Ensemble Filtering for High Dimensional Non-linear State Space Models

Jing Lei and Peter Bickel
Department of Statistics, UC Berkeley

August 31, 2009

Abstract

We consider non-linear state space models in high-dimensional situations, where the two common tools for state space models both have difficulties. The Kalman filter variants are seriously biased due to non-Gaussianity and the particle filter suffers from the "curse of dimensionality".

Inspired by a regression perspective on the Kalman filter, a novel approach is developed by combining the Kalman filter and particle filter, retaining the stability of the Kalman filter in large systems as well as the accuracy of particle filters in highly non-linear systems. Its theoretical properties are justified under the Gaussian linear models as an extension of the Kalman filter. Its performance is tested and compared with other methods on a simulated chaotic system which is used widely in numerical weather forecasting.

1 Introduction

A state space model (SSM) typically consists of two time series $\{X_t : t \geq 1\}$ and $\{Y_t : t \geq 1\}$ defined by the following model:

$$\begin{aligned} X_{t+1} &= f_t(X_t, U_t), \quad f_t(\cdot,\cdot) : \mathbb{R}^p \times [0,1] \mapsto \mathbb{R}^p, \\ (Y_t | X_t = x) &\sim g(\cdot; x), \quad g(\cdot;\cdot) : \mathbb{R}^q \times \mathbb{R}^p \mapsto \mathbb{R}^+. \end{aligned} \tag{1}$$

where U_t is a random variable independent of everything else with uniform distribution on $[0,1]$ and $g(\cdot; x)$ is a density function for each x. The state variable X_t evolves according to the dynamics $f_t(\cdot, U_t)$ is usually of interest but never directly observed. Instead it can only be learned indirectly through the observations Y_t. SSM has been widely used in sciences and engineering including signal processing, public health, ecology, economics and geophysics. For a comprehensive summary, please see [12, 17]. A central problem in SSM is the filtering problem: assume that $f(\cdot,\cdot)$ and $g(\cdot;\cdot)$ are known, how can one approximate the distribution of X_t given the observations $Y_1^t := (Y_1, \ldots, Y_t)$ and the initial distribution of X_0, for every $t \geq 1$? A related problem of much practical interest is the tracking problem: for a realization of the SSM, how can one locate the current hidden state X_t based on the past observations Y_1^t? Usually the filtered expectation $\hat{X}_t = E(X_t|Y_1^t, X_0)$ can be used to approximate X_t.

A closed form solution to the filtering problem is available only for few special cases such as the Gaussian linear model (Kalman filter). The Kalman filter variants for non-linear dynamics includes the extended Kalman filter (EKF), the unscented Kalman filter (UKF, [16]) and the Ensemble Kalman filter (EnKF, [10]). The ensemble Kalman filter (EnKF), a combination of the sequential Monte Carlo (SMC, see below) and the Kalman filter, mostly used in geophysical data assimilation, has performed successfully in high dimensional models [11].

Despite the ease of implementation of the Kalman filter variants, they might still be seriously biased because the accuracy of the Kalman filter update requires the linearity of the observation function and the Gaussianity of the distribution of X_t given Y_1^{t-1}, both of which might fail in reality. Another class of ensemble filtering technique is the sequential Monte Carlo (SMC [22]) method, or the particle filter (PF, [14]).

The basic idea of the PF (also the EnKF) is using a discrete set of n weighted particles to represent the distribution of X_t, where the distribution is updated at each time by changing the particle weights according to their likelihoods. It can be shown that the PF is consistent under certain conditions, e.g., when the hidden Markov chain $\{X_t : t \geq 1\}$ is ergodic and the state space is compact [18], whereas the EnKF in general is not [19, 20].

A major challenge arises when p, q are very large in model (1) while n is relatively small. In typical climate models p can be a few thousands with n being only a few tens or hundreds. Even Kalman filter variants cannot work on the whole state variable because it is hard to estimate very large covariance matrices. It is also known that the particle filter suffers from the "curse of dimensionality" due to its nonparametric nature [3] even for moderately large p. As a result, dimension reduction must be employed in the filtering procedure. For example, a widely employed technique in geophysics is "localization": the whole state vector and observation vector are decomposed into many overlapping local patches according to their physical location. Filtering is performed on each local patch and the local updates are pieced back to get the update of the whole state vector. Such a scheme works for the EnKF but not for the PF because the former keeps track of each particle whereas the PF involves a reweighting/resampling step in the update of each local patch and there is no straightforward way of reconstructing the whole vector since the correlation among the patches is lost in the local reweighting/resampling step.

