proof of Theorem 5.6. □

6. Conclusion and Discusion

In the preceding pages, we have established a simple analytic framework for validating two important nonlinear stability properties of ensemble based Kalman filters, namely boundedness and geometric ergodicity for the signal-ensemble process. These are important properties in practice, as they guarantee the filter processes will remain bounded on an infinite time horizon and that initialization errors in the algorithm dissipate exponentially quickly in time.

In Section 3, upper bounds for the signal-ensemble processes are obtained via a simple Lyapnuov function argument. In particular we show that, if the signal satisfies the *observable energy criterion*, introduced in Assumption 3.1, then one can construct a Lyapnuov function for the signal-ensemble process. The sub-level sets for this Lyapnuov function are only compact in the observed directions. Hence this can be thought of as an upper bound for the observable ensemble $\{HV_n^{(k)}\}_{k=1}^K$. This Lyapnuov function argument is used to construct upper bounds for EnKF, ETKF and EAKF.

Section 4 discusses the applicability of the observable energy criterion. Heuristically speaking, systems with strong dissipation in kinetic energy and complete observations as well as suitable spatial observations will satisfy the observable energy criterion, therefore their EnKF and ESQF ensemble will be bounded uniformly in time. On the other hand, systems without observable energy criterion have the potential to diverge to machine infinity in finite time, which is known as catastrophic filter divergence. This phenomenon has been captured in previous numerical experiments [11, 4, 13]. In a separate article, the authors have constructed a concrete nonlinear system with kinetic energy dissipation but without observable energy dissipation, whose EnKF ensemble diverges to infinity exponentially fast with large probability [21].

In Section 5, geometric ergodicity is established for the signal-ensemble processes of EnKF, ETKF and EAKF. This is achieved by combining the existence of a Lyapnunov function (with compact sub-level sets) with a minorization condition. The existence of a Lyapnuov function can be guaranteed using the results from Section 5, but any other Lyapnuov function would suffice.

While our results have shown that systems with dissipative observable energy will produce stable filter processes, it is hard to believe that Assumption 3.1 is necessary for stability. As discussed earlier, numerical evidence suggests that ensemble based Kalman filters are typically very stable and catastrophic filter divergence only happens when the observations are sparse. The stability and ergodicity of the filter processes here are inherited from the original nonlinear system. This type of inheritance phenomenon generally exists for many filter processes, for example Kalman filter preserves linear stability, and optimal filters preserves absolute regularity of general Markov processes [31, 34, 15, 16]. In order to extend our results, one might seek new dynamical properties that can be inherited through the Kalman assimilation step.

In obtaining our results, we use few properties of the forecast covariance matrix other than positivity. On the one hand this lends generality to our results, in that the upper bounds hold for a broad class of ensemble methods, including covariance inflation schemes. On the other hand, if we had more control over the covariance one might hope to gain more control over the filter ensemble. Thus, it is quite natural to ask whether one can "inflate" the covariance adaptively in order to guarantee stability properties, even in the case where the observation H is of low rank. This question will be investigated by the authors in a subsequent paper [22]. Similar nonlinear stability and ergodicity results as developed here are also valid for finite ensemble Kalman filters which are continuous in time [17] and will be reported elsewhere.

Acknowledgments

This research is supported by the MURI award grant N-000-1412-10912, where A.J.M. is the principal investigator, while X.T.T. is supported as a postdoctoral fellow. D.K. is supported as a Courant Instructor. We thank Ian Grooms and Tom Trogdon for their discussions over some topics of this paper.

Appendix A. Elementary claims

Lemma Appendix A.1. *Let A be a positive semidefinite symmetric matrix, then the following holds:*

$$0 \preceq A(A+I)^{-1} \preceq I, \quad 0 \preceq (A+I)^{-1} \preceq I, \quad 0 \preceq (A+I)^{-1} \preceq A^{-1} ,$$

where $A \preceq B$ means that $B - A$ is positive semidefinite.

