Jana de Wiljes | Xin T. Tong

Analysis of a localised nonlinear ensemble Kalman Bucy filter with complete and accurate observations

Suggested citation referring to the original publication:
Nonlinearity 33 (2020) 9, 4752–4782
DOI https://doi.org/10.1088/1361-6544/ab8d14
ISSN 1361-6544 (online), 0951-7715 (print)

Journal article | Version of record

OPEN ACCESS

IOP Publishing | London Mathematical Society Nonlinearity

Nonlinearity **33** (2020) 4752–4782 https://doi.org/10.1088/1361-6544/ab8d14

Analysis of a localised nonlinear ensemble Kalman Bucy filter with complete and accurate observations

Jana de Wiljes[1,2,4] **and Xin T Tong**[3]

[1] Universität Potsdam, Institut für Mathematik, Karl-Liebknecht-Str. 24/25, D-14476 Potsdam, Germany
[2] University of Reading, Department of Mathematics and Statistics, Whiteknights, PO Box 220, Reading, RG6 6AX, United Kingdom
[3] National University Singapore, 10 Lower Kent Ridge Road, 119076, Singapore

E-mail: wiljes@uni-potsdam.de

Received 27 August 2019, revised 25 March 2020
Accepted for publication 24 April 2020
Published 28 July 2020

Abstract

Concurrent observation technologies have made high-precision real-time data available in large quantities. Data assimilation (DA) is concerned with how to combine this data with physical models to produce accurate predictions. For spatial–temporal models, the ensemble Kalman filter with proper localisation techniques is considered to be a state-of-the-art DA methodology. This article proposes and investigates a localised ensemble Kalman Bucy filter for nonlinear models with short-range interactions. We derive dimension-independent and component-wise error bounds and show the long time path-wise error only has logarithmic dependence on the time range. The theoretical results are verified through some simple numerical tests.

Keywords: data assimilation, stability and accuracy, dimension independent bound, localisation, high dimensional, filter, nonlinear,

Mathematics Subject Classification numbers: 65P40, 65Z05, 62F15, 62L10.

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

1. Introduction

With the advancement of technology, we now have access to vast amounts of high-precision data in many areas of science. It is important to develop robust and efficient tools to combine

[4] Author to whom any correspondence should be addressed.

1361-6544/20/094752+31$33.00 © 2020 The Author(s). Published by IOP Publishing Ltd & London Mathematical Society Printed in the UK

the available data with refined large-scale physical models. This study is known as data assimilation (DA) and typically the goal is to produce accurate real-time estimations of the current state of the system.

In geophysical problems, the considered models often have vast spatial scales, therefore millions of state variables are needed to store information at different locations. Such high dimensionality poses a severe challenge to DA methodologies, since the associated computations are expensive and direct global uncertainty quantification tends to be erroneous. Over the last two decades, various computationally feasible approaches have been developed with practical success [7, 14, 24, 27]. One of the most popular algorithms among these is the ensemble Kalman filter (EnKF). It has been first derived in [7] and heavily advanced and employed in the field of numerical weather prediction. To combat dimensionality issues arising due to the extent of the spatial domain, the so called localisation techniques are often employed for the EnKF [9, 26]. The key motivation behind localisation is that many systems exhibit a natural decrease in spatial correlation. This can guide artificial tunings of the empirical covariance matrix to avoid spurious correlations.

The empirical success of EnKF has aroused great interest in understanding the underlying theoretical properties [1, 2, 15, 29]. EnKFs can be interpreted as Monte Carlo implementations of the Kalman filter [8, 12, 13, 17] which is derived for linear prediction and observation models. Therefore most theoretical studies of EnKFs assume a linear setting [4, 6, 21, 22, 28]. Existing analysis of EnKFs for nonlinear models concern mostly the boundedness of algorithm outputs [15, 16, 29], which is not helpful in understanding EnKF performance. The only exception is a recent work [5], where accuracy and stability results have been derived assuming abundant and accurate observations. However, the results there do not consider localisation, and hence they require the sample size to be larger than the state dimension. This is infeasible in practice.

This paper intends to close the aforementioned gaps, i.e. nonlinearity and high dimensionality in filter performance analysis, by investigating a localised ensemble Kalman–Bucy filter (l-EnKBF). Following [5], we assume abundant and accurate observations are available. Since most geophysical models are formulated through partial differential equations or their discretisations, the associated prediction dynamics often have a short interaction range. This is often paired with a short decorrelation length in the localisation technique to reduce the potential spurious long-range correlations. Under these assumptions, we show that l-EnKBF estimation error for *each component is bounded independent of the overall dimension*, both in the sense of mean square and the moment generating function. Such result does not exist in literature for DA analysis, based on our knowledge. Some related dimension-independent error analysis can be found in [21, 28], but the error estimates are implicit and the models are assumed to be linear. Moreover, we also show the long time path-wise error has a logarithmic dependence on the time range, which is much weaker than the square root dependence in [5]. All these results indicate l-EnKBF has stable and accurate estimation skills.

In section 2 the underlying setting is outlined and the considered l-EnKBF will be defined. Upper and lower bounds for the empirical second moment are derived in section 3.1. Then point-wise and path-wise bounds for the mean squared error and a Laplace type condition are derived in section 3.2 in the l_2 sense, and in section 3.3 in the component-wise sense. We allocate the proofs of our results in the appendix. In section 4, the numerical sensitivity of an implementation of the considered l-EnKBF with respect to the underlying assumptions is tested for the Lorenz 96 system.

Throughout the article we assume $(\|\cdot\|, \langle\cdot,\cdot\rangle)$ denotes the l_2-norm with its corresponding inner product. Given a matrix $A \in \mathbb{R}^{m\times n}$ the l_2-operator norm is defined as

$$\|A\| = \max_{\|x\|=1}\|Ax\| = \sqrt{\lambda_{\max}(A^\top A)},$$

where $\lambda_{\max}$ denotes the largest eigenvalue of a matrix. The following two matrix norms are also useful to us:

$$\|A\|_1 = \max_{1\leqslant j\leqslant n}\sum_{i=1}^{m}|A_{i,j}|$$

$$\|A\|_{\max} = \max_{ij}|[A]_{i,j}|$$

where both $A_{i,j}$ and $[A]_{i,j}$ denote the entries of the matrix A. The bracket notation is necessary to denote matrix entries such as $[A^{-1}]_{i,j}$ or $[AB]_{i,j}$. Given two symmetric matrices A and B then $A \succeq B$ implies the matrix $A - B$ is positive semidefinite, which is equivalent to $v^{\mathrm{T}}(A-B)v \geqslant 0$ for all $v \in \mathbb{R}^n$. Given a covariance matrix Γ the Mahalanobis norm is defined by $\|v\|_\Gamma^2 = v^{\mathrm{T}}\Gamma^{-1}v$. Lastly, in order to describe the smallness of certain quantities, we use the big theta notation. In particular, a quantity a_ϵ is $\Theta(\epsilon^p)$, if there is a ϵ-independent constant $C > 0$ and $c > 0$ so that $c\epsilon^p \leqslant a_\epsilon \leqslant C\epsilon^p$.

2. Problem setup

In this paper, we consider a continuous-time filtering problem, formulated by

$$\begin{aligned} \mathrm{d}X_t &= f(X_t)\,\mathrm{d}t + \sqrt{2}\sigma\mathrm{d}W_t,\\ \mathrm{d}Y_t &= HX_t\mathrm{d}t + R\mathrm{d}B_t. \end{aligned} \tag{1}$$

In (1), $X_t \in \mathbb{R}^{N_x}$ represents the system we try to recover. We assume its initial distribution is given by $X_0 \sim \pi_0$. Its dynamics is driven by a deterministic forcing described by a map $f: \mathbb{R}^{N_x} \to \mathbb{R}^{N_x}$ and a stochastic forcing term $\sqrt{2}\sigma\mathrm{d}W_t$. We assume linear noisy observations $Y_t \in \mathbb{R}^{N_y}$ of the system are available. In (1), the matrices σ and R are positive definite matrices, and $W_t \in \mathbb{R}^{N_x}$ and $B_t \in \mathbb{R}^{N_y}$ are independent Wiener processes.

In many spatial models, each model component is representing a state information at one spatial location. This introduces a natural distance between two indices, which we will denote as $\mathbf{d}$. As a simple example, For example, if the indices are representing themself on the interval $[1, n]$, then $\mathbf{d}(i,j)$ can be taken as $|i-j|$. For another example, if the indices are representing equally spaced points on a length n circle, then $\mathbf{d}(i,j)$ be taken as $\min\{|i-j|, n-|i-j|\}$.

We will use $x_i(t)$ to denote the ith component of X_t, so $X_t = [x_1(t), \ldots, x_{N_x}(t)]^{\mathrm{T}}$. We will also use f_i and $w_i(t)$ to denote the ith component of f and W_t. For notational simplicity, we will often write $x_i(t)$ as x_i and $w_i(t)$ as w_i, whenever their dependence on time is evident. Then the SDE that x_i follows is given by

$$\mathrm{d}x_i = f_i(X_t)\mathrm{d}t + \sqrt{2}\sigma\mathrm{d}w_i. \tag{2}$$

Note that different components are interacting through the drift term, as $f_i(X_t)$ could have dependence on $x_j(t)$ for $j \neq i$. But in many physical processes, such interactions are of short range, meaning the dependence of $f_i(X_t)$ on $x_j(t)$ decays with $\mathbf{d}(i,j)$. More generally, this can be formulated as

Assumption 2.1 (short range interaction). There is a sequence of Lipschitz constants $\mathcal{F}_k$, such that for any $X = [x_1, \ldots, x_{N_x}]$ and $X' = [x'_1, \ldots, x'_{N_x}]$, the following holds

$$|f_i(X) - f_i(X')| \leqslant \sum_{j=1}^{N_x} \mathcal{F}_{\mathbf{d}(i,j)} |x_j - x'_j|.$$

To have a single number controlling the overall stability of the system, we will consider the largest row sum of these Lipschitz constants and define

$$C_f := \max_i \sum_{j=1}^{N_x} \mathcal{F}_{\mathbf{d}(i,j)}. \tag{3}$$

We will assume that C_f is a constant independent of the dimension N_x. This can be verified if $\mathcal{F}_k$ decays to zero exponentially with increasing k. In section 4, we demonstrate how to verify assumption 2.1 on the Lorenz 96 model, assuming all components are bounded.

In computational models, assumption 2.1 often holds if the spatial resolution is at the same scale of the spatial correlation length. A large N_x indicates that the spatial domain size is large. It is worthwhile mentioning that, it is also possible to obtain a high dimensional model with a moderate size spatial domain, if one use very small spatial resolution. But assumption 2.1 is unlikely to hold in such a setting, and localisation techniques are not meant to resolve such high dimensionality. One should use dimension reduction techniques instead [21]. The difference between these two high dimensional settings are discussed in [23, 30].

2.1. Localised ensemble Kalman–Bucy filter

Here we will consider a deterministic EnKBF first proposed in [2] that has been shown to be the time limit of a broad class of ensemble square root filters [18]. Let $\{X_t^i\}_{i=1,\ldots,M}$ be the ensemble of particles which describe the uncertainty of X_t. To run the considered algorithm, each of the particle is initialised at a random location from π_0 and then driven by the following dynamics

$$\mathrm{d}X_t^i = f(X_t^i)\mathrm{d}t + \sigma^2 P_t^{-1}(X_t^i - \overline{X}_t)\mathrm{d}t - \frac{1}{2}P_t H^{\mathrm{T}}(RR^{\mathrm{T}})^{-1}(HX_t^i\mathrm{d}t + H\overline{X}_t\,\mathrm{d}t - 2\mathrm{d}Y(t)). \tag{4}$$

In (4), the sample mean and covariance are defined by

$$\overline{X}_t = \frac{1}{M}\sum_{i=1}^{M} X_t^i, \quad P_t = \frac{1}{M-1}\sum_{i=1}^{M}(X_t^i - \overline{X}_t)(X_t^i - \overline{X}_t)^{\mathrm{T}}.$$

The posterior distribution of X_t conditioned on $Y_{s\leqslant t}$ is then approximated by the Gaussian distribution $\mathcal{N}(\overline{X}_t, P_t)$. It is important to mention that for a linear drift f the EnKBF in (4) converges to the KBF for $M \to \infty$. Further a mean-field limit has been derived for the nonlinear drift scenario [5]. Note that mean field limits of EnKFs for a nonlinear setting have also been derived in [19].

When the dimension is high, EnKBF is in general ill-defined and it can perform poorly. This is because of two reasons. First, the rank of P_t is at maximum $M-1$. So if $M \ll N_x$, P_t is singular and its numerical approximated inverse is usually unstable. Second, by random matrix theory, it is known that if X_t^i are i.i.d. samples from a Gaussian distribution $\mathcal{N}(0, P)$, in order for the covariance sampler error in l_2-norm $\|P - P_t\|$ to be small, one needs $M = O(N_x)$. In other words, P_t is a very inaccurate approximation of the true posterior covariance when $M \ll N_x$ [28].