To sum up it is desirable to have an non-linear filtering method that is easily localizable like the EnKF and adaptive to non-linearity and non-Gaussianity like the PF. In this paper we propose a nonlinear filter that combines the advantages of both the EnKF and the PF. This is a filter that keeps track of each particle and use direct particle transformation like the EnKF while using importance sampling as the PF to avoid serious bias. The new filter, which we call the Non-Linear Ensemble Adjustment Filter (NLEAF), is indeed a further combination of the EnKF and the PF in that it uses a moment-matching idea to update the particles while using importance sampling to estimate the posterior moments. It is conceptually and practically simple and performs competitively in simulations. Single step consistency can be shown for certain Gaussian linear models.

In Section 2 we describe EnKF and PF with emphasis on the issue of dimension reduction. The NLEAF method is described and the consistency issue is discussed in Section 3. In Section 4 we present the simulation results on two chaotic system.

2 Background of ensemble filtering

2.1 Ensemble filtering at a single time step

Since the filtering methods considered in this paper are all recursive, from now on we focus on a single time step and drop the time index t whenever there is no confusion. Let X_f denote the variable $(X_t|Y_1^{t-1})$ where the subindex f stands for "forecast", and Y denote Y_t. Let X_u denote the conditional random variable $(X_t|Y_1^t)$.

Suppose the forecast ensemble $\{x_f^{(i)}\}_{i=1}^n$ is a random sample from X_f, and the observation $Y = y$ is also available. There are two inference tasks in the filtering/tracking procedure:

(a) Estimate $E(X_u)$ to locate the current state.

(b) Generate the updated ensemble $\{x_u^{(i)}\}_{i=1}^n$, i.e., a random sample from X_u, which will be used to generate the forecast ensemble at next time.

2.2 The ensemble Kalman filter [9, 10, 11]

We first introduce the Kalman filter. Assuming a Gaussian forecast distribution: , and a linear observation model

$$\begin{aligned} X_f &\sim N(\mu_f, \Sigma_f), \\ Y &= HX_f + \epsilon, \quad \epsilon \sim N(0, R), \end{aligned} \tag{2}$$

then $X_u = (X_f|Y)$ is still Gaussian:

$$X_u \sim N(\mu_u, \Sigma_u),$$

where

$$\mu_u = \mu_f + K(y - H\mu_f), \quad \Sigma_u = (I - KH)\Sigma_f, \tag{3}$$

and

$$K = \Sigma_f H^T (H\Sigma_f H^T + R)^{-1} \tag{4}$$

is the *Kalman gain.*

The EnKF approximates the forecast distribution by a Gaussian with the empirical mean and covariance, then updates the parameters using the Kalman filter formula. Recall the two inference tasks listed in Section 2.1. The estimation of $E(X_u)$ is straightforward using the Kalman filter formula. To generate the updated ensemble, a naïve (and necessary if in the Gaussian case) idea is to sample directly from the updated Gaussian distribution. This will, as verified widely in practice, lose much information in the forecast ensemble, such as skewness, kurtosis, clustering, etc. Instead, in the EnKF update, the updated ensemble is obtained by shifting and re-scaling the forecast ensemble. A brief EnKF algorithm is described as below:

The EnKF procedure

1. Estimate $\hat{\mu}_f$, $\hat{\Sigma}_f$.

2. Let $\hat{K} = \hat{\Sigma}_f H^T (H\hat{\Sigma}_f H^T + R)^{-1}$.

3. $\hat{\mu}_u = (I - \hat{K}H)\hat{\mu}_f + \hat{K}y$.

4. $x_u^{(i)} = x_f^{(i)} + \hat{K}(y - Hx_f^{(i)} - \epsilon^{(i)})$, with $\epsilon^{(i)} \overset{iid}{\sim} N(0, R)$.[1]

[1] In step 4 there is another update scheme which does not use the random perturbations $\epsilon^{(i)}$. This deterministic update, also known as the Kalman square-root filter, is usually used to avoid sampling error when the ensemble size is very small [1, 5, 29, 27, 20].

5. The next forecast ensemble is obtained by plugging each particle into the dynamics: $x_{t+1,f}^{(i)} = f_t(x_u^{(i)}, u_i)$, $i = 1, \ldots, n$.

Under model (2) the updated ensemble is approximately a random sample from X_u and that $\hat{\mu}_u \to \mu_u$ as $n \to \infty$. The method would be biased if the model (2) does not hold [13]. Large sample asymptotic results can be found in [19], where the first two moments of the EnKF are shown to be consistent under the Gaussian linear model, see also [20].