Proof. Since $A(A+I)^{-1} + (A+I)^{-1} = I$, it suffices to show $0 \preceq (A+I)^{-1} \preceq I$. Since A is positive semidefinite and symmetric, it can be diagonalized through an orthogonal matrix Ψ, i.e. $A = \Psi D \Psi^T$ with D being diagonal. Then based on $(A+I)^{-1} = \Psi(D+I)^{-1}\Psi^T$, it is elementary to conclude our lemma. □

Lemma Appendix A.2. *By Young's inequality, for any $\epsilon > 0$ and $x, y \in \mathbb{R}^d$, the following holds:*

$$|x+y|^2 = |x|^2 + |y|^2 + 2\langle x, y\rangle \leq (1+\epsilon^2)|x|^2 + (1+\epsilon^{-2})|y|^2.$$

Lemma Appendix A.3. *For any two $d \times d$ positive semidefinite symmetric matrices A and B, the following holds:*

$$tr(A(A+B)^{-1}) = tr((A+B)^{-1}A) \leq tr((A+B)^{-1}(A+B)) = d.$$

Lemma Appendix A.4.

$$S = \widehat{S}[I_K + (K-1)^{-1}\widehat{S}^T H^T H \widehat{S}]^{-1/2}\Theta(\vec{F})[\widehat{S}\widehat{S}^T]^{-1/2}\widehat{S}$$

Proof. Let N be a matrix with its columns form an orthogonal basis of the null space of $\widehat{S}$, i.e.

$$N = [e_1, \dots, e_u], \quad \widehat{S}N = 0, \quad n + \text{rank}(\vec{F}) = d.$$

Then we can find F and X such that

$$\widehat{S}\widehat{S}^T = [N, F]\begin{bmatrix} 0 & 0 \\ 0 & \Sigma \end{bmatrix}\begin{bmatrix} N^T \\ F^T \end{bmatrix}, \quad I_K + (K-1)^{-1}\widehat{S}^T H^T H\widehat{S} = [N, F]\begin{bmatrix} 0 & 0 \\ 0 & \Sigma \end{bmatrix}\begin{bmatrix} N^T \\ F^T \end{bmatrix}$$

□

Appendix B. Discrete time formulation

In many applications, the underlying dynamical system is given by an ODE or SDE, which in general can be written as

$$du_t = \psi(u_t) + \Sigma dW_t.$$

However, the noisy observations of these systems are usually made not in continuous time, but rather sequentially in time, with time interval h. Therefore it is natural to see the stochastic process instead as a stochastic sequence $U_n = u_{nh}$ as we do in Section 2.1. Theoretically speaking, we can always transform the SDE for u_t into the nonlinear update map of U_n as in (2.1), since if we write the transition kernel of process u_t from time 0 to time h as $K_h(u, dv)$, then it suffices to let

$$\Psi_h(u) = \mathbb{E}(u_h | u_0 = u) = \int vK_h(u, dv), \quad \zeta_n = u_{nh} - \Psi_h(u_{(n-1)h}).$$

The reader should notice that the introduction of the nonlinear map Ψ_h and random sequence ζ_n is just for the convenience of our formal illustration and rigorous proof. In order to do the forecast step of EnKF or ESQF in practice, it is not necessary to find the concrete formulation of Ψ_h; it suffices to simulate the SDE starting from each posterior ensemble $V_{n-1}^{(k)}$ from time 0 to h, and let the forecast ensemble $\widehat{V}_n^{(k)}$ be the realization of the simulation at time h.

It is also easy to obtain an energy principle, Assumption 2.2, for the discrete time version $U_n = u_{nh}$, as long as the original SDE satisfies that $\langle u, \psi(u)\rangle \leq -\gamma|u|^2 + k$ for some $\gamma, k > 0$. Because then the generator of the square norm $|u_t|^2$ would satisfy

$$\mathcal{L}|u_t|^2 = 2\langle u_t, \psi(u_t)\rangle + \mathrm{tr}(\Sigma\Sigma^T) \leq -2\gamma|u_t|^2 + 2k + \mathrm{tr}(\Sigma\Sigma^T),$$

which by Grönwall's inequality and Dynkin's formula yields

$$\begin{aligned}\mathbb{E}|u_h|^2 &\leq e^{-2\gamma h}|u_0|^2 + \int_0^h e^{-2\gamma(h-s)}(2k + \mathrm{tr}(\Sigma\Sigma^T))ds\\ &\leq e^{-2\gamma h}|u_0|^2 + h(2k + \mathrm{tr}(\Sigma\Sigma^T)).\end{aligned}$$