In practice, one popular way to resolve the issues mentioned above is to apply covariance localisation. Mathematically, this operation can be formulated as replacing P_t in (4) with

$P^{\mathrm{L}}_t = P_t \circ \phi$. Here $\circ$ denotes the component-wise product or Schur product, so the components of P^{L}_t are defined as

$$[P^{\mathrm{L}}_t]_{i,j} := [P_t]_{i,j}\phi_{i,j}. \tag{5}$$

The symmetric matrix ϕ here is called a localisation matrix. Its components are nonnegative. They are of value 1 at the diagonal, and decay to zero extremely fast along the off diagonal direction. One popular choice takes the form of $\phi_{i,j} = \rho(\frac{\mathbf{d}(i,j)}{l})$, where ρ is a function from (4.10) in [9]

$$\rho(x) = \begin{cases} -\dfrac{1}{4}x^5 + \dfrac{1}{2}x^4 + \dfrac{5}{8}x^3 - \dfrac{5}{3}x^2 + 1, & |x| \leqslant 1; \\ \dfrac{1}{12}x^5 - \dfrac{1}{2}x^4 + \dfrac{5}{8}x^3 + \dfrac{5}{3}x^2 - 5x + 4 - \dfrac{2}{3x}, & 1 \leqslant |x| \leqslant 2; \\ 0 & 2 \leqslant |x|. \end{cases} \tag{6}$$

where l denotes the typical decorrelation length, which we assume to be independent of N_x. We will consider again the largest row sum of ϕ, and define

$$C_\phi := \max_i \sum_{j=1}^{N_x} \phi_{i,j}. \tag{7}$$

We will assume C_ϕ is a constant independent of the dimension N_x. This is true for most practical localisation matrices including (6).

When the true covariance matrix is spatially localised, P^{L}_t is a much better covariance estimator, because the localisation operation eliminates spurious long distance correlation errors [3]. Moreover, the localisation operation improves the rank, so P^{L}_t is often full rank and invertible. But this is not guaranteed in general. So for the rigorousness of this exposition, we use the following inversion

Definition 2.2. If all diagonal entries of P_t are nonzero, then its diagonal inverse (DI) is given by

$$[P^\dagger_t]_{i,i} = [P_t]^{-1}_{i,i}, \quad [P^\dagger_t]_{i,j} = 0, \quad \forall i, j = 1, \dots, n, \ i \neq j.$$

Note that it satisfies the following for all $i = 1, \dots, n$

$$[P^\dagger_t P_t]_{i,i} = [P^\dagger_t P_t]_{i,i} = 1. \tag{8}$$

In the original EnKBF formulation (4), we replace P_t with P^{L}_t and P^{-1}_t with $P^\dagger_t$ and we obtain the localised EnKBF (l-EnKBF):

$$\mathrm{d}X^i_t = f(X^i_t)\mathrm{d}t + \sigma^2 P^\dagger_t(X^i_t - \overline{X}_t)\mathrm{d}t - \frac{1}{2}P^{\mathrm{L}}_t H^{\mathrm{T}}(RR^{\mathrm{T}})^{-1}(HX^i_t\mathrm{d}t + H\overline{X}_t\,\mathrm{d}t - 2\mathrm{d}Y_t). \tag{9}$$

As a remark, the using of $P^\dagger_t$ simplifies the theoretical derivation in below, since we can verify that $P^\dagger_t$ is well defined (see lemma 3.2 below). Meanwhile, it is an open question on how to generalise our results to other versions of pseudo inverse for P_t.

2.2. Abundant and accurate observations

When the observation sources are abundant, H in (1) can be assumed to be of rank N_x, and there is an $H^- \in \mathbb{R}^{N_x \times N_y}$ such that $H^- H = I_{N_x}$. We can consider the following transformation

$$\widetilde{X}_t = \sigma^{-1} X_t, \quad \widetilde{f}(X) = \sigma^{-1} f(\sigma X), \quad \widetilde{Y}_t = \sigma^{-1} H^- Y_t, \quad \widetilde{R} = \sigma^{-1} H^- R,$$

then $\widetilde{X}_t$ and $\widetilde{Y}_t$ follow the SDE in below

$$\begin{aligned} \mathrm{d}\widetilde{X}_t &= \sigma^{-1}\mathrm{d}X_t = \widetilde{f}(\widetilde{X}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \\ \mathrm{d}\widetilde{Y}_t &= \sigma^{-1}H^-\mathrm{d}Y_t = \sigma^{-1}X_t + \sigma^{-1}H^- R\mathrm{d}B_t = \widetilde{X}_t\mathrm{d}t + \widetilde{R}\mathrm{d}B_t. \end{aligned} \tag{10}$$

If we apply l-EnKBF (9) to the transformed system $(\widetilde{X}_t, \widetilde{Y}_t)$, then the sample mean and covariance matrices will follow

$$\overline{\widetilde{x}}_t = \sigma^{-1}\overline{X}_t, \quad \widetilde{P}_t = \sigma^{-2}P_t, \quad \widetilde{P}^L_t = \sigma^{-2}P^{\mathrm{L}}_t,$$

while $\widetilde{P}^\dagger_t$ can be taken as $\sigma^2 P^\dagger_t$. Then the dynamics of each l-EnKBF particle will satisfy

$$\begin{aligned} \mathrm{d}\widetilde{X}^i_t &= \widetilde{f}(\widetilde{X}^i_t)\mathrm{d}t + \widetilde{P}^\dagger_t(\widetilde{X}^i_t - \overline{\widetilde{x}}_t)\mathrm{d}t - \frac{1}{2}\widetilde{P}^{\mathrm{L}}_t(\widetilde{R}\widetilde{R}^{\mathrm{T}})^{-1}(\widetilde{X}^i_t\mathrm{d}t + \overline{\widetilde{x}}_t\,\mathrm{d}t - 2\mathrm{d}\widetilde{Y}_t) \\ &= \sigma^{-1}f(X^i_t)\mathrm{d}t + \sigma P^\dagger_t(X^i_t - \overline{X}_t)\mathrm{d}t - \frac{1}{2}\sigma^{-1}P^{\mathrm{L}}_t H^{\mathrm{T}}(RR^{\mathrm{T}})^{-1}(HX^i_t\mathrm{d}t + H\overline{X}_t\,\mathrm{d}t - 2\mathrm{d}Y_t) \\ &= \sigma^{-1}\mathrm{d}X^i_t. \end{aligned}$$

It is evident that the theoretical properties of X^i_t will be the same as the ones of $\widetilde{X}^i_t$.

Note that (10) corresponds to the original model (1) with $\sigma = 1$ and $H = I$. This is a much simplified parameter setting for followup discussion. And from the above derivation, there is no sacrifice of generality by focussing on it. Under this setting, the l-EnKBF formula will be simplified as

$$\mathrm{d}X^i_t = f(X^i_t)\mathrm{d}t + P^\dagger_t(X^i_t - \overline{X}_t)\mathrm{d}t - \frac{1}{2}P^{\mathrm{L}}_t\Omega_R(X^i_t\mathrm{d}t + \overline{X}_t\,\mathrm{d}t - 2\mathrm{d}Y_t), \quad \Omega_R := (RR^{\mathrm{T}})^{-1}.$$

When the observations are accurate and independent, the observation noise covariance RR^{T} is a diagonal matrix with small components. We will use ϵ to describe their order. In summary, we have made the following assumption

Assumption 2.3. Through a linear transformation, we assume (1) is transformed to

$$\begin{aligned} \mathrm{d}X_t &= f(X_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t, \\ \mathrm{d}Y_t &= X_t\mathrm{d}t + R\mathrm{d}B_t \end{aligned}$$

Moreover we assume for an $\epsilon > 0$ that $\Omega = \epsilon(RR^{\mathrm{T}})^{-1}$ is diagonal, and bounded by constants $\omega_{\min} I \preceq \Omega \preceq \omega_{\max} I$.

Note that assumption 2.3 implies that $RR^{\mathrm{T}} = \Theta(\epsilon)$. In other words we assume that the squared observation error covariance matrix is of order ϵ.

By replacing Ω_R with $\epsilon^{-1}\Omega$, the l-EnKBF formula is written as

$$\mathrm{d}X^i_t = f(X^i_t)\mathrm{d}t + P^\dagger_t(X^i_t - \overline{X}_t)\mathrm{d}t - \frac{1}{2\epsilon}P^{\mathrm{L}}_t\Omega(X^i_t\mathrm{d}t + \overline{X}_t\,\mathrm{d}t - 2\mathrm{d}Y_t). \tag{11}$$

Since $\overline{X}_t = \frac{1}{M}\sum_{i=1}^{M} X_t^i$, the sample mean process follows the following dynamics

$$\mathrm{d}\overline{X}_t = \overline{f}_t \mathrm{d}t - \epsilon^{-1} P_t^{\mathrm{L}} \Omega(\overline{X}_t\, \mathrm{d}t - \mathrm{d}Y_t), \quad \overline{f}_t := \frac{1}{M}\sum_{i=1}^{M} f(X_t^i). \tag{12}$$

So if we denote $\Delta X_t^i = X_t^i - \overline{X}_t$, it follows the ordinary differential equation (ODE)

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta X_t^i = f(X_t^i) - \overline{f}_t + P_t^{\dagger}\Delta X_t^i - \frac{1}{2\epsilon}P_t^{\mathrm{L}}\Omega\Delta X_t^i.$$

Because the sample covariance $P_t = \frac{1}{M-1}\sum_{i=1}^{M}\Delta X_t^i(\Delta X_t^i)^{\mathrm{T}}$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}P_t = (F_t + F_t^{\mathrm{T}}) + (P_t^{\dagger}P_t + P_t P_t^{\dagger}) - \frac{1}{2\epsilon}(P_t^{\mathrm{L}}\Omega P_t + P_t\Omega P_t^{\mathrm{L}}) \tag{13}$$

where $F_t := \frac{1}{M-1}\sum(X_t^i - \overline{X}_t)(f(X_t^i) - \overline{f}_t)^{\mathrm{T}}$.

3. Main results

We present our main theoretical results for the l-EnKBF in (9) in this section. To keep the discussion concise, we allocate the technical verifications to the appendix.

3.1. Wellposedness and stability

Before the accuracy of the filter can be addressed it is crucial to check if the l-EnKBF can blow-up or collapse. In other words, we will demonstrate that the filter is stable, such that there are upper and lower bounds for P_t. The upper bound is established by the following:

Lemma 3.1. *Under assumptions 2.1 and 2.3, suppose P_t^{L} evolving in time according to (13) exists, the following holds*

$$\|P_t\|_{\max} \leqslant \lambda_{\max} := \frac{2\epsilon}{\omega_{\min}}\left(\sqrt{C_f^2 + \frac{3\omega_{\min}}{\epsilon}}\right), \quad \forall t > t'_* := \frac{\omega_{\min}\epsilon}{\lambda_{\max}}.$$

And for all $t > 0$, $\|P_t\|_{\max} \leqslant \max\{\|P_0\|_{\max}, \lambda_{\max}\}$. It is clear that when C_f and $\omega_{\min}$ are constants, $\lambda_{\max}(\epsilon) = \Theta(\sqrt{\epsilon})$, $t'_* = \Theta(\sqrt{\epsilon})$.

In [5] the bound depends explicitly on M (as the Frobenius norm is used to derive the bound). Here a different route is taken which results in a bound independent of M.

To ensure that the filter does not collapse, it is crucial to have a lower bound on the covariance. This comes as a reverse of lemma 3.1. For this purpose, we denote

$$\|P_t\|_{\min} = \min\{[P_t]_{i,i}, i = 1, \ldots, N_x\}.$$

It should be noted that $\|P_t\|_{\min}$ is not a norm, and we choose this notation just for its symmetry with $\|P_t\|_{\max}$.

Lemma 3.2. *Under assumptions 2.1 and 2.3, suppose P_t^{L} evolves in time according to (13), the following holds for sufficiently small $\epsilon > 0$*

$$\|P_t\|_{\min} \geqslant \lambda_{\min} := \frac{\epsilon}{3\lambda_{\max}\omega_{\max}C_\phi}, \quad \forall t > t_* := \frac{\omega_{\min}\epsilon}{\lambda_{\max}} + 3\lambda_{\min}.$$

$$\|P_t\|_{\min} \geqslant \min\left\{\|P_0\|_{\min}, \frac{\epsilon}{2\omega_{\min}\max\{\|P_0\|_{\max}, \lambda_{\max}\}}\right\} > 0, \quad \forall t > 0.$$

It is clear that when C_f and $\omega_{\min}$ are constants, $\lambda_{\min}(\epsilon) = \Theta(\sqrt{\epsilon}), t_* = \Theta(\sqrt{\epsilon})$.

Since $P_t^\dagger$ is well defined as long as $\|P_t\|_{\min} > 0$, using the same proof as in theorem 2.3 of [5], we can show that the l-EnKBF given by (9) has a strong solution:

Corollary 3.3. *Suppose the initial ensemble is selected so that $\|P_0\|_{\min} > 0$. Then the l-EnKBF filter is well defined for all $t > 0$.*

3.2. Error analysis in l_2 norm

As the next step we consider the accuracy of l-EnKBF in terms of the l_2 norm. Since the filter estimate with the ensemble mean, the error is its deviation from the truth, $e_t = X_t - \overline{X}_t$. While it has already been shown in [5] $\|e_t\|^2$ is of order $N_x\sqrt{\epsilon}$ through tail probability, our new result extends this estimate to the Laplace transforms. Moreover we show the path-wise maximum has the logarithm scaling with time, indicating the filter is highly stable in terms of error.

Theorem 3.4. *Let $e_t = X_t - \overline{X}_t$ be the filter error of l-EnKBF (11). Under assumptions 2.1 and 2.3, if $\tilde{\phi} := \phi - \rho I \succeq \mathbf{0}$ for a constant $\rho > 0$, then for any fixed $t_0 > 0$ there are strictly positive constants ϵ_0, c and C such that for every $\epsilon \in (0, \epsilon_0)$,*

(1) *When $t > t_0$,* $\quad \mathbb{E}\|e_t\|^2 \leqslant C\sqrt{\epsilon}N_x$.

(2) *For any $0 < \lambda < c\epsilon^{-1/2}$,*

$$\limsup_{t\to\infty} \mathbb{E}\, \exp(\lambda\|e_t\|^2) \leqslant 2\, \exp(4C\lambda N_x\sqrt{\epsilon}).$$

(3) *For any $T > t_0$, the following holds*

$$\mathbb{E}_{t_0}\left[\sup_{t_0\leqslant t\leqslant T}\|e_t\|^2\right] \leqslant \|e_{t_0}\|^2 + C\sqrt{\epsilon}N_x + C\sqrt{\epsilon}\, \log(CT/\sqrt{\epsilon})$$

Here $\mathbb{E}_{t_0}$ denotes conditional expectation with respect to information available at time t_0.