2.3 The particle filter

The particle filter [14, 22] also approximates the distribution of X_f by a set of particles. It differs from the EnKF in that instead of assuming a Gaussian and linear model, it reweights the particles according to their likelihood. Formally, one simple version of the PF acts as the following:

A simple version of the particle filter

1. Compute weight $W_i = \frac{g(y;x_f^{(i)})}{\sum_{j=1}^n g(y;x_f^{(i)})}$ for $i = 1, \ldots, n$.
2. The updated mean $\hat{\mu}_u = \frac{\sum_{i=1}^n x_f^{(i)} g(y;x_f^{(i)})}{\sum_{i=1}^n g(y;x_f^{(i)})}$.
3. Generate n random samples $x_u^{(1)}, \ldots, x_u^{(n)}$ i.i.d from $\{x_f^{(i)}\}_{i=1}^n$ with probability $P(X_u^{(1)} = x_f^{(i)}) = W_i$ for $i = 1, \ldots, n$.

It can be shown [18] that under strong conditions such as compactness of the state space and mixing conditions of the dynamics, the particle approximation of the forecast distribution is consistent in L_1 norm uniformly for all $1 \le t \le T_n$, for $T_n \to \infty$ subexponentially in n. However, it is well-known that the PF has a tendency to collapse (also known as sample degeneracy) especially in high-dimensional situations, see [21], and rigorous results in [3]. It is suggested that the ensemble size n needs to be at least exponential in p to avoid collapse.

Another fundamental difference between the PF and the EnKF is that in the PF, $x_u^{(i)}$ is generally not directly related to the $x_f^{(i)}$ because of reweighting/resampling. Recall that in the EnKF update, each particle is updated explicitly and $x_u^{(i)}$ does correspond to $x_f^{(i)}$. This difference materializes in the dimension reduction as discussed below.

2.4 Dimension reduction via localization

Dimension reduction becomes necessary for both EnKF and PF when X and Y are high dimensional, e.g., in numerical weather forecasting X and Y represents the underlying and observed weather condition. It is usually the case that the coordinates of the state vector X and observation Y are physical quantities measured at different grid points in the physical space. Therefore it is reasonable to assume that two points far away in the physical space have little correlation, and the corresponding coordinates of the state vector can be updated independently using only the "relevant" data [15, 4, 25, 2]. Formally, let $X = (X(1), ..., X(p))^T$. One can decompose the index set $\{1, \ldots, p\}$ into L (possibly overlapping) local windows $N_1, \ldots, N_L$ such that $|N_l| \ll p$ and $\bigcup_l N_l = \{1, \ldots, p\}$, and correspondingly decompse $\{1, \ldots, q\}$ into $\{N'_1, \ldots, N'_L\}$ such that $|N'_l| \ll q$ and $\bigcup_l N'_l = \{1, \ldots, q\}$. Let $X_f(N_l)$ denote the subvector of X_f consisting of the coordinates in N_l, and similarly define $Y(N'_l)$. $Y(N'_l)$ is usually chosen as the *local observation* of local state vector $X_f(N_l)$.

The localization of the EnKF is straightforward: For each local window N_l and its corresponding local observation window N'_l, one can apply the EnKF on $\{x_f^{(i)}(N_l)\}_{i=1}^n$ and $y(N'_l)$ with local observation matrix $H(N'_l, N_l)$, which is the corresponding submatrix of H. In the L local EnKF updates, each coordinate of X might be updated in multiple local windows. The final update is a convex combination of these multiple updates. Such a localized EnKF has been successfully implemented in the Lorenz 96 system (a 40 dimensional chaotic system, see Section 4) with the sample (ensemble) size being only 10 [25]. The localization idea will be further explained in Section 3.1. To be clear, we summarize the localized EnKF as simply L parallel runs of EnKF plus a piecing step:

The localized EnKF

1. For $l = 1, \ldots, L$, run the EnKF on $\{x_f^{(i)}(N_l)\}_{i=1}^n$ and $y(N'_l)$, with local observation matrix $H(N'_l, N_l)$. Store the results: $\hat{\mu}_u(N_l)$ and $\{x_u^{(i)}(N_l)\}_{i=1}^n$.
2. For each $j = 1, \ldots, p$, let $\hat{\mu}_u(j) = \sum_{l:j\in N_l} w_{j,l}\hat{\mu}_u(N_l; j)$, and $x_u^{(i)}(j) = \sum_{l:j\in N_l} w_{j,l} x_u^{(i)}(N_l; j)$, where $X(N_l; j)$ is the coordinate of $X(N_l)$ that corresponds to $X(j)$, and $w_{j,l} \geq 0$, $\sum_{l:j\in N_l} w_{j,l} = 1$.