Moreover, when the stochastic covariance matrix Σ is nonsingular, by [35] the transition kernel $K_h(u, dv)$ is equivalent to the Lebesgue measure on $\mathbb{R}^d$; also by Theorem 38.16 of [36], this density is smooth both in u and v. As a consequence, the nondegnerate system noise condition, Assumption 5.3, holds because $\{(u, v) : |u| \leq M_1, |v| \leq M_2\}$ is compact. More generally, Assumption 5.3 should hold for discrete time formulation of hypoelliptic systems where certain controllability conditions are satisfied, please refer to [19] and the reference therein.

Appendix C. EAKF with rank deficiency

In this section, we demonstrate that, when there is rank deficiency, in order to achieve the posterior covariance relation (2.7) for EAKF, it is necessary to use specific eigen-decompositions in the formulation of EAKF. In particular, certain choices of eigen-decomposition will result in a violation of (2.7). We will also show that the formulation employed in Section 5.4.1 does always satisfy (2.7).

The following example illustrates that not all eigen-decompoisiton will satisfy (2.7). Let

$$\widehat{S} = \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix}, \quad H = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

then the SVD of $\widehat{S}$ is given by

$$\widehat{S} = Q\Lambda R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

Notice that $\Lambda Q^T H^T H Q \Lambda = 0$, so G can be chosen as any orthonormal matrix. One possible choice of matrix G is R^T and $D = 0$. However, the spread update with $G = R^T$ following the classical formulation is

$$S = Q\Lambda G^T(I + D)^{-1/2}\Lambda^\dagger Q^T \widehat{S} = \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 \end{bmatrix} I_2 \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ 0 & 0 \end{bmatrix}.$$

This posterior spread S does not match the posterior covariance relation (2.7), which requires that $SS^T = \widehat{S}\widehat{S}^T$ since $H = 0$ indicates there is no observation. It is necessary for us to choose $G = I_2$ here, which produces the right posterior covariance,

$$S = Q\Lambda G^T(I+D)^{-1/2}\Lambda^\dagger Q^T\widehat{S} = \begin{bmatrix}\sqrt{2} & 0\\ 0 & 0\end{bmatrix} I_2 \begin{bmatrix}\frac{1}{\sqrt{2}} & 0\\ 0 & 0\end{bmatrix} I_2 \begin{bmatrix}1 & -1\\ 0 & 0\end{bmatrix} = \begin{bmatrix}1 & -1\\ 0 & 0\end{bmatrix}.$$

In general, as long as $G_2R_1^T = R_2G_1^T = 0$ holds in the formulation of EAKF, then the posterior covariance relation (2.7) holds with formulation (5.6). One can see that from

$$SS^T = Q_1\Lambda_1G_1^T(I+D_1)^{-1}G_1\Lambda_1Q_1^T = Q_1\Lambda_1[I+\Lambda_1Q_1^TH^THQ_1\Lambda_1]^{-1}\Lambda_1Q_1^T$$

which satisfies our need because of the following algebraic identity with $A = Q_1\Lambda_1$

$$[I+(K-1)^{-1}A^TH^THA]^{-1} = I-(K-1)^{-1}A^TH^T(I+HAA^TH^T)^{-1}HA.$$

On the practical side, in mathematical computing packages, directly applying the classical EAKF formulation (5.7) will produce the same result as the formula we used (5.6). To see this, we note that

$$\Lambda^TQ^TH^THQ\Lambda = \begin{bmatrix}\Lambda_1 & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}Q_1^T\\ Q_2^T\end{bmatrix}H^TH\begin{bmatrix}Q_1 & Q_2\end{bmatrix}\begin{bmatrix}\Lambda_1 & 0\\ 0 & 0\end{bmatrix} = \begin{bmatrix}\Lambda_1Q_1^TH^THQ_1\Lambda_1 & 0\\ 0 & 0\end{bmatrix}.$$