Note that the $\epsilon^{1/2}$ scaling is sharp. This can be understood best if one applies the Kalman–Bucy filter to (1) with $f(X) = \mathbf{0}$, $H = I_{N_x}$ and $R = \sqrt{\epsilon}I_{N_x}$, the posterior covariance P_t follows the ODE $\frac{\mathrm{d}}{\mathrm{d}t}P_t = 2I_{N_x} - \epsilon^{-1}P_t^2$. It is easy to show that P_t will converge to the limit $P_\infty = \sqrt{2\epsilon}I_{N_x}$, which is of order $\epsilon^{1/2}$ as well.

3.3. Analysis for component-wise error

While theorem 3.4 provides an estimate $\|e_t\|^2$, the estimate has a scaling of N_x because $\|e_t\|^2$ is the sum of N_x component errors. From theorem 3.4, it is impossible to indicate the error of one specific component, or whether this component's error is independent of the dimension N_x. This section shows that with a stronger structure assumption on the localisation matrix, we can derive dimension-independent bounds for each individual component.

Assumption 3.5. The localisation matrix ϕ is diagonally dominant. In other words, there is a $q < 1$ such that

$$\sum_{j\neq i} \phi_{i,j} \leqslant q.$$

Moreover, the interaction between components can be dominated by a constant $C_{\mathcal{F}}$-multiple of the matrix structure ϕ:

$$\mathcal{F}_{\mathbf{d}(i,j)} \leqslant C_{\mathcal{F}}\phi_{i,j} \quad \forall i, j.$$

Since $\mathcal{F}_d$ usually decays to zero quickly in practice, so $C_{\mathcal{F}}$ are likely to be found. Using lemma A.1 it is easy to show ϕ satisfying assumption 3.5 will have $\phi \succeq (1-q)I$, meaning $\tilde{\phi} = \phi - qI$ is positive semidefinite. In other words, assumption 3.5 is stronger than assumption for ϕ imposed in theorem 3.4. In general, ϕ is not always diagonally domain. However, this can hold if one choose small localisation length l. For example, for the Gaspari–Cohn [9] distance matrix ϕ, it will be diagonally dominant if $l \leqslant 1.4$. In other words, assumption 3.5 is likely to hold if the components of model represent spatial information of distant apart.

With assumption 3.5, we can reproduce theorem 3.4 type of result for individual component.

Theorem 3.6. *Let $e_t = X_t - \overline{X}_t$ be the filter error of l-EnKBF (11). Under assumptions 2.1, 2.3, and 3.5, for any fixed $t_0 > 0$ there are constants c and C such that for sufficiently small $\epsilon > 0$,*

(1) *When $t > t_0$, for any index i, $\mathbb{E}[[e_t]_i^2] \leqslant C\sqrt{\epsilon}$.*

(2) *For any $0 < \lambda < c\epsilon^{-1/2}$ and index i,*

$$\limsup_{t\to\infty} \mathbb{E}\, \exp(\lambda [e_t]_i^2) \leqslant 2\, \exp(4C\lambda\sqrt{\epsilon}).$$

(3) *For any $T > t_0$, the following holds for all i*

$$\mathbb{E}_{t_0}\left[\sup_{t_0\leqslant t\leqslant T} [e_t]_i^2\right] \leqslant \max_i\{[e_{t_0}]_i^2\} + C\sqrt{\epsilon}\, \log(T/\sqrt{\epsilon}).$$

Here $\mathbb{E}_{t_0}$ denotes conditional expectation with respect to information available at time t_0.

(4) *For any $T > t_0$,*

$$\mathbb{E}_{t_0}\left[\max_i \sup_{t_0\leqslant t\leqslant T} [e_t]_i^2\right] \leqslant \max_i\{[e_{t_0}]_i^2\} + C\sqrt{\epsilon}\, \log(N_x T/\sqrt{\epsilon}).$$

Remark. If $Z_1, \ldots, Z_n$ are i.i.d. samples of a Gaussian distribution, a rough estimate of $\max_i\{Z_i\}$ is of order log$n$. And when system has short range interaction, its components are tend to be independent when they are far apart. Likewise, when a system is stationary, it is close to independent with it self in a distance past. The filter error process happens to have both of these two properties. That is why we have the scaling of $\log(N_x T)$ in claim (4).

4. Numerical investigation

Lastly the theoretical findings are numerically verified by means of the stochastically perturbed Lorenz 96 system (L96) [20]. The evolution of each spatial component is given by

$$\mathrm{d}x_s(t) = f_s(X(t))\mathrm{d}t + \sqrt{2}\mathrm{d}W_s(t)$$

for $s \in \{1, \dots, N_x\}$. Here

$$f_s(X(t)) = \left(x_{s+1}(t) - x_{s-2}(t)\right) x_{s-1}(t) - x_s(t) + 8, \tag{14}$$

and spatial periodicity is assumed, i.e., $x_{-1}(t) = x_{N_x-1}(t)$, $x_0(t) = x_N(t)$ and $x_{N_x+1}(t) = x_1(t)$. Numerically generated trajectories of (14) are typically bounded in the l_∞ norm, i.e.,

$$|x_s(t)| \leqslant C = 40 \tag{15}$$

for all s for the Lorenz 96 system. In other words, the solution of (14) is largely indifferent from a soft-truncated version $\mathrm{d}X_s(t) = \hat{f}_s(X(t))\mathrm{d}t + \sqrt{2}\mathrm{d}W_s(t)$, where

$$\tilde{f}(x_s(t)) = 1_{\|X(t)\|_\infty \leqslant C} \left(x_{s+1}(t) - x_{s-2}(t)\right) x_{s-1}(t) - x_s(t). \tag{16}$$

Then note that when $\|X(t)\|_\infty \geqslant C$, $|\tilde{f}_s(X(t)) - \tilde{f}_s(X'(t))| = |x'_s(t) - x_s(t)|$; when $\|X(t)\|_\infty \leqslant C$,

$$\begin{aligned}
|\tilde{f}_s(X(t)) - \tilde{f}_s(X'(t))| &= |\left(x_{s+1}(t) - x_{s-2}(t)\right) x_{s-1}(t) - x_s(t) - [\left(x'_{s+1}(t) - x'_{s-2}(t)\right) x'_{s-1}(t) - x'_s(t)]| \\
&\leqslant |(x_{s+1}(t) - x'_{s+1}(t))x_{s-1}(t)| + |(x_{s-1}(t) - x'_{s-1}(t))x'_{s+1}(t)| \\
&\quad + |(x'_{s-2}(t) - x_{s-2}(t))x'_{s-1}(t)| + |(x'_{s-1}(t) - x_{s-1}(t))x_{s-2}(t)| + |x'_s(t) - x_s(t)| \\
&\leqslant C|x_{s+1}(t) - x'_{s+1}(t)| + C|x_{s-1}(t) - x'_{s-1}(t)| + C|x'_{s-2}(t) - x_{s-2}(t)| \\
&\quad + C|x'_{s-1}(t) - x_{s-1}(t)| + |x'_s(t) - x_s(t)|.
\end{aligned}$$

Therefore, assumption 2.1 is fulfilled with $\mathcal{F}_{\mathbf{d}(i,j)} = 0$ for $\mathbf{d}(i,j) > 2$ where $\mathbf{d}(i,j) = \min\{|i - j|, |i + n - j|, |j + n - i|\}$, and (16) has only short range interactions. While we will only simulate (14) in below, we expect the associated filter behaviour will be similar to the one in (16). Further the entries of the localisation matrix ϕ are set to

$$\phi_{i,j} = \rho\left(\frac{\mathbf{d}(i,j)}{l}\right)$$

using the Gaspari–Cohn function (5) for ρ and setting the localisation radius to $l = 1.4$. Note that this choice of localisation radius ensures that ϕ is diagonally dominant, i.e., assumption 3.5 is fulfilled. It is important to note that this choice is not necessarily the *optimal*[5] value for the considered system yet the chosen value is sufficient to obtain reasonable MSE values of the expected order. Further we choose the model noise variance to be $\sigma = 1$ and the observation operator H to be the identity matrix which is in line with assumption 2.3. Three test scenarios are considered to numerically verify the sensitivity of the l-EnKBF with respect to the dimension N_x, time interval size T and the measurement error ϵ.

[5] Here optimality can for example be associated with the lowest MSE.

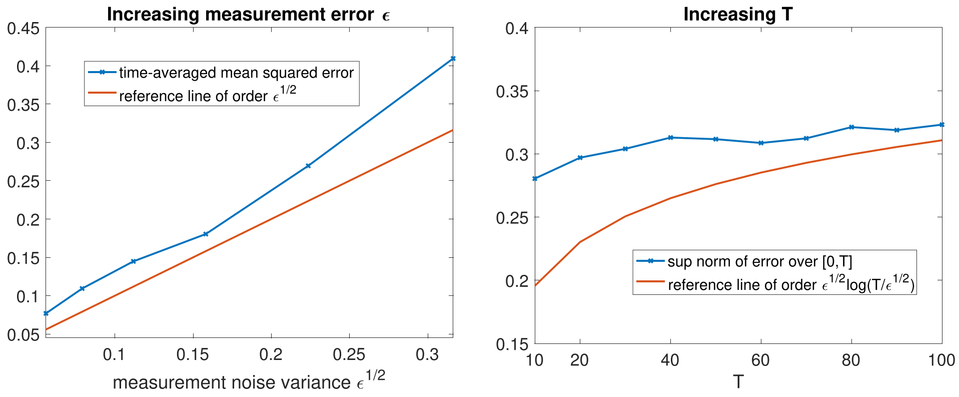

Figure 1. Time-averaged MSE as a function of the measurement error variance ϵ is displayed in the left panel. The right panel displays the estimated $\sup_{t\in[0,T]} e_t$ for varying T.

4.1. Sensitivity with respect to ϵ

In the first test scheme, the expected filtering error is approximated via a time-averaged MSE for different measurement error values

$$\epsilon \in \{0.003\,125, 0.006\,25, 0.025, 0.05, 0.1\}.$$

In order to emulate a continuous setting the steps size is chosen to be $\mathrm{d}t = 10^{-7}$ and the number of steps 10^7. The dimension of the state space is set to be $N_x = 40$, which is a standard choice of the Lorenz 96 model. The l-EnKBF is implemented with $M = 10$ ensemble members. The results are displayed in the left panel of figure 1. Note that the MSE is normalised with respect to the dimension, i.e., is divided by N_x. The test run confirms that the numerical growth rate with respect to an increasing ϵ is in line with theoretical order of the expected error derived in claim (1) of theorem 3.4.

4.2. High dimensional testcase

In the second test scheme, the robustness with respect to state space dimension is investigated. In particular we consider the case where the number of ensemble members M is comparatively small and kept fixed for increasing dimensions. Thus the imbalance between ensemble size and dimension of the state space grows with increasing N_x. More precisely we run the filter for $N_x \in \{40, 240, 440, 640, 840, 1040\}$ with $M = 10$ and $\epsilon = 0.003\,125$. The resulting time-averaged MSE after 10^6 steps with step size $\mathrm{d}t = 10^{-7}$ are displayed in the left panel of figure 2. As state in claim (1) of theorem 3.4 the error grows linearly with N_x. Further we numerically verify that the time-averaged error of the individual components, i.e.,

$$\frac{1}{T}\sum_{t=1}^{T} [e_t]_i^2(t) \tag{17}$$

are dimension independent (see right panel of figure 2) as stated in claim (1) of theorem 3.6. Note that we fixed the considered component of the state vector to be $i = 11$ while other index choice produces largely the same results.

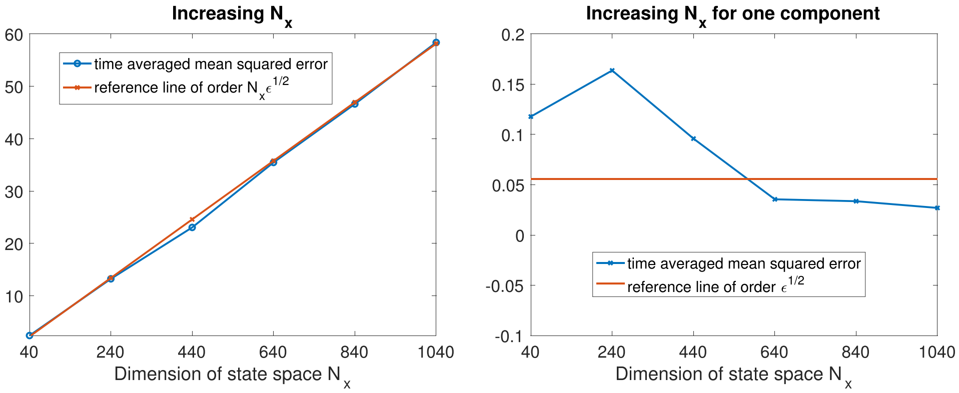

Figure 2. Time-averaged MSE as a function of the state dimension N_x is displayed in the left panel whereas the time-averaged MSE for one fixed component for increasing N_x is shown in the right panel.

4.3. Uniform error for bounded time interval

In the final test scheme, we consider a setting with a growing number of steps 10^6 to 10^7 for a fixed step size $\mathrm{d}t = 10^{-5}$ resulting in filter runs for different time values $T \in \{10, \ldots, 100\}$. Note that the step size is set to be slightly larger than in the previous examples so that the range of considered T values is more interesting. Further the measurement error variance is set to $\epsilon = 0.01$ and the dimension of the state space is $N_x = 40$. We simulate the filtering process 30 times and record the filter error $\mathrm{e}^j(t), j = 1, \ldots, 30$, for each simulation. We plot the averaged path-wise l_2-square error up to T, which is

$$\frac{1}{30}\sum_{j=1}^{30} \max_{t\in[0,T]} \|\mathrm{e}^j(t)\|^2$$

in the right panel of figure 1. The dominating part[6] $C\sqrt{\epsilon}\ \log(CT/\sqrt{\epsilon})$ of the theoretical order of claim (3) of theorem 3.4 is plotted as a reference slope.