The choices of local windows N_l, N'_l and combination coefficients $w_{j,l}$ can be pre-determined since in many applications there are simple and natural choices. They can also be chosen in a data-driven fashion. For example, as we will explain later, the Kalman filter is essentially a linear regression of X on Y. Therefore for each coordinate of X one can use sparse regression techniques to select the most relevant coordinates in Y. Similarly the choice of $w_{j,l}$ in the algorithm can be viewed as a problem of combining the predictions from multiple regression models and can be calculated from the data [6, 30, 7]. We will return to this issue in Section 3.1.

Table 1: A quick comparison of the EnKF and the PF.

	consistent	stable	localizable
EnKF	✗	✓	✓
PF	✓	✗	✗

On the other hand, such a dimension reduction scheme is not applicable to the PF because each particle is reweighted differently in different local windows. In words, the reweighting breaks the strong connection of a single particle update in different local windows and it is not clear how to combine the updated particle across the local windows. This can be viewed as a form of sample degeneracy: in high dimension situations, a particle might be plausible in some coordinates but absurd in other coordinates.

So far, the properties of the EnKF and the PF can be summarized as in Table 1, where the only check mark for the PF is higher accuracy. A natural idea to reduce the bias of EnKF is to update the mean of X using importance sampling as in the PF. Meanwhile, a possible improvement of the PF is avoiding the reweighting/resampling step. One possibility is generating an ensemble using direct transformations on each particle as in the EnKF. In the next section we present what we call the "nonlinear ensemble adjustment filter" (NLEAF) as a combination of the EnKF and the PF. Some relevant works [4, 8] also have the flavor of combining the EnKF and the PF, but both involve some form of resampling, which is typically undesirable in high-dimensional situations.

3 The NonLinear Ensemble Adjustment Filter (NLEAF)

3.1 A regression perspective of the EnKF

In equation (4), the Kalman gain K^T is simply the linear regression coefficient of X_f on Y. In fact, from Model (2) we have $\mathrm{Cov}(X, Y) = \Sigma_f H^T$ and $\mathrm{Var}(Y) = H\Sigma_f H^T + R$, therefore $K^T = \mathrm{Var}(Y)^{-1}\mathrm{Cov}(Y, X_f)$. The conditional expectation of X_f given y is

$$\mu_f + K(y - H\mu_f) := m_1(y).$$

Let $y^{(i)} = Hx_f^{(i)} + \epsilon^{(i)}$ be an observation given $X_f = x_f^{(i)}$ then $(x_f^{(i)}, y^{(i)})$ is a random sample from the joint distribution of (X_f, Y). $\hat{m}_1(\cdot) = \hat{\mu}_f + \hat{K}(\cdot - H\hat{\mu}_f)$ is an estimator of

$m_1(\cdot)$. The update step of the EnKF can be written as

$$x_u^{(i)} = \hat{m}_1(y) + x_f^{(i)} - \hat{m}_1(y^{(i)}). \tag{5}$$

Under Model (2) we have that $(X_f - m_1(y)|Y = y) \sim N(0, \Sigma_u)$ where Σ_u does not depend on y. Note further that $\big(x_f^{(i)}, y^{(i)}\big) \sim (X_f, Y)$, so $x_f^{(i)} - m_1(y^{(i)})$ is a random draw from $N(0, \Sigma_u)$. Therefore $x_u^{(i)} = m_1(y) + x_f^{(i)} - m_1(y^{(i)})$ is a random draw from $N(\mu_u, \Sigma_u)$ by noting that $m_1(y) = \mu_u$, which validates the update formula (5).

The procedure described above is an abstraction of the EnKF which can be viewed as a solution to the sampling problem of generating a random sample of $(X_f|Y = y)$ given a sample of X_f. Classical approaches to this problem includes rejective sampling and importance sampling (with possibly a resampling step). However, the approach described above uses direct transformations on the particles $x_f^{(i)}$, with randomness involved only in generating $y^{(i)}$. This procedure is effective in the sense that each particle in the forecast ensemble correspond to exactly one particle in the updated ensemble, without sample degeneracy.

3.2 A general NLEAF framework

Based on the discussion above, an effective way of updating the ensemble is directly transforming each particle so that the transformed particles have the desired distribution. In a Gaussian linear model, it suffices to adjust the mean by a simple shift as in equation (5) and the posterior variance is implicitly obtained by generating the random number $x_f^{(i)} - \hat{m}_1(y^{(i)})$. For general models where the likelihood function $g(y; x)$ and the forecast distribution are not Gaussian, it is too much to ask for the transformation to achieve the exact posterior distribution. Instead, it is more practical to achieve only the correct posterior moments. This simple idea leads to the NonLinear Ensemble Adjustment Filter (NLEAF).