In the default setting of most mathematical programs, the eigenvalue decomposition will arrange the eigenvalues in decreasing order, and due to the block structure, producing the orthonormal matrix G and eigenvalue matrix D to be

$$G = \begin{bmatrix}G_1 & 0\\ 0 & I_{K-r}\end{bmatrix},\quad D = \begin{bmatrix}D_1 & 0\\ 0 & 0\end{bmatrix}.$$

Then S is given by the following,

$$Q\Lambda G^T(I+D)^{-1/2}\Lambda^\dagger Q^T\widehat{S} = [Q_1\Lambda_1\quad 0]\begin{bmatrix}G_1^T & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}(I+D_1)^{-1/2} & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}\Lambda_1^{-1} & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}Q_1^T\\ Q_2^T\end{bmatrix}Q_1\Lambda_1R_1,$$

which is the same as $Q_1\Lambda_1G_1^T(I+D_1)^{-1/2}R_1$.

Appendix D. Perturbation theory for matrices

In [33], the perturbation theory for matrices are carefully studied. Here we collect some useful results for our study of the controllability of ESRF, where all the section numbers and page numbers mentioned below are referring to [33] if not otherwise specified. One thing the reader must be careful is that most results in [33] are for univariate perturbation, while our interest lies in multivariate perturbation. So we generally follow the strategy of Section

II.5.7, that is trying to express matrices through contour integral of the resolvent, and restrict to cases with stable spectrum structure when eigenprojections are involved (hence why we pick a specific $\vec{F}^*$ in Lemma 5.9).

The most fundamental tool used in [33] is the resolvent of a symmetric positive semidefinite matrix C, given in Section I.5.2 by

$$R(\zeta, C) = (C - \zeta)^{-1}.$$

Clearly $R(\zeta, C)$ is well defined at any ζ not inside the spectrum of C. Moreover if C depends $\mathbf{C}^k$ upon on a multivariate $\vec{F}$, using the derivative formula for matrix inversion, $R(\zeta, C)$ depends $\mathbf{C}^k$ upon $\vec{F}$ as well. The resolvent can be used in the Dunford-Taylor integral to give simple expression for functions of matrix C, see section I.5.6. For example in the computation of ETKF, we need to compute $(I + C)^{-1/2}$, which can be expressed as

$$(I + C)^{-1/2} = \frac{1}{2\pi i} \int_{\Gamma} (1 + \zeta)^{-1/2} R(\zeta, C) d\zeta.$$

Here Γ is a closed loop in the positive half of the complex plain that encloses all eigenvalues of C. Then by the differentiability of $R(\zeta,C)$, it is very easy to see the following lemma

Lemma Appendix D.1. *Let $C(\vec{F})$ be a symmetric semi positive definite matrix that depends $\mathbf{C}^k$ upon $\vec{F} \in \mathbb{R}^n$, then $(I + C(\vec{F}))^{-1/2}$ exists and depends $\mathbf{C}^k$ upon $\vec{F}$.*

When both C and its perturbation are symmetric, the projection to the eigenspace of a particular eigenvalue λ, which is called the eigenprojection (Section I.5.3), can be defined through a contour integral of the resolvent using Cauchy's integral formula:

$$P_\lambda = -\frac{1}{2\pi i} \int_{\Gamma_\lambda} R(\zeta, C) d\zeta,$$

where Γ_λ is a path in the complex half-plane $\{z \in \mathbb{C} : \mathrm{Re}(z) > -1\}$ that encloses λ but not any other eigenvalues. As a reminder, when the eigenvalue λ splits into different branches after perturbation, the perturbed eigenprojection $P_\lambda(\vec{F})$ will be the sum of the eigenprojections of all branches, and therefore it is also known as the total projection of the λ-group (Section II.2.1). This is a slightly nasty case and we try to avoid it by picking a point where the spectrum is stable. In particular, if λ is a simple eigenvalue, P_λ is the orthogonal projection to the eigenvector, and the splitting singularity does not exist for λ. Because $R(\zeta, C)$ is smooth or $\mathbf{C}^k$ w.r.t. $\vec{F}$ and the eigenvalues are continuous with respect to perturbations, the eigenprojection P_λ is smooth or $\mathbf{C}^k$ w.r.t. $\vec{F}$. In particular, the directional derivative in a direction f can be computed as