Note that the numerically obtain error is in line with the theoretical order and thus is verifying the logarithmic dependence of the uniform bound on time T.

5. Conclusion

In this paper, the earlier derived stability and accuracy results for the EnKBF are extended for systems with $N_x \gg M$ via localisation. Further the upper bound for the covariance is independent of the number of ensemble members M and the derived path-wise bounds have a better scaling with respect to the time T. Moreover it is shown that the accuracy in the individual components is independent of the state dimension N_x and a Laplace type condition is obtained. Natural extensions include partially observed processes and misspecified drift functions $f(x_t, \lambda)$ with unknown parameter λ. Moreover the presented ideas can be used for the analysis of properties of multilevel ensemble Kalman filters [10, 11] or of consistent filters,

[6] For the considered N_x, ϵ and T, the other dominating component $C\sqrt{\epsilon}N_x$ is not large but can of course become significant for $N_x \gg 0$.

such as the ensemble transform particle filter [25] or the feedback particle filter [31], for finite number of ensemble members.

Acknowledgments

The research of JdW is funded by Deutsche Forschungsgemeinschaft (DFG) - SFB1294/1 - 318763901 Project (A02) 'Long-time stability and accuracy of ensemble transform filter algorithms'. JdW was also supported by ERC Advanced Grant ACRCC (Grant 339390) and by the Simons CRM Scholar-in-Residence Program. The research of XTT is funded by Singapore MOE AcRF Tier 1 funding, grant number R-146-000-292-114. Further the authors would like to thank the anonymous referees for their constructive criticism.

Appendix A. Proof for filter wellposedness and stability

A.1. Matrix norms and Riccati equation

To start, we have several norm inequalities which are utilised in this paper.

Lemma A.1. *For any $N \times N$ matrix A, the following holds*

$$\|A\|_{\max} \leqslant \|A\|, \tag{18}$$

$$\|A\| \leqslant \sqrt{\|A\|_1 \|A^{\mathrm{T}}\|_1}. \tag{19}$$

Proof. Inequality (18) follows via

$$\|A\|_{\max} = \max_{i,j} |[A]_{i,j}| = \max_{i,j} |[e_t]_i^{\mathrm{T}} A e_j| \leqslant \|A\|,$$

where $[e_t]_i$ and e_j are the ith and jth standard Euclidean basis vector. Inequality (19) follows from [23] lemma B.2. □

Lemma A.2. *Let P, Q and ϕ be positive, symmetric and semidefinite $N_x \times N_x$ matrices and $[\phi]_{i,i} = 1$ for all i. Then*

(1) *For all i, $[(P \circ \phi)Q]_{i,i} = [P(Q \circ \phi)]_{i,i}$.*
(2) *If $P \preceq Q$, then $P \circ \phi \preceq Q \circ \phi$.*
(3) $\|P \circ \phi\|_{\max} = \|P\|_{\max} = \max_i\{[P]_{i,i}\}$
(4) $\|P \circ \phi\| \leqslant \|P \circ \phi\|_1 \leqslant C_\phi \|P\|_{\max}$, *where $C_\phi = \max_i \sum_j |\phi_{i,j}|$.*

Proof. claim (1). Just note that

$$[(P \circ \phi)Q]_{i,i} = \sum_k P_{i,k}\phi_{i,k}Q_{k,i} = \sum_k P_{i,k}\phi_{k,i}Q_{k,i} = [P(Q \circ \phi)]_{i,i}.$$

□

Proof. claim (2). Due to the linearity of the Schur product, it suffices to show that $0 \preceq P \circ \phi$. This is known as the Schur product theorem, which can be verified using the following identity, which holds for all N_x-dimensional vectors u, with D_u being the diagonal matrix where its diagonal entries are the same as u:

$$u^{\mathrm{T}}(P\circ\phi)u = \mathrm{tr}(PD_u\phi D_u) = \mathrm{tr}(P^{\frac{1}{2}}D_u\phi^{\frac{1}{2}}\phi^{\frac{1}{2}}D_uP^{\frac{1}{2}}) \geqslant 0.$$

□

Proof. claim (3). Since P is a positive semidefinite matrix for each i and j, it follows that

$$\langle [e_t]_i - e_j, P([e_t]_i - e_j)\rangle = \langle [e_t]_i, P[e_t]_i\rangle - 2\langle [e_t]_i, Pe_j\rangle + \langle e_j, Pe_j\rangle \geqslant 0,$$

where $[e_t]_i$ and e_j are the i and jth standard Euclidean basis vector. This implies

$$2\langle [e_t]_i, Pe_j\rangle \leqslant \langle [e_t]_i, P[e_t]_i\rangle + \langle e_j, Pe_j\rangle \leqslant 2\max_k [P]_{k,k}.$$

In other words in a positive semidefinite matrix the maximal values are reached on the diagonal. Note that the Schur product $P\circ\phi$ is a positive semidefinite as well, so it is maximal matrix entries are also assumed on the diagonal. Since ϕ is set to $[\phi]_{i,i} = 1$ for all i the Schur product does not alter the diagonal entries of P thus $\|P\circ\phi\|_{\max} = \|P\|_{\max}$.

□

Proof. claim (4). Recall that inequality (19) implies $\|A\| \leqslant \|A\|_1$ for any symmetric matrix A which yields the first half of claim (4), since $P\circ\phi$ is symmetric. The other half can be obtained by

$$\|P\circ\phi\| \leqslant \|P\circ\phi\|_1 = \max_i \left|\sum_j \phi_{i,j}P_{i,j}\right| \leqslant \max_i \sum_j |\phi_{i,j}|\|P\|_{\max} = C_\phi\|P\|_{\max}.$$

□

In this paper, we often concern Riccati type of stochastic equation. In particular, we are often interested in finding bounds for the maximum entry of the solution. To do so, we employ a comparison principle, which generates bounds by comparing with another ODE. In particular, we have the following lemma.

Lemma A.3. *Suppose $X_t = [x_1(t), \ldots, x_n(t)]$ jointly follows an ODE, $\frac{\mathrm{d}}{\mathrm{d}t}X_t = F(X_t)$. Let $m_t = \max_{1\leqslant i\leqslant n'}\{x_i(t)\}$, where n' can be smaller than n. Let i_t be the smallest index i such that $x_i(t) = m_t$. Suppose there is a continuous function $g(x,t)$ such that for any $t \geqslant 0$,*

$$\frac{\mathrm{d}}{\mathrm{d}t}x_{i_t}(t) \leqslant g(x_{i_t}(t), t).$$

Suppose y_t satisfies $\frac{\mathrm{d}}{\mathrm{d}t}y_t = g(y_t, t) + \delta_0$ for a fixed $\delta_0 > 0$ and $y_0 > m_0$, then the following hold

(1) *For all $t > 0$, $y_t > m_t$.*

(2) *Suppose $g(x,t) = g(x) = -\frac{c}{\epsilon}x^2 + bx + a - \delta_0$, where a, b, c are constants. Let*

$$\Delta_\epsilon := 2\sqrt{\frac{b^2\epsilon^2}{4c^2} + \frac{a\epsilon}{c}} = \Theta(\sqrt{\epsilon}).$$

If $y_0 > 0$, then $y_t \leqslant \max\{\Delta_\epsilon, y_0\}$ for all $t > 0$. Moreover, when $t > t_* = \frac{c\epsilon}{\Delta_\epsilon} = \Theta(\sqrt{\epsilon})$, $y_t \leqslant \Delta_\epsilon$.

(3) *Suppose z_t is a process such that $0 < z_t < D$ for all $t > 0$, and $z_t \leqslant \Delta_\epsilon$ for $t > t_*$, where Δ_ϵ, t_* are positive quantities of order $\Theta(\sqrt{\epsilon})$. Suppose $g(x,t) = \alpha\sqrt{-xz_t} - \frac{\beta}{\epsilon}xz_t - \gamma - \delta_0$, where α, β, γ are all positive constants. Then if $y_0 < 0$,*

$$y_t \leqslant -\min\left\{|y_0|, \frac{\gamma^2}{\alpha^2 D}, \frac{\gamma\epsilon}{2\beta D}\right\}, \quad \forall t > 0.$$

Moreover $|y_t| > c_\epsilon$ for all $t > t_* + \frac{3c_\epsilon}{\gamma} = \Theta(\sqrt{\epsilon})$, where

$$c_\epsilon := \min\left\{\frac{\gamma^2}{9\alpha^2\Delta_\epsilon}, \frac{\gamma\epsilon}{3\beta\Delta_\epsilon}\right\} = \Theta(\sqrt{\epsilon}).$$

Proof. claim (1). Let $t_1 = \inf\{t > 0, y_t \leqslant m_t\}$. By continuity of m_t and y_t, $t_1 > 0$. Suppose t_1 is finite, then $y_{t_1} = m_{t_1}$. Therefore

$$\frac{\mathrm{d}}{\mathrm{d}t}x_{i_{t_1}}(t_1) \leqslant g(x_{i_{t_1}}(t), t_1) = g(y_{t_1}, t_1) = \frac{\mathrm{d}}{\mathrm{d}t}y(t_1) - \delta_0.$$

This indicate for sufficiently small $\delta > 0$,

$$x_{i_{t_1}}(t_1 - \delta) > x_{i_{t_1}}(t_1) - \delta g(x_{i_{t_1}}(t), t_1) - \frac{1}{2}\delta\delta_0 > y(t_1) - \delta g(y(t_1), t_1) + \frac{1}{2}\delta\delta_0 > y(t_1 - \delta).$$

This contradicts with the definition of t_1. Therefore $t_1 = \infty$. □

Proof. claim (2). First we denote the root of $g(x,t) + \delta_0 = 0$ as

$$y_\pm = -\frac{b\epsilon}{2c} \pm \sqrt{\frac{b^2\epsilon^2}{4c^2} + \frac{a\epsilon}{c}}.$$

It is easy to check that $y_+ > 0 > y_-$, while $\Delta_\epsilon = y_+ - y_-$. Note that $g(x) + \delta_0 \leqslant 0$ when $y_t \geqslant \Delta_\epsilon \geqslant y_+$. So y_t is decreasing when y_t is above Δ_ϵ.

Next note that y_t is the solution of a Riccati differential equation. The solution to the Riccati ODE is given by has the explicit formulation

$$\frac{y_t - y_-}{y_t - y_+} = \exp\left(\frac{c}{\epsilon}t(y_+ - y_-)\right)\left(\frac{y_0 - y_-}{y_0 - y_+}\right) \tag{20}$$

If $y_0 < y_+$, it is easy to check that (20) always take negative value, meaning $y_t < y_+ < y_+ - y_- = \Delta_\epsilon$ for all $t > 0$. When $y_0 > y_+$ and $t > t_*$, from (20) leads to

$$\frac{y_t - y_-}{y_t - y_+} \geqslant \exp\left(\frac{c}{\epsilon}t(y_+ - y_-)\right) \geqslant 2,$$

so $y_t \leqslant y_+ - y_- = 2\sqrt{\frac{b^2\epsilon^2}{4c^2} + \frac{a\epsilon}{c}} = \Theta(\sqrt{\epsilon})$. □

Proof. claim (3). Note that $g(x,t)+\delta_0<0$ when $|x|<\min\left\{\frac{\gamma^2}{\alpha^2 D},\frac{\gamma\epsilon}{2\beta D}\right\}$, so y_t will be decreasing if it is above $-\min\left\{\frac{\gamma^2}{\alpha^2 D},\frac{\gamma\epsilon}{2\beta D}\right\}$. This leads to the first part of the claim.

Next note that when $0\geqslant x\geqslant -c_\epsilon$ and $t>t_*$, $g(x,t)\leqslant -\frac{\gamma}{3}$. So y_t will be decreasing with rate at least $\frac{\gamma}{3}$ when $0\geqslant y_t\geqslant -c_\epsilon$ and $t>t_*$, this leads to our claim.