$$\mathcal{D}_f P_\lambda = P_\lambda(\mathcal{D}_f C)[\sum_{\eta \neq \lambda} (\lambda - \eta)^{-1} P_\eta] + [\sum_{\eta \neq \lambda} (\lambda - \eta)^{-1} P_\eta](\mathcal{D}_f C) P_\lambda. \tag{D.1}$$

Here the summation of η is over all spectrum of C that is not λ. (See page 88 equation (2.14) because all the eigenvalues are semi simple when C is symmetric, there S is given by (5.28) and (5.32) on page 42).

The perturbation to the eigenvectors is studied through transformation functions (Section II.4.2). Consider a perturbation $C(x)$ parametrized by a scalar x such that $C = C(0)$, then the transformation matrix is defined through an ODE:

$$\dot{U}(x) = \sum_{\lambda} \dot{P}_\lambda(x) P_\lambda(x) U(x), \quad U(0) = I_d. \tag{D.2}$$

One can similarly define $U^{-1}(x)$ and consequently show that $U(x)P_\lambda(x)U^{-1}(x) = P_\lambda(0)$ for all eigenvalues λ of C. Moreover, when C and M are symmetric, $U(x)$ is actually an orthogonal matrix (Section II.6.2), so if no eigenvalues splitting occur after the perturbation, then $U(x)$ actually provides a transformation of the eigenvectors of C to the ones of $C(x)$. Therefore, for symmetric perturbation $C(\vec{F})$ small enough, we can define a transformation matrix $U(\vec{F})$ as $U(1)$ with $C_{\vec{F}}(x) = C + x(C(\vec{F}) - C)$. If $C(\vec{F})$ depends $\mathbf{C}^k$ on $\vec{F}$, then because the coefficients of the ODE (D.2) depends $\mathbf{C}^k$ on $\vec{F}$, then so is the solution

$$U(\vec{F}) = \exp\left(\int_0^1 \sum_{\lambda} \dot{P}_{\lambda,\vec{F}}(x) P_{\lambda,\vec{F}}(x) dx\right), \quad \text{with} \quad P_{\lambda,\vec{F}}(x) = -\frac{1}{2\pi i}\int_{\Gamma_\lambda} R(\zeta, C_{\vec{F}}(x)) d\zeta$$

and $\dot{P}_{\lambda,\vec{F}}(x)$ given by (D.1). Notice that $C_{\vec{F}}(\epsilon x) = C_{\epsilon\vec{F}}(x)$, and $D_{\epsilon\vec{F}}C = \epsilon^{-1}D_{\vec{F}}C$ so $\dot{P}_{\lambda,\epsilon\vec{F}} = \epsilon^{-1}\dot{P}_{\lambda,\epsilon\vec{F}}$ using a change of variable formula we get

$$\begin{aligned} U(\epsilon\vec{F}) &= \exp\left(\int_0^1 \sum_{\lambda} \dot{P}_{\lambda,\epsilon\vec{F}}(x) P_{\lambda,\epsilon\vec{F}}(x) dx\right) \\ &= \exp\left(\int_0^\epsilon \sum_{\lambda} \dot{P}_{\lambda,\vec{F}}(x) P_{\lambda,\vec{F}}(x) dx\right). \end{aligned}$$

So the directional derivative of $U(\vec{F})$ at $\vec{F} = 0$ is

$$\begin{aligned} \mathcal{D}_f U(\vec{F})\Big|_{\vec{F}=0} &= \frac{d}{d\epsilon}\exp\left(\int_0^\epsilon \sum_{\lambda} \dot{P}_{\lambda,f}(x) P_{\lambda,f}(x) dx\right)\Big|_{\epsilon=0} = \sum_{\lambda} \dot{P}_{\lambda,f} P_{\lambda,f} \\ &= \sum_{\lambda}\left[P_\lambda(\mathcal{D}_f C)[\sum_{\eta\neq\lambda}(\lambda-\eta)^{-1}P_\eta] + [\sum_{\eta\neq\lambda}(\lambda-\eta)^{-1}P_\eta](\mathcal{D}_f C)P_\lambda\right]P_\lambda. \end{aligned} \tag{D.3}$$

Formula (D.3) can also be presented in a more concrete matrix form through its image over eigenvectors. Suppose the eigenvalue decomposition of a symmetric matrix C be given as $C = \Psi^T\Sigma\Psi$, we consider the perturbation of $\Psi(\vec{F}) = \Psi U(\vec{F})^T$ near $\vec{F}^*$. Let ψ_i be one of the left eigenvectors associated with the i-th eigenvalue λ_l, then P_{λ_i} is give by matrix $\Psi_i^T\Psi_i$, where Ψ_i is Ψ with rows not associated with λ_l removed. Then using $\psi_i\Psi_j^T\Psi_j = 1_{i=j}\psi_i$ we have

$$\mathcal{D}_M\psi_i = \psi_i(\mathcal{D}_f C)\sum_{j\neq i}(\lambda_i - \lambda_j)^{-1}\Psi_j^T\Psi_j. \tag{D.4}$$

To conclude, we have shown that

Lemma Appendix D.2. *Let $C(\vec{F})\in\mathbb{R}^{d\times d}$ be a symmetric semi positive definite matrix that depends $\mathbf{C}^1$ upon $\vec{F}\in\mathbb{R}^n$, then around any point $\vec{F}^*$, there is an orthogonal transformation map $U(\vec{F})$ that depends $\mathbf{C}^1$ upon $\vec{F}$ with directional derivative given by* (D.3) *or matrix representation* (D.4)*. It characterizes the perturbation of eigen-projection by*

$$U(\vec{F})^T P_\lambda(\vec{F})U(\vec{F}) = P_\lambda(\vec{F}^*).$$

In particular, when λ is a simple eigenvalue of $C(\vec{F}^*)$, $U(\vec{F})$ transforms the eigenvector of $C(\vec{F}^*)$ associated with λ to an eigenvector of $C(\vec{F})$ associated with the perturbed λ.

References

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

[2] C. H. Bishop, B. J. Etherton, and S. J. Majumdar. Adaptive sampling with the ensemble transform kalman filter. part i: Theoretical aspects. *Monthly weather review*, 129(3):420–436, 2001.

[3] J. L. Anderson. An ensemble adjustment kalman filter for data assimilation. *Monthly weather review*, 129(12):2884–2903, 2001.

[4] A. J. Majda and J. Harlim. *Filtering complex turbulent systems.* Cambridge University Press, Cambridge, UK, 2012.

[5] E. Kalnay. *Atmospheric modeling, data assimilation, and predictability.* Cambridge university press, 2003.

[6] G. Nævdal and et al. Reservoir monitoring and continuous model updating using ensemble kalman filter. *SPE journal*, 10(1):66–74., 2005.

[7] R. Furrer and T. Bengtsson. Estimation of hign-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis*, 98:227–255, 2007.

[8] D. M. Livings, S. L. Dance, and N. K. Nichols. Unbiased ensemble square root filters. *Physica D*, 237:1021–1081, 2008.

[9] F. Le Gland, Monbet V., and V. D. Tran. Large sample asymptotics for the ensemble kalman filter. *Oxford Handbook of Nonlinear Filtering*, 2010.

[10] J. Mandel, L. Cobb, and J. D. Beezley. On the convergence of the ensemble Kalman filter. *Applications of Mathematics*, 56(6):533–541, 2011.

[11] A.J. Majda and J. Harlim. Catastrophic filter divergence in filtering nonlinear dissipative systems. *Comm. Math. Sci.*, 8:27–43, 2008.

[12] J. Harlim and A.J. Majda. Filtering turbulent sparesly observed geophysical flows. *Mon. Wea. Rev.*, 138:1050–1083, 2010.