□

A.2. Upper bounds for sample covariance

Proof of lemma 3.1. Recall that P_t is positive semidefinite, therefore by lemma A.2

$$\|P_t\|_{\max}=\max\{[P_t]_{i,i}, i=1,\ldots,N_x\},$$

where the components of P_t follows an ODE (13). Therefore, in order to apply lemma A.3 claim (1), it suffices to investigate the ODE deriving the component with the maximal value. Suppose at time t, $[P_t]_{k,k}=\|P_t\|_{\max}$ for certain k. Considering the time evolution of $[P_t^{\mathrm{L}}]_{k,k}$ given by (13), it is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}[P_t]_{k,k}=[F_t+F_t^{\mathrm{T}}]_{k,k}+[P_t^{\dagger}P_t+P_tP_t^{\dagger}]_{k,k}-\frac{1}{2\epsilon}[P_t^{\mathrm{L}}\Omega P_t+P_t\Omega P_t^{\mathrm{L}}]_{k,k}. \qquad (21)$$

First note that $[F_t]_{k,k}=\frac{1}{M-1}\sum_i(x_k^i-\overline{x}_k)(f_k(X_t^i)-\overline{f}_k)$, where by assumption 2.1 we have

$$|f_k(X^i)-\overline{f}_k|\leqslant\frac{1}{M}\sum_{j=1}^{M}|f_k(X_t^i)-f_k(X_t^j)|\leqslant\frac{1}{M}\sum_{j=1}^{M}\sum_{l=1}^{N_x}\mathcal{F}_{\mathbf{d}(k,l)}|x_l^i-x_l^j|.$$

This leads to

$$\begin{aligned}[F_t]_{k,k}&\leqslant\frac{1}{M-1}\sum_{i=1}^{M}|x_k^i-\overline{x}_k||f_k(X_t^i)-\overline{f}_k|\\&\leqslant\frac{1}{(M-1)M}\sum_{i,j,l,m}\mathcal{F}_{\mathbf{d}(k,l)}|x_k^i-x_k^m||x_l^i-x_l^j|\\&\leqslant\sum_{l=1}^{N_x}\mathcal{F}_{\mathbf{d}(k,l)}\sqrt{\frac{\sum_{i,m}|x_k^i-x_k^m|^2}{M(M-1)}}\sqrt{\frac{\sum_{i,j}|x_l^i-x_l^j|^2}{M(M-1)}}\\&\leqslant\sum_{l=1}^{N_x}\mathcal{F}_{\mathbf{d}(k,l)}[P_t]_{k,k}\leqslant C_f[P_t]_{k,k}.\end{aligned}\qquad (22)$$

Also, note that $[P_t^{\dagger}P_t]_{k,k}=[P_t^{\dagger}]_{k,k}[P_t]_{k,k}=1$ due to definition 2.2, so

$$[P_t^{\dagger}P_t]_{k,k}=1. \qquad (23)$$

Lastly, we have

$$[P_t^{\mathrm{L}}\Omega P_t]_{k,k}=\sum_{i=1}^{N_x}[P_t^{L}]_{k,i}\Omega_{i,i}[P_t]_{i,k}=\sum_{i=1}^{N_x}[P_t]_{k,i}^2\phi_{i,k}\Omega_{i,i}\geqslant\Omega_{k,k}[P_t]_{k,k}^2\geqslant\omega_{\min}[P_t]_{k,k}^2. \qquad (24)$$

Insert (22)–(24) to (21), we find

$$\frac{\mathrm{d}}{\mathrm{d}t}[P_t]_{k,k} \leqslant 2C_f[P_t]_{k,k} + 2 - \frac{\omega_{\min}}{\epsilon}[P_t]_{k,k}^2.$$

Therefore lemma A.3 claim (1) applies with

$$g(x,t) = 2C_f x + 2 - \frac{\omega_{\min}}{\epsilon}x^2.$$

Let $\delta_0 = 1$, lemma A.3 claim (2) yields the result of this lemma. □

A.3. Lower bounds for sample covariance

Proof of lemma 3.2. The proof is similar to the one of lemma 3.1, but we need to change sign, because

$$-\|P_t\|_{\min} = \max\{-[P_t]_{i,i}, i = 1, \ldots, N_x\}.$$

By lemma A.3 claim (1), we assume at time t, $\|P_t\|_{\min} = [P_t]_{k,k}$ and investigate the ODE that $-[P_t]_{k,k}$ follows. It is given by the inverse of (21). Following same procedures prior to (22), we have

$$[F_t]_{k,k} \geqslant -\sum_{l=1}^{N_x} \mathcal{F}_{\mathbf{d}(k,l)} \sqrt{\frac{\sum_{i,m}|x_k^i - x_k^m|^2}{M(M-1)}} \sqrt{\frac{\sum_{i,j}|x_l^i - x_l^j|^2}{M(M-1)}} \geqslant -C_f\sqrt{[P_t]_{k,k}\|P_t\|_{\max}} \quad (25)$$

(23) remains the same. Finally recall that in (24), we have

$$[P_t^{\mathrm{L}}\Omega P_t]_{k,k} = \sum_{i=1}^{N_x}[P_t]_{k,i}^2\phi_{i,k}\Omega_{i,i} \leqslant \sum_{i=1}^{N_x}\omega_{\max}\phi_{i,k}[P_t]_{k,k}\|P_t\|_{\max} \leqslant \omega_{\max}C_\phi[P_t]_{k,k}\|P_t\|_{\max}. \quad (26)$$

Insert (23), (25) and (26) into (21), we find

$$-\frac{\mathrm{d}}{\mathrm{d}t}[P_t]_{k,k} \leqslant 2C_f\sqrt{-(-[P_t]_{k,k})\|P_t\|_{\max}} + \frac{\omega_{\max}}{\epsilon}C_\phi\|P_t\|_{\max}[P_t]_{k,k} - 2.$$

So we can apply lemma A.3 claim (3) with $\delta_0 = 1$ and

$$g(x,t) = 2C_f\sqrt{-x\|P_t\|_{\max}} + \frac{\omega_{\max}}{\epsilon}C_\phi x[P_t]_{k,k} - 2.$$

This gives us the claimed result. □

Appendix B. Proof for filter error analysis in l_2 norm

B.1. Evolution of component-wise error

Before we prove the statements of theorem 3.4 we consider the following auxiliary lemma which will be used several times throughout the remainder of the paper.

Lemma B.1. *Let $[e_t]_j$ be the jth component of the filter error e_t. Then the following holds*

$$\mathrm{d}[e_t]_j^2 \leqslant \left(-\alpha_t[e_t]_j^2 - 2\epsilon^{-1}\sum_{i=1}^{j}[P_t \circ \tilde{\phi}]_{j,i}[e_t]_i[e_t]_j + \sum_{i\neq j}\mathcal{F}_{\mathbf{d}(i,j)}|[e_t]_i|^2 + \beta_t\right)\mathrm{d}t + \mathrm{d}[\mathcal{M}_t]_j. \tag{27}$$

In (27), *$\mathcal{M}_t$ is a N_x dimensional martingale with components being*

$$\mathrm{d}[\mathcal{M}_t]_j = 2\sqrt{2}[e_t]_j\mathrm{d}W_j - 2\epsilon^{-1/2}[e_t]_j[P_t^L\Omega^{1/2}\mathrm{d}B]_j.$$

In (27), *α_t and β_t are two real valued processes given by*

$$\alpha_t := 2\epsilon^{-1}\rho\|P_t\|_{\min} - C_f - 1,$$
$$\beta_t := C_f^2\|P_t\|_{\max} + 2 + \epsilon^{-1}C_\phi^2\omega_{\max}\|P_t\|_{\max}^2.$$

By lemmas 3.1 *and* 3.2, *the following holds for all $t \geqslant 0$*

$$\alpha_t \geqslant \alpha^* = -C_f - 1,$$

$$\beta_t \leqslant \beta^* := C_f^2\max\{\|P_0\|_{\max}, \lambda_{\max}\} + 2 + \epsilon^{-1}C_\phi^2\omega_{\max}\max\{\|P_0\|_{\max}^2, \lambda_{\max}^2\} = \Theta(\epsilon^{-1}).$$

When $t \geqslant t_*$, these bounds can be further improved to

$$\alpha_t \geqslant \alpha_* := 2\epsilon^{-1}\rho\lambda_{\min} - C_f - 1 = \mathcal{O}(\epsilon^{-\frac{1}{2}}),$$
$$\beta_t \leqslant \beta_* := C_f^2\lambda_{\max} + 2 + \epsilon^{-1}C_\phi^2\omega_{\max}\lambda_{\max}^2 = \Theta(1).$$

Proof. Recall the evolution of X_t and $\overline{X}_t$ are given by $\mathrm{d}X_t = f(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$ and

$$\mathrm{d}\overline{X}_t = \overline{f}_t\mathrm{d}t - \epsilon^{-1}P_t^{\mathrm{L}}\Omega(\overline{X}_t\,\mathrm{d}t - \mathrm{d}Y_t) = \overline{f}_t\mathrm{d}t - \epsilon^{-1}P_t^{\mathrm{L}}\Omega(\overline{X}_t\,\mathrm{d}t - X_t\,\mathrm{d}t - \sqrt{\epsilon}\Omega^{-1/2}\mathrm{d}B_t).$$

The evolution of the error $e_t = X_t - \overline{X}_t$ is given by the difference between the two, namely

$$\mathrm{d}e_t = (f(X_t) - \overline{f}_t - \epsilon^{-1}P_t^{\mathrm{L}}\Omega e_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t - \epsilon^{-1/2}P_t^{\mathrm{L}}\Omega^{1/2}\mathrm{d}B_t.$$

The jth component of this differential equation is given by

$$\mathrm{d}[e_t]_j = (f_j(X_t) - \overline{f}_j - \epsilon^{-1}[P_t^{\mathrm{L}}\Omega e_t]_j)\mathrm{d}t + \sqrt{2}\mathrm{d}W_j - \epsilon^{-1/2}[P_t^{\mathrm{L}}\Omega^{1/2}\mathrm{d}B]_j,$$

where $\overline{f}_j$ denotes the jth component of $\overline{f}_t$. Ito's formula implies that

$$\mathrm{d}[e_t]_j^2 = \big(2(f_j(X_t) - \overline{f}_j - \epsilon^{-1}[P_t^{\mathrm{L}}\Omega e_t]_j)[e_t]_j + 2 + \epsilon^{-1}[P_t^{\mathrm{L}}\Omega P_t^{\mathrm{L}}]_{jj}\big)\,\mathrm{d}t$$
$$+ 2\sqrt{2}[e_t]_j \mathrm{d}W_j - \epsilon^{-1/2} 2[e_t]_j [P_t^{\mathrm{L}}\Omega^{1/2}\mathrm{d}B]_j. \tag{28}$$

To continue, note that

$$\begin{aligned} |f_j(X_t) - \overline{f}_j||[e_t]_j| &= \left| f_j(X_t) - \frac{1}{M}\sum_{i=1}^{M} f_j(X_t^i)\right| \, |[e_t]_j| \\ &\leqslant \frac{1}{M}\sum_{i=1}^{M} |f_j(X_t) - f_j(X_t^i)||[e_t]_j| \\ &\leqslant \frac{1}{M}\sum_{i=1}^{M} |f_j(\overline{X}_t) - f_j(X_t^i)||[e_t]_j| + |f_j(X_t) - f_j(\overline{X}_t)||[e_t]_j|. \end{aligned} \tag{29}$$

By assumption 2.1, the second part of (29) can be bounded easily by

$$\begin{aligned} |f_j(X_t) - f_j(\overline{X}_t)||[e_t]_j| &\leqslant \sum_{i=1}^{N_x} \mathcal{F}_{\mathbf{d}(i,j)}|[X_t - \overline{X}_t]_i||[e_t]_j| = \sum_{i=1}^{N_x} \mathcal{F}_{\mathbf{d}(i,j)}|[e_t]_i||[e_t]_j| \\ &\leqslant \frac{1}{2}C_f|[e_t]_j|^2 + \frac{1}{2}\sum_{i\neq j}\mathcal{F}_{\mathbf{d}(i,j)}|[e_t]_i|^2. \end{aligned} \tag{30}$$

To bound the first part of (29), we note by assumption 2.1 and Cauchy Schwarz,

$$\begin{aligned} \frac{1}{M}\sum_{i=1}^{M} |f_j(\overline{X}_t) - f_j(X_t^i)| &\leqslant \sum_{i,k=1}^{M,N_x} \frac{\mathcal{F}_{\mathbf{d}(k,j)}}{M}|[X_t^i - \overline{X}_t]_k| \\ &\leqslant \sqrt{\sum_{k=1}^{N_x}\frac{\mathcal{F}_{\mathbf{d}(k,j)}}{M^2}\sum_{i=1}^{M}|[X_t^i - \overline{X}_t]_k|^2} \\ &\leqslant \sqrt{\frac{1}{M}C_f^2(M-1)[P_t]_{k,k}} \leqslant C_f\|P_t\|_{\max}^{\frac{1}{2}}. \end{aligned} \tag{31}$$

Then multiplication with $|[e_t]_j|$ with (31) yields

$$\begin{aligned} \frac{1}{M}\sum_{i=1}^{M} |f_j(\overline{X}_t) - f_j(X_t^i)||[e_t]_j| &\leqslant \frac{1}{2}(\frac{1}{M}\sum_{i=1}^{M} |f_j(\overline{X}_t) - f_j(X_t^i)|)^2 + \frac{1}{2}|[e_t]_j|^2 \\ &\leqslant \frac{1}{2}C_f^2\|P_t\|_{\max} + \frac{1}{2}|[e_t]_j|^2. \end{aligned}$$

Plug these into (29), we find

$$|f_j(X_t) - \overline{f}_j||[e_t]_j| \leqslant \frac{1}{2}C_f^2\|P_t\|_{\max} + \frac{1}{2}|[e_t]_j|^2 + \frac{1}{2}C_f|[e_t]_j|^2 + \frac{1}{2}\sum_{i\neq j}\mathcal{F}_{\mathbf{d}(i,j)}|[e_t]_i|^2. \tag{32}$$

Next, we deal with $[P_t^{\mathrm{L}}\Omega e_t]_j$ in (28). Define $\tilde{\phi} := \phi - \rho I$ and obtain the following equality

$$[P_t^{\mathrm{L}}\Omega e_t]_j = \sum_{i=1}^{N_x}[P_t \circ \phi]_{j,i}\Omega_{i,i}[e_t]_i = \rho[P_t]_{j,j}\Omega_{j,j}[e_t]_j + \sum_{i=1}^{N_x}[P_t \circ \tilde{\phi}]_{j,i}\Omega_{i,i}[e_t]_i. \quad (33)$$

Also note that

$$[P_t^{\mathrm{L}}\Omega P_t^{\mathrm{L}}]_{j,j} = \sum_i \Omega_{i,i}\phi_{i,j}^2[P_t]_{i,j}^2 \leqslant \|P_t\|_{\max}^2\omega_{\max}\sum_i \phi_{i,j}^2 \leqslant C_\phi^2\omega_{\max}\|P_t\|_{\max}^2. \quad (34)$$

Plug (32)–(34) into (28), we obtain

$$\begin{aligned}\mathrm{d}[e_t]_j^2 \leqslant \Bigg(&(1 + C_f - 2\epsilon^{-1}\rho\|P_t\|_{\min})[e_t]_j^2 - 2\epsilon^{-1}\sum_{i=1}^{N_x}[P_t \circ \tilde{\phi}]_{j,i}[e_t]_i[e_t]_j \\ &+ \sum_{i\neq j}\mathcal{F}_{\mathbf{d}(i,j)}|[e_t]_i|^2 + C_f^2\|P_t\|_{\max} + 2 + \epsilon^{-1}C_\phi^2\omega_{\max}\|P_t\|_{\max}^2\Bigg)\mathrm{d}t \\ &+ 2\sqrt{2}[e_t]_j\mathrm{d}W_j - 2\epsilon^{-1/2}[e_t]_j[P_t^{\mathrm{L}}\Omega^{1/2}\mathrm{d}B]_j. \end{aligned} \quad (35)$$

□

B.2. Two technical lemmas

Lemma B.2 (Grönwall's inequality). *Suppose a real value process u_t satisfies the following for $t \geqslant t_0$ and constants α and β:*

$$\mathrm{d}u_t \leqslant (-\alpha u_t + \beta)\mathrm{d}t + \mathrm{d}M_t$$

for some martingale M_t. It follows that for any $t \geqslant t_0$

$$\mathbb{E}_{t_0}u_t \leqslant u_{t_0}\,\exp(-\alpha(t-t_0)) + \frac{\beta}{\alpha}(1 - \exp(-\alpha(t-t_0))).$$

When α and β are both positive, we have further that

$$\mathbb{E}_{t_0}u_t \leqslant u_{t_0}\,\exp(-\alpha(t-t_0)) + \frac{\beta}{\alpha}.$$

Proof. Consider $u'_t = \exp(\alpha(t-t_0))u_t$. Then its evolution follows

$$\mathrm{d}u'_t = \alpha u_t\,\mathrm{d}t + \exp(\alpha(t-t_0))\,\mathrm{d}u_t \leqslant \exp(\alpha(t-t_0))\beta + \exp(\alpha(t-t_0))\mathrm{d}M_t.$$

Integrating both hands from t_0 to t, then take conditional expectation we have

$$\mathbb{E}_{t_0}u'_t = u'_{t_0} + \frac{\beta}{\alpha}(\exp(\alpha(t-t_0)) - 1).$$

This leads to our claim. □

Lemma B.3. *For a positive random variable X, if there are constants $A \geqslant 2, B \geqslant 0$ such that $\mathbb{P}(X > M) \leqslant A\ \exp(-\lambda M) + \exp(\lambda B - \lambda M)$ holds for all $M > 0$, then*

$$\mathbb{E}[X] \leqslant \frac{1 + \log\,2A}{\lambda} + B.$$

Proof. Note that if we let $C = \frac{1}{\lambda}\log A + B$, which is the point the quantile upper bound takes value 1,

$$\begin{aligned}\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x)\,\mathrm{d}x = \int_C^\infty \mathbb{P}(X > x)\,\mathrm{d}x + \int_0^C \mathbb{P}(X > x)\,\mathrm{d}x\\ \leqslant \int_C^\infty (A + \exp(\lambda B))\exp(-\lambda x)\mathrm{d}x + \int_0^C 1\,\mathrm{d}x\\ = \frac{A + \exp(\lambda B)}{\lambda}\exp(-\lambda C) + C = \frac{1 + \log(A + \exp(\lambda B))}{\lambda}.\end{aligned}$$

Finally, since $A \leqslant A\exp(\lambda B), \exp(\lambda B) \leqslant A\exp(\lambda B)$, so $\log(A + \exp(\lambda B) \leqslant \lambda B + \log 2A$. □

B.3. Proof of theorem 3.4

Proof. claim (1). Note that $\tilde{\phi} \succeq 0$ and thus $\sum_{i,j}[P_t \circ \tilde{\phi}]_{j,i}[e_t]_i[e_t]_j \geqslant 0$, and $\sum_i \mathcal{F}_{\mathbf{d}(i,j)} \leqslant C_f$. So utilising lemma B.1 and summing over all j on both sides of (27) yields

$$\mathrm{d}\|e_t\|^2 \leqslant \left(C_f - \alpha_t\right)\|e_t\|^2\,\mathrm{d}t + N_x\beta_t\,\mathrm{d}t + \mathrm{d}\mathcal{M}'_t, \tag{36}$$

where the martingale is given by

$$d\mathcal{M}'_t = \sum_{j=1}^{N_x} 2\sqrt{2}[e_t]_j\mathrm{d}W_j - 2r^{-1/2}[e_t]_j[P_s^{\mathrm{L}}\Omega^{1/2}\mathrm{d}B]_j = 2\sqrt{2}e_t^{\mathrm{T}}\mathrm{d}W_t - 2r^{-1/2}e_t^{\mathrm{T}}P_t^{\mathrm{L}}\Omega^{1/2}\mathrm{d}B_t.$$

For $t \in [0, t_*]$, (36) can be further upper-bounded by

$$\mathrm{d}\|e_t\|^2 \leqslant \left(C_f - \alpha^*\right)\|e_t\|^2\,\mathrm{d}t + N_x\beta^*\,\mathrm{d}t + \mathrm{d}\mathcal{M}'_t.$$

Employing Gronwall's inequality, there is a constant D such that

$$\mathbb{E}\|e_{t_*}\|^2 \leqslant \exp((2C_f + 1)t_*)\mathbb{E}\|e_0\|^2 + N_x\beta^*\frac{\exp((2C_f + 1)t_*) - 1}{2C_f + 1} = \Theta(\epsilon^{-1}).$$

For $t \geqslant t_*$, (36) can be further upper-bounded by

$$\mathrm{d}\|e_t\|^2 \leqslant -\alpha'_*\|e_t\|^2\,\mathrm{d}t + N_x\beta_*\,\mathrm{d}t + \mathrm{d}\mathcal{M}'_t.$$

Employing Gronwall's inequality, we find that with $\alpha'_* = \alpha_* - C_f$,

$$\mathbb{E}\|e_t\|^2 \leqslant \mathbb{E}\|e_{t_*}\|^2 \exp(-\alpha'_*(t-t_*)) + \frac{N_x\beta_*}{\alpha'_*}(1-\exp(-\alpha'_*(t-t_*))). \tag{37}$$

Since $\alpha'_* = \alpha_* - C_f = \Theta(\epsilon^{-1/2})$, $\beta_* = \mathcal{O}(1)$, and $\epsilon^{-1}\exp(-\lambda\epsilon^{-1/2}) = o(1)$ for any $\lambda > 0$, so we have proved for claim (1). □

Proof. claim (2). First we note the quadratic variation of the martingale term $\mathcal{M}'_t$ is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle\mathcal{M}'\rangle_t = 8\|e_t\|^2 + 4\epsilon^{-1}\|\Omega^{1/2}P_t^{\mathrm{L}}e_t\|^2 \leqslant (8+4\epsilon^{-1}\omega_{\max}\|P_t^{\mathrm{L}}\|^2)\|e_t\|^2.$$

So by Ito's formula on $\exp(\lambda\|e_t\|^2)$, the following holds with $\alpha'_t = \alpha_t - C_f$,

$$\begin{aligned}\mathrm{d}\exp(\lambda\|e_t\|^2) &\leqslant ((-\lambda\alpha'_t\|e_t\|^2 + \lambda\beta_t N_x)\mathrm{d}t + \lambda\mathrm{d}\mathcal{M}'_t)\exp(\lambda\|e_t\|^2) + \frac{1}{2}\lambda^2\exp(\lambda\|e_t\|^2)\mathrm{d}\langle\mathcal{M}\rangle_t\,\mathrm{d}t\\ &\leqslant (-\gamma_t\|e_t\|^2 + \lambda\beta_t N_x)\exp(\lambda\|e_t\|^2)\mathrm{d}t + \lambda\exp(\lambda\|e_t\|^2)\mathrm{d}\mathcal{M}'_t.\end{aligned}$$

where

$$\gamma_t = \lambda\alpha'_t - 4\lambda^2 - 2\lambda^2\omega_{\max}\epsilon^{-1}\|P_t^{\mathrm{L}}\|^2.$$

By lemmas 3.1 and 3.2, we have for all $t > 0$

$$-\gamma_t \leqslant \gamma^* = \lambda(2C_f+1) + 4\lambda^2 + 2\lambda^2\omega_{\max}C_\phi^2\max\{\|P_0\|^2_{\max}, \lambda^2_{\max}\},$$

and for $t \geqslant t_*$

$$\lambda \leqslant \lambda_* = \frac{\alpha'_*}{8+4\omega_{\max}C_\phi^2\lambda^2_{\max}} = \Theta(\epsilon^{-1/2}).$$

$$\gamma_t \geqslant \lambda\alpha'_* - 4\lambda^2 - 2\lambda^2\omega_{\max}C_\phi^2\lambda^2_{\max} \geqslant \frac{1}{2}\lambda\alpha'_*.$$

For $t \leqslant t_*$, by Gronwall's inequality we have

$$\mathbb{E}\exp(\lambda\|e_{t_*}\|^2) \leqslant \exp((\gamma^* + \lambda\beta^*N_x)t_*)\exp(\lambda\|e_0\|^2). \tag{38}$$

And when $t \geqslant t_*$,

$$\mathrm{d}\exp(\lambda\|e_t\|^2) \leqslant (\lambda\beta_*N_x - \frac{1}{2}\lambda\alpha'_*\|e_t\|^2)\exp(\lambda\|e_t\|^2)\mathrm{d}t + \lambda\exp(\lambda\|e_t\|^2)\mathrm{d}\mathcal{M}'_t. \tag{39}$$

Note that when $\frac{1}{4}\lambda\alpha'_*\|e_t\|^2 \leqslant \lambda\beta_*N_x$,

$$\exp(\lambda\|e_t\|^2) \leqslant \exp\left(\frac{4\lambda N_x\beta_*}{\alpha'_*}\right),$$

we obtain

$$(\lambda\beta_* N_x - \frac{1}{2}\lambda\alpha'_*\|e_t\|^2)\exp(\lambda\|e_t\|^2) \leqslant -\lambda\beta_* N_x\ \exp(\lambda\|e_t\|^2) + 2\lambda\beta_* N_x\ \exp\left(\frac{4\lambda N_x\beta_*}{\alpha'_*}\right).$$

Otherwise, when $\frac{1}{4}\lambda\alpha'_*\|e_t\|^2 \geqslant \lambda\beta_* N_x$, we have

$$(\lambda\beta_* N_x - \frac{1}{2}\lambda\alpha'_*\|e_t\|^2)\exp(\lambda\|e_t\|^2) \leqslant -\frac{1}{4}\lambda\alpha'_*\|e_t\|^2\ \exp(\lambda\|e_t\|^2) \leqslant -\lambda\beta_* N_x\ \exp(\lambda\|e_t\|^2).$$

In summary, we always have

$$(\lambda\beta_* N_x - \frac{1}{2}\lambda\alpha'_*\|e_t\|^2)\exp(\lambda\|e_t\|^2) \leqslant -\lambda\beta_* N_x\ \exp(\lambda\|e_t\|^2) + 2\lambda\beta_* N_x\ \exp\left(\frac{4\lambda N_x\beta_*}{\alpha'_*}\right). \tag{40}$$

Inserting (40) in (39) yields

$$\mathrm{d}\ \exp(\lambda\|e_t\|^2) \leqslant \left[-\lambda\beta_* N_x\ \exp(\lambda\|e_t\|^2) + 2\lambda\beta_* N_x\ \exp\left(\frac{4\lambda N_x\beta_*}{\alpha'_*}\right)\right]\mathrm{d}t + \lambda\ \exp(\lambda\|e_t\|^2)\mathrm{d}\mathcal{M}'_t.$$

After applying Grönwall’s inequality and (38) we obtain the following

$$\begin{aligned}\mathbb{E}[\exp(\lambda\|e_t\|^2)] &\leqslant \exp(-\lambda\beta_* N_x(t-t_*))\mathbb{E}[\exp(\lambda\|e_{t_*}\|^2)] + 2\ \exp\left(\frac{4\lambda N_x\beta_*}{\alpha_*}\right)\\ &\leqslant \exp(-(\gamma^* + \lambda\beta^* N_x)t_* - \lambda\beta_* N_x(t-t_*))\mathbb{E}[\exp(\lambda\|e_0\|^2)] + 2\ \exp\left(\frac{4\lambda N_x\beta_*}{\alpha'_*}\right).\end{aligned}$$

When $t \to \infty$, this leads to claim (2):

$$\limsup_{t\to\infty}\mathbb{E}\ \exp(\lambda\|e_t\|^2) \leqslant 2\ \exp\left(\frac{4\lambda N_x\beta_*}{\alpha'_*}\right).$$

□

Proof. claim (3). We consider function

$$g(x) = (\lambda\beta_* N_x - \frac{1}{2}\lambda\alpha'_* x)\exp(\lambda x)$$

By finding the critical point, it is easy to see

$$g(x) \leqslant g\left(\frac{2\beta_* N_x}{\alpha'_*} - \frac{1}{\lambda}\right) = \frac{\alpha'_*}{2\mathrm{e}}\exp\left(\frac{2\lambda_*\beta N_x}{\alpha'_*}\right) = :G_* = \Theta(\epsilon^{-1/2})$$

Combine this with (39), we find

$$\mathrm{d}\ \exp(\lambda\|e_t\|^2) \leqslant G_* \mathrm{d}t + \lambda\ \exp(\lambda\|e_t\|^2)\mathrm{d}\mathcal{M}'_t, \quad \forall t \geqslant t_*.$$

So by Dynkin's formula, if we let $\tau = \min\{t : t \geqslant t_0, \|e_t\|^2 \geqslant M\}$, then

$$\mathbb{E}_{t_0}\ \exp(\lambda\|e_{T\wedge\tau}\|^2) \leqslant \exp(\lambda\|e_{t_0}\|^2) + \mathbb{E}\int_{t_0}^{T\wedge\tau} G_*\ \mathrm{d}t \leqslant \exp(\lambda\|e_{t_0}\|^2) + G_* T.$$

By Markov inequality we have

$$\begin{aligned}\mathbb{P}(\max_{t_0\leqslant t\leqslant T}\|e_t\|^2 \geqslant M) = \mathbb{P}_{t_0}(\|e_{T\wedge\tau}\|^2 \geqslant M) &\leqslant \frac{\mathbb{E}_{t_0}\ \exp(\lambda\|e_{T\wedge\tau}\|^2)}{\exp(\lambda M)} \\ &\leqslant \frac{\alpha'_* T}{2\mathbb{e}}\exp\left(\frac{2\lambda\beta_* N_x}{\alpha'_*} - \lambda M\right) + \exp(\lambda\|e_{t_0}\|^2 - \lambda M).\end{aligned}$$

Then by lemma B.3,

$$\mathbb{E}_{t_0}\max_{t_0\leqslant t\leqslant T}\|e_T\|^2 \leqslant \frac{1}{\lambda} + \frac{2\beta_* N_x}{\alpha'_*} + \frac{1}{\lambda}\log\left(\frac{\alpha'_* T}{\mathbb{e}}\right) + \frac{1}{\lambda} + \|e_{t_0}\|.$$

We take $\lambda = \lambda_* = \Theta(\epsilon^{-\frac{1}{2}})$ to obtain our claimed result. □

Appendix C. Proof for component-wise filter error analysis

C.1. Component-wise Lyapunov weights

In order to bound $[e_t]_i^2$ in long time, it is necessary to build a Lyapunov function for it. The main challenge here is that dynamics of $[e_t]_i^2$ is coupled with the error of other components. The idea is here to find a weight vector v^i so that $E_t^i = \sum_j v_j^i [e_t]_j^2$ is a Lyapunov function. The design of v^i happens to relate to the structure of ϕ, and can be expressed as the Green function of a Markov chain.