[13] G. Gottwald and A. J. Majda. A mechanism for catastrophic filter divergence in data assimilation for sparse observation networks. *Nonlin. Processes Geophys*, 20:705–712, 2013.

[14] D. Crisan and B. Rozovskii. *The Oxford handbook of nonlinear filtering.* Oxford university press, Oxford, 2011.

[15] X. T. Tong and R. Van Handel. Ergodicity and stability of the conditional distributions of nondegenerate markov chains. *Ann. Appl. Probab.*, 22(4):1495–1540, 2012.

[16] X. T. Tong and R. Van Handel. Conditional ergodicity in inifite dimension. *Annals of Probability*, 42(6):2243–2313, 2014.

[17] D. T. B. Kelly, K. J. Law, and A. M. Stuart. Well-posedness and accuracy of the ensemble kalman filter in discrete and continuous time. *Nonlinearity*, 27:2579–2603, 2014.

[18] S. Meyn and R. Tweedie. *Markov chains and stochastic stability.* Springer-Verlag, 1993.

[19] J. C. Mattingly and A. M. Stuart. Geometric ergodicity of some hypo-elliptic diffustion for particle motions. *Markov Process. Related Fields*, 8(2):199–214, 2002.

[20] J. C. Mattingly, A. M. Stuart, and D. J. Higham. Ergodicity for sdes and approximations: Locally lipschitz vector fields and degenerate noise. *Stochastic processes and their applications*, 101(2):185–232, 2002.

[21] D. Kelly, A. J. Majda, and X. T. Tong. Rigorous catastrophic filter divergence with large probability. In preparation.

[22] X. T. Tong, A. J. Majda, and D. Kelly. Stability of ensemble kalman filter with covariance inflation. In preparation.

[23] G. Evensen. Sampling strategies and square root analysis schemes for the enkf. *Ocean dynamics*, 54(6):539–560, 2004.

[24] J. L. Anderson and S. L. Anderson. A monte carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Mon. Weather Rev.*, 127:2741–2758, 1999.

[25] T. M. Hamill, C. Whitaker, and C. Snyder. Distance-dependent filtering of background error covariance estimates in an ensemble kalman filter. *Mon. Weather Rev.*, 129:2776–2790, 2001.

[26] A. J. Majda, R. V. Abramov, and M. J. Grote. *Information theory and stochastics for multiscale nonlinear systems*, volume 25. American Mathematical Soc., 2005.

[27] A. J. Majda and X. T. Tong. Geometric ergodicity for piecewise contracting processes with applications for tropical stochastic lattice models. Accepted by Comm. Pure Appl. Math, 2014.

[28] A. J. Majda, C. Franzke, and D. T. Crommelin. Normal forms for reduced stochastic climate models. *Proc. Natl. Acad. Sci.*, 16(10):3649–3653, 2009.

[29] A. J. Majda and M. Branicki. Lessons in uncertainty quantification for turbulent dynamical system. *Discrete and Continuous Dynamical Systems*, 32(9):3133–3221, 2013.

[30] A. J. Majda and Y. Lee. Conceptual dynamical model for turbulence. *Proc. Natl. Acad. Sci.*, 111(18):6548–6553, 2014.

[31] A. H. Jazwinski. *Stochastic Processes and Filtering Theory.* Academic Press, New York, 1972.

[32] M. K. Tippett, J. L. Anderson, C. H. Bishop, T. M. Hamill, and C. Whitaker. Ensemble square root filters. *Mon. Wea. Rev.*, 131:1485–1490, 2003.

[33] T. Kato. *A Short Introduction to Perturbation Theory for Linear Operators.* Springer-Verlag, 1982.

[34] K. Reif, S. Günther, and E. Yaz. Stochastic stability of the discrete-time extended kalman filter. *IEEE Trans on Automatic Control*, 44(4), 1999.

[35] D. W. Strook and S. R. S. Varadhan. On the support of diffusion processes with applications to the strong maxium principle. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, 3:333 – 359, 1972.

[36] L. C. G. Rogers and D. Williams. *Diffusions, Markov processes, and martingales. Vol. 2. Itô calculus.*, volume 2. Cambridge University Press, 2000.