Lemma C.1. *Under assumption 3.5. Let T be a random variable of geometric-q distribution, that is*

$$\mathbb{P}(T = n) = (1-q)q^{n-1}, \quad n = 1, 2, \ldots.$$

Consider a Markov chain X_t on the points $\{1, \ldots, N_x\}$. Its transition probability is given by

$$\mathbb{P}(X_{t+1} = j | X_t = i) = \begin{cases} \dfrac{1}{q}\phi_{i,j} & j \neq i \\ 1 - \dfrac{1}{q}\displaystyle\sum_{j\neq i}\phi_{i,j} & j = i. \end{cases}$$

Fix an index $i \in \{1, \ldots, N_x\}$. Define vector v^i, where its components are given by

$$v_j^i = \mathbb{E}\left(\sum_{k=1}^{T}\mathbf{1}_{X_k=i}\middle| X_1 = j\right).$$

Then v^i satisfies the following properties

(1) $v^i_j \geqslant 0, \quad \forall j$ *and in specific* $v^i_i \geqslant 1 - q$.
(2) *For all index j,* $\sum_{l\neq j}\phi_{j,l}v^i_l \leqslant v^i_j$.
(3) $\sum_{j=1}^{N_x} v^i_j \leqslant 1$.

Proof. claim (1). Since $\sum_{k=1}^{T} \mathbf{1}_{X_k=i} \geqslant 0$ a.s., so $v^i_j \geqslant 0$. Moreover,

$$v^i_i = \mathbb{E}\left(\sum_{k=1}^{T} \mathbf{1}_{X_k=i} \middle| X_1 = i\right) \geqslant \mathbb{E}\left(\mathbf{1}_{T=1,X_1=i} \middle| X_1 = i\right) = 1 - q.$$

□

Proof. claim (2). Next, by doing a first step analysis of Markov chain, we find that

$$v^i_j = (1-q)\cdot \mathbf{1}_{j=i} + q\left(1 - \frac{1}{q}\sum_{l\neq j}\phi_{j,l}\right) v^i_j + q\cdot\frac{1}{q}\sum_{l\neq j}\phi_{j,l}v^i_l. \tag{41}$$

Since $\sum_{\,l\neq j}\phi_{j,l} \leqslant q < 1$, we have

$$v^i_j \geqslant \sum_{l\neq j}\phi_{l,j}v^i_l.$$

□

Proof. claim (3). We sum (41) over all j and obtain

$$\begin{aligned}\sum_{j=1}^{N_x} v^i_j &= (1-q) + q\sum_{j=1}^{N_x}\left(1 - \frac{1}{q}\sum_{l\neq j}\phi_{j,l}\right) v^i_j + \sum_{j=1}^{N_x}\sum_{l\neq j}\phi_{j,l}v^i_l \\ &\leqslant 1 - q + \sum_{j=1}^{N_x}\sum_{l\neq j}\phi_{j,l}v^i_l = 1 - q + \sum_{l=1}^{N_x} v^i_l\left(\sum_{j\neq l}\phi_{j,l}\right).\end{aligned}$$

Therefore we have

$$(1-q)\sum_{j=1}^{N_x} v^i_j \leqslant \sum_{j=1}^{N_x}(1 - \sum_{j\neq l}\phi_{j,l})v^i_j \leqslant 1 - q,$$

which leads to our claim. □

C.2. Proof of theorem 3.6

Proof. claim (1). Recall that lemma B.1 has shown that

$$\mathrm{d}[e_t]^2_i \leqslant \left(-\alpha_t[e_t]^2_i - 2\epsilon^{-1}\sum_{j=1}[P_t\circ\tilde{\phi}]_{i,j}[e_t]_i[e_t]_j + \sum_{j\neq i}\mathcal{F}_{\mathbf{d}(i,j)}|[e_t]_j|^2 + \beta_t\right)\mathrm{d}t + \mathrm{d}[\mathcal{M}_t]_i. \tag{42}$$

Recall that $\tilde{\phi} = \phi - \rho I$. In the following, we use $P_{j,i}$ to denote the (j,i)-th component of P_t. Then by Cauchy Schwartz and Young's inequality

$$
\begin{aligned}
-2[P_t \circ \tilde{\phi}]_{i,j}[e_t]_i[e_t]_j = -2\phi_{j,i}P_{j,i}[e_t]_i[e_t]_j &\leqslant -2\phi_{j,i}\sqrt{P_{j,j}}[e_t]_j\sqrt{P_{i,i}}[e_t]_i \\
&\leqslant \phi_{i,j}(P_{j,j}[e_t]_j^2 + P_{i,i}[e_t]_i^2), \quad \text{for } j \neq i.
\end{aligned}
$$

Then note that

$$
-2P_{i,i}\phi_{i,i}[e_t]_i^2 + \sum_{i\neq j}\phi_{i,j}P_{i,i}[e_t]_i^2 \leqslant (q-2)P_{i,i}[e_t]_i^2 < -P_{i,i}[e_t]_i^2,
$$

so (42) leads to

$$
\mathrm{d}[e_t]_i^2 \leqslant \left(\sum_{j\neq i}(\mathcal{F}_{\mathbf{d}(i,j)} + \epsilon^{-1}\phi_{i,j}P_{j,j})[e_t]_j^2 - \alpha_t[e_t]_i^2 - \epsilon^{-1}P_{i,i}[e_t]_i^2 + \beta_t\right)\mathrm{d}t + \mathrm{d}[\mathcal{M}_t]_i. \tag{43}
$$

We denote the vector $E_t = [e_1^2, e_2^2, \ldots\ e_N^2]^T$. Further we define vector v^i, of which the component is given by lemma C.1. Denote

$$
E_t^i = \langle v^i, E_t\rangle, \quad \mathcal{M}_t^i = \langle v^i, \mathcal{M}_t\rangle.
$$

Then the SDE of E_t^i can be bounded by a linear combination of (43), which is

$$
\begin{aligned}
\mathrm{d}E_t^i &\leqslant \sum_{j=1}^{N_x}\left(-\alpha_t v_j^i[e_t]_j^2 - \epsilon^{-1}v_j^i(P_{j,j}[e_t]_j^2 - \sum_{l\neq j}\phi_{j,l}P_{l,l}\mathrm{e}_l^2) + \sum_{l\neq j}v_j^i\mathcal{F}_{\mathbf{d}(j,l)}e_l^2\right) + \beta_t + \mathrm{d}\mathcal{M}_t^i \\
&= \sum_{j=1}^{N_x}\left(-\alpha_t v_j^i[e_t]_j^2 - \epsilon^{-1}(v_j^iP_{j,j}[e_t]_j^2 - \sum_{l\neq j}v_l^i\phi_{l,j}P_{j,j}[e_t]_j^2) + \sum_{l\neq j}\mathcal{F}_{\mathbf{d}(j,l)}v_l^i[e_t]_j^2\right) + \beta_t + \mathrm{d}\mathcal{M}_t^i \\
&\leqslant \sum_{j=1}^{N_x}\left(-\alpha_t v_j^i[e_t]_j^2 v_j^i + C_{\mathcal{F}}\phi_{j,l}v_l^i[e_t]_j^2\right) + \beta_t + \mathrm{d}\mathcal{M}_t^i. \qquad (44) \\
&\leqslant \sum_{j=1}^{N_x}(-\alpha_t + C_{\mathcal{F}})v_j^i[e_t]_j^2 + \beta_t + \mathrm{d}\mathcal{M}_t^i = (-\alpha_t + C_{\mathcal{F}})E_t^i + \beta_t + \mathrm{d}\mathcal{M}_t^i. \qquad (45)
\end{aligned}
$$

We have used claims (2) and (3) of lemma C.1 at (44) and (45).

Between time 0 and t_ϵ, recall the upper bound in lemma B.1, apply Gronwall's inequality

$$
\mathbb{E}E_{t_\epsilon}^i \leqslant \exp((C_{\mathcal{F}} - \alpha^*)t_\epsilon)\left(\mathbb{E}E_0^i + \frac{\beta^*}{C_{\mathcal{F}} - \alpha^*}\right).
$$

Then after t_ϵ, for any t, apply Gronwall's inequality

$$\begin{aligned}\mathbb{E}E_t^i &\leqslant \exp((C_{\mathcal{F}}-\alpha_*)t_\epsilon)\mathbb{E}E_{t_\epsilon}^i+\frac{\beta_*}{\alpha_*-C_{\mathcal{F}}}\\ &\leqslant \exp(C_{\mathcal{F}}t-\alpha^*t_\epsilon-\alpha_*(t-t_\epsilon))\left(\mathbb{E}E_0^i+\frac{\beta^*}{C_{\mathcal{F}}-\alpha^*}\right)+\frac{\beta_*}{\alpha_*-C_{\mathcal{F}}}.\end{aligned}$$

Recall that in lemma B.1, $\alpha_*=\Theta(\epsilon^{-1/2}),\beta*=\Theta(\epsilon^{-1}),\alpha*=\beta_*=\Theta(1)$. So if $t>t_0$, for certain constants c and C

$$-(C_{\mathcal{F}}t-\alpha^*t_\epsilon-\alpha_*(t-t_\epsilon))\geqslant c\epsilon^{-1/2},\quad \mathbb{E}E_0^i\leqslant\max_i\{|[e_t]_i(0)|^2\}\sum_j v_j^i\leqslant C,$$

$$\frac{\beta^*}{C_{\mathcal{F}}-\alpha^*}\leqslant C\epsilon^{-1},\quad \frac{\beta_*}{\alpha_*-C_{\mathcal{F}}}\leqslant C\epsilon^{1/2}.$$

Therefore when ϵ is small enough, $\mathbb{E}E_t^i\leqslant 2C\sqrt{\epsilon}$, which is our claim (1). □

Proof. claim (2). First recall the individual martingale driving E_t^i is given by

$$\mathrm{d}\mathcal{M}_t^i=\sum_j v_j^i\sqrt{8}[e_t]_j\mathrm{d}W_j-2v_j^i\epsilon^{-1/2}[e_t]_j[P_t^{\mathrm{L}}\Omega^{1/2}\mathrm{d}B]_j.$$

The corresponding quadratic variation is bounded by

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\langle\mathcal{M}^i\rangle_t &= 8\sum_{j=1}^{N_x}[e_t]_j^2(v_j^i)^2+4\epsilon^{-1}\sum_{j=1}^{N_x}(v_j^i)^2[e_t]_j^2\sum_{l=1}^{N_x}[P_t^L]_{j,l}^2[\Omega]_{l,l}\\ &\leqslant 8\sum_{j=1}^{N_x}[e_t]_j^2v_j^i+4\omega_{\max}\epsilon^{-1}\|P_t\|_{\max}^2\sum_{j=1}^{N_x}v_j^i[e_t]_j^2\leqslant 4\beta_tE_t^i.\end{aligned}$$

Denote $\alpha_t'=\alpha_t-C_{\mathcal{F}}$, (which is slightly different from the one in the proof of theorem 3.4) then recall from (45) we have

$$\mathrm{d}E_t^i\leqslant-\alpha_t'E_t^i\,\mathrm{d}t+\beta_t\mathrm{d}t+\mathrm{d}\mathcal{M}_t^i.$$

By Ito's formula on $\exp(\lambda E_t^i)$, we have

$$\begin{aligned}\mathrm{d}\,\exp(\lambda E_t^i)&\leqslant(-\frac{1}{2}\lambda\alpha_t'E_t^i+4\lambda\beta_t)\mathrm{d}t+\lambda d\mathcal{M}_t^i)\exp(\lambda E_t^i)+\frac{1}{2}\lambda^2\,\exp(\lambda E_t^i)\mathrm{d}\langle\mathcal{M}^i\rangle_t\\ &\leqslant(-\frac{1}{2}(\lambda\alpha_t'-4\lambda^2\beta_t)E_t^i+\lambda\beta_t)\exp(\lambda E_t^i)\mathrm{d}t+\lambda\,\exp(\lambda E_t^i)\mathrm{d}\mathcal{M}_t^i.\quad(46)\end{aligned}$$

From time 0 to t_ϵ, by lemma B.1,

$$\alpha_t'=\alpha_t-C_{\mathcal{F}}\geqslant\alpha^*-C_{\mathcal{F}},\quad \beta_t\leqslant\beta^*,$$

by Gronwall's inequality, for all i

$$\mathbb{E}\,\exp(\lambda E^i_{t_\epsilon}) \leqslant \exp(t_\epsilon(-\frac{1}{2}\lambda(\alpha^* - C_{\mathcal{F}}) + 2\lambda^2\beta^* + \frac{1}{2}\lambda\beta^*))\exp(\lambda \max_i\{|\mathrm{e}^i_t(0)|^2\}) < \infty. \tag{47}$$

When $t > t_\epsilon$, lemma B.1 further shows that

$$\alpha'_t = \alpha_t - C_{\mathcal{F}} \geqslant \alpha'_* := \alpha_* - C_{\mathcal{F}}, \quad \beta_t \leqslant \beta_*.$$

Consider $\lambda \leqslant \lambda_* = \frac{\alpha'_*}{8\beta_*}$, then

$$-\frac{1}{2}(\lambda\alpha_* - 4\lambda^2\beta_*) = -\frac{1}{4}\lambda\alpha_*.$$

Then for $t > t_\epsilon$ and $\lambda < \lambda_\epsilon$, we have the following upper bound from (46)

$$\mathrm{d}\,\exp(\lambda E^i_t) \leqslant \left(-\frac{1}{4}\lambda\alpha'_* E^i_t + \lambda\beta_*\right)\exp(\lambda E^i_t)\mathrm{d}t + \lambda\,\exp(\lambda E^i_t)\mathrm{d}\mathcal{M}^i_t. \tag{48}$$

When ϵ is small enough, $\alpha'_* > 0$. Then if $\frac{1}{8}\lambda\alpha'_* E^i_t \leqslant \lambda\beta_*$,

$$\left(-\frac{1}{4}\lambda\alpha'_* E^i_t + \lambda\beta_*\right)\exp(\lambda E^i_t) + \frac{1}{8}\lambda\alpha'_*\,\exp(\lambda E^i_t) \leqslant 2\lambda\beta_*\,\exp(8\lambda\beta_*/\alpha'_*).$$

If $\frac{1}{8}\lambda\alpha'_* E^i_t \geqslant \lambda\beta_*$,

$$\left(-\frac{1}{4}\lambda\alpha'_* E^i_t + \lambda\beta_*\right)\exp(\lambda E^i_t) \leqslant -\frac{1}{8}\lambda\alpha'_*\,\exp(\lambda E^i_t).$$

In summary, we always have

$$\left(-\frac{1}{4}\lambda\alpha'_* E^i_t + \lambda\beta_*\right)\exp(\lambda E^i_t) \leqslant -\frac{1}{8}\lambda\alpha'_*\,\exp(\lambda E^i_t) + 2\lambda\beta_*\,\exp(8\lambda\beta_*/\alpha'_*).$$

Plug this into (48), we have

$$\mathrm{d}\,\exp(\lambda E^i_t) \leqslant \left(-\frac{1}{8}\lambda\alpha'_*\,\exp(\lambda E^i_t) + 2\lambda\beta_*\,\exp(8\lambda\beta_*/\alpha'_*)\right)\mathrm{d}t + \lambda\,\exp(\lambda E^i_t)\mathrm{d}\mathcal{M}^i_t. \tag{49}$$

So Gronwall's inequality and implies for $t \geqslant t_\epsilon$

$$\mathbb{E}\,\exp(\lambda E^i_t) \leqslant \exp\left(-\frac{1}{8}\lambda\alpha'_*(t - t_\epsilon)\right)\mathbb{E}\,\exp(\lambda E^i_{t_\epsilon}) + 16\frac{\beta_*}{\alpha'_*}\exp(8\lambda\beta_*/\alpha'_*)$$

The first term on the right converges to zero as $t \to \infty$ because of bound (47). We have our claim (2) because of $\beta_* = \Theta(1)$, $\alpha'_* = \Theta(\epsilon^{-1/2})$, moreover $E^i_t \geqslant v^i_i[e_t]^2_i \geqslant (1-q)[e_t]^2_i$ by lemma C.1 claim (1).

□

Proof. claim (3). We consider function

$$g(x) = \left(-\frac{1}{4}\lambda\alpha'_* x + \lambda\beta_*\right)\exp(\lambda x)$$

and by finding the critical point, it is easy to see

$$g(x) \leqslant g\left(\frac{4\beta_*}{\alpha'_*} - \frac{1}{\lambda}\right) = \frac{\alpha_*}{4\mathbb{e}} \exp\left(\frac{4\lambda\beta_*}{\alpha'_*}\right) =: G_*.$$

Plug this into (48), we find for all $t > t_0$,

$$\mathrm{d}\ \exp(\lambda E^i_t) \leqslant G_* \mathrm{d}t + \lambda\ \exp(\lambda E^i_t)\mathrm{d}\mathcal{M}^i_t.$$

So by Dynkin's formula, if we let $\tau_i = \min\{t : E^i_t \geqslant M\}$, then

$$\mathbb{E}_{t_0}[\exp(\lambda E^i_{T\wedge\tau})] \leqslant \exp(\lambda E^i_{t_0}) + \mathbb{E}_{t_0}[\int_{t_0}^{T\wedge\tau} G_*\ \mathrm{d}t] \leqslant \exp(\lambda E^i_{t_0}) + G_* T.$$

Recall that $E^i_t \geqslant v^i_i[e_t]^2_i \geqslant (1-q)[e_t]^2_i$. By Markov inequality

$$\begin{aligned}\mathbb{P}\left(\sup_{t_0\leqslant t\leqslant T}\{[e_t]^2_i\} \geqslant \frac{M}{1-q}\right) &\leqslant \mathbb{P}\left(\sup_{t_0\leqslant t\leqslant T} E^i_t \geqslant M\right)\\ &\leqslant \frac{\mathbb{E}\ \exp(\lambda E^i_{T\wedge\tau})}{\exp(\lambda M)}\\ &\leqslant \frac{\alpha_* T}{4\mathbb{e}} \exp\left(\frac{4\lambda\beta_*}{\alpha'_*} - ((1-q)\lambda)\frac{M}{(1-q)}\right)\\ &\quad + \exp\left(\lambda E^i_{t_0} - ((1-q)\lambda)\frac{M}{(1-q)}\right).\end{aligned}$$

Note that $E^i_{t_0} = \sum_{j=1}^{N_x} v^i_j[e_{t_0}]^2_j \leqslant \max_j[e_{t_0}]^2_j$.Then by lemma B.3, we have

$$\mathbb{E}[\sup_{t_0\leqslant t\leqslant T}\{[e_t]^2_i\}] \leqslant \frac{1}{(1-q)\lambda} + \frac{4\beta_*}{\alpha'_*(1-q)} + \frac{1}{\lambda}\log\left(\frac{\alpha'_* T}{2e}\right) + \max_i[e_{t_0}]^2_j.$$

We have claim (3) because $\beta_* = \Theta(1)$, $\alpha'_* = \Theta(\epsilon^{-1/2})$ and taking $\lambda = \lambda_\epsilon = \Theta(\epsilon^{-1/2})$. □

Proof. claim (4). We note

$$\begin{aligned}\mathbb{P}\left(\max_i \sup_{t_0\leqslant t\leqslant T}\{[e_t]^2_i\} \geqslant \frac{M}{1-q}\right) &\leqslant \sum_{i=1}^{N_x}\mathbb{P}\left(\sup_{t_0\leqslant t\leqslant T}\{[e_t]^2_i\} \geqslant \frac{M}{1-q}\right)\\ &\leqslant \sum_{i=1}^{N_x}\mathbb{P}\left(\sup_{t_0\leqslant t\leqslant T} E^i_t \geqslant M\right)\\ &\leqslant \sum_{i=1}^{N_x}\frac{\mathbb{E}\ \exp(\lambda E^i_{T\wedge\tau})}{\exp(\lambda M)}\\ &\leqslant \frac{N_x\alpha_* T}{4\mathbb{e}} \exp\left(\frac{4\lambda\beta_*}{\alpha'_*} - ((1-q)\lambda)\frac{M}{(1-q)}\right)\\ &\quad + N_x\ \exp\left(\lambda E^i_{t_0} - ((1-q)\lambda)\frac{M}{(1-q)}\right).\end{aligned}$$

Then by lemma B.3 and $E^i_{t_0} \leqslant \sum_{j=1}^{N_x} v^i_j[e_{t_0}]_j^2 \leqslant \max_j\{[e_{t_0}]_j^2\}$

$$\mathbb{E}_{t_0} \max_i \sup_{t_0\leqslant t\leqslant T}\{[e_t]_i^2\} \leqslant \frac{1}{(1-q)\lambda} + \frac{4\beta_*}{\alpha_*(1-q)} + \frac{1}{\lambda}\log\left(\frac{\alpha_* N_x T}{\mathrm{e}}\right) + \log N_x + \max_j\{[e_{t_0}]_j^2\}.$$

We take $\lambda = \lambda_\epsilon = \Theta(\epsilon^{-\frac{1}{2}})$ to obtain our claimed result. □

ORCID iDs

Jana de Wiljes https://orcid.org/0000-0002-9636-1147
Xin T Tong https://orcid.org/0000-0002-8124-612X

References

[1] Amezcua J, Kalnay E, Ide K and Reich S 2014 Ensemble transform Kalman–Bucy filters *Q. J. R. Meteorol. Soc.* **140** 995–1004

[2] Bergemann K and Reich S 2012 An ensemble Kalman–Bucy filter for continuous data assimilation *Meteorol. Z.* **21** 213–9

[3] Bickel P J and Levina E 2008 Regularized estimation of large covariance matrices *Ann. Stat.* **36** 199–227

[4] Bishop A N and Del Moral P 2017 On the stability of Kalman–Bucy diffusion *SIAM J. Control Optim.* **55** 4015–47

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

[6] Del Moral P and Tugaut J 2018 On the stability and the uniform propagation of chaos properties of ensemble Kalman–Bucy filters *Ann. Appl. Probab.* **28** 790–850

[7] Evensen G 2006 *Data Assimilation: The Ensemble Kalman Filter* (Berlin: Springer)

[8] Le Gland F, Monbet V and Tran V-D 2009 Large sample asymptotics for the ensemble Kalman filter *INRIA Research Report* RR-7014 inria-00409060

[9] Gaspari G and Cohn S E 1999 Construction of correlation functions in two and three dimensions *Q. J. R. Meteorol. Soc.* **125** 723–57

[10] Hoel H and Kody J H 2016 Law, and Raul Tempone. Multilevel ensemble Kalman filtering *SIAM J. Numer. Anal.* **54** 1813–39

[11] Hoel H, Shaimerdenova G and Tempone R 2020 Multilevel ensemble Kalman filtering with local-level Kalman gains (arXiv:2002.00480)

[12] Jazwinski A H 1970 *Stochastic Processes and Filtering Theory* (New York: Academic)

[13] Kalman R E 1960 A new approach to linear filtering and prediction problems *Trans. ASME* **82** 35–45

[14] Kalnay E 2002 *Atmospheric Modeling, Data Assimilation and Predictability* (Cambridge: Cambridge University Press)

[15] Kelly D, Majda A J and Tong X T 2015 Concrete ensemble Kalman filters with rigorous catastrophic filter divergence *Proc. Natl Acad. Sci. USA* **112** 10589–94

[16] Kelly D T, Law K J H and Stuart A 2014 Well-posedness and accuracy of the ensemble Kalman filter in discrete and continuous time *Nonlinearity* **27** 2579–604

[17] Evan K and Mandel J 2015 Convergence of the square root ensemble kalman filter in the large ensemble limit *SIAM/ASA J. Uncertain. Quantification* **3** 1–17

[18] Lange T and Stannat W 2019 On the continuous time limit of ensemble square root filters (arXiv:1910.12493)

[19] Law K J H, Tembine H and Tempone R 2016 Deterministic mean-field ensemble kalman filtering *SIAM J. Sci. Comput.* **38** A1251–79

[20] Lorenz E N 1996 Predictibility: a problem partly solved *Proc. Seminar on Predictibility* vol 1 (Reading: ECMWF) pp 1–18

[21] Majda A and Tong X T 2018 Performance of ensemble Kalman filters in large dimensions *Commun. Pure Appl. Math.* **71** 892–937

[22] Mandel J, Cobb L and Beezley J D 2011 On the convergence of the ensemble Kalman filter *Appl. Math.* **56** 533–41

[23] Morzfeld M, Tong X T and Marzouk Y M 2019 Localisation for MCMC: sampling high-dimensional posterior distributions with local structure *J. Comput. Phys.* **310** 1–28

[24] Oliver D, Reynolds A and Liu N 2008 *Inverse Theory for Petroleum Reservoir Characterization and History Matching* (Cambridge: Cambridge University Press)

[25] Reich S 2013 A nonparametric ensemble transform method for Bayesian inference *SIAM J. Sci. Comput.* **35** A2013–24

[26] Reich S and Cotter C J 2015 *Probabilistic Forecasting and Bayesian Data Assimilation* (Cambridge: Cambridge University Press)

[27] Simo S 2013 *Bayesian Filtering and Smoothing* (Cambridge: Cambridge University Press)

[28] Tong X T 2018 Performance analysis of local ensemble Kalman filter *J. Nonlinear Sci.* **28** 1397–442

[29] Tong X T, Majda A J and Kelly D 2016 Nonlinear stability and ergodicity of ensemble based Kalman filters *Nonlinearity* **29** 657

[30] Tong X T, Morzfeld M and Marzouk Y M 2020 MALA-within-Gibbs samplers for high-dimensional distributions with sparse conditional structure *SIAM J. Sci. Comput.* accepted (arXiv: 1908.09429)

[31] Yang T, Mehta P G and Meyn S P 2013 Feedback particle filter *IEEE Trans. Autom. Control* **58** 2465–80