NLEAF of order S

1. For $s = 1, \ldots, S$, where S is a pre-chosen positive integer, estimate the conditional sth moment $m_s(y) := E(X_f^s|y)$. Denote the estimates by $\hat{m}_s$, $s = 1, \ldots, S$.

2. For $i = 1, \ldots, n$, find $\xi_i(\cdot)$ such that

$$E(\xi_i^s(X_f)|y^{(i)}) \approx \hat{m}_s(y), \quad s = 1, \ldots, S,$$

   where the function $\xi_i(\cdot)$ might depend on y, $y^{(i)}$, $\{x_f^{(i)}\}_{i=1}^n$ and $\hat{m}_s$, $s = 1, \ldots, S$.

3. The updated particle is $x_u^{(i)} = \xi_i(x_f^{(i)})$. The updated mean is simply $\hat{\mu}_u = \hat{m}_1(y)$.

The conditional moments can be estimated using importance sampling as in Step 1-3 of the particle filter. For example, the conditional mean and covariance can be estimated as following:

$$\hat{m}_1(y) = \frac{\sum_{i=1}^n g(y; x_f^{(i)}) x_f^{(i)}}{\sum_{i=1}^n g(y; x_f^{(i)})}, \tag{6}$$

$$\hat{m}_2(y) = \frac{\sum_{i=1}^n g(y|x_f^{(i)})(x_f^{(i)} - \hat{m}_1(y))(x_f^{(i)} - \hat{m}_1(y))^T}{\sum_{i=1}^n g(y; x_f^{(i)})}. \tag{7}$$

If the likelihood $g(\cdot;\cdot)$ is not known explicitly (eg, y is generated by a black-box function), one may use regression methods to estimate the conditional moments. For example, the EnKF uses a linear regression of X_f on Y to find $\hat{m}_1(y)$. However, under general models, one might need more general methods, such as polynomial regressions, to avoid serious bias. This idea is further explained in Section 4.2.

When $s > 2$, estimating the higher order conditional moments will require a large ensemble and the estimates are very sensitive to outliers. Moreover, there is no clear choice of ξ when $s \geq 3$. For the rest of this paper we focus on the simple cases $S = 1$ and $S = 2$.

3.2.1 The first order NLEAF ($S = 1$)

The first order NLEAF algorithm is a direct generalization of the EnKF:

1. Generate $y^{(i)} \sim g(\cdot; x_f^{(i)})$, for $i = 1, \ldots, n$.

2. Estimate $m_1(\cdot)$ by $\hat{m}_1(\cdot) = \frac{\sum_{i=1}^n x_f^{(i)} g(\cdot; x_f^{(i)})}{\sum_{i=1}^n g(\cdot; x_f^{(i)})}$.

3. Updated mean $\hat{\mu}_u = \hat{m}_1(y)$. Updated particle $x_u^{(i)} = \hat{m}_1(y) + x_f^{(i)} - \hat{m}_1(y^{(i)})$.

This approach is valid if $\mathcal{L}\left(X_f - m_1(y^{(i)})|y^{(i)}\right) \approx \mathcal{L}\left(X_f - m_1(y)|y\right)$, where $\mathcal{L}(X)$ denotes the distribution of the random variable X. That is, $\mathcal{L}(X_f|y)$ depends on y mostly in terms of the mean. A simple example is the Gaussian linear model, where only the posterior mean depends on y. One can also expect such a situation when the likelihood $g(y; x)$ has a lighter tail than the forecast distribution X_f. To formalize, let

$$\eta = \sup_{y', y} \mathrm{TV}\left(\mathcal{L}(X_f - m_1(y')|y'), \mathcal{L}(X_f - m_1(y)|y)\right),$$

where $\mathrm{TV}(\mathcal{L}_1, \mathcal{L}_2) = \sup_A |P_{\mathcal{L}_1}(A) - P_{\mathcal{L}_2}(A)|$ denotes the total variation distance between two distributions $\mathcal{L}_1$ and $\mathcal{L}_2$. Then the smaller η is, the better is the approximation given by the first order NLEAF. To state a rigorous result, we need the following technical conditions on the likelihood function $g(x; y)$ which make the argument simple.

(A0) X_f has density function $f(\cdot) > 0$.

(A1) $0 < g(x; y) \leq M < \infty$ for all (x, y), $\sup_{x \in \mathbb{R}^p, y \in K} |xg(x; y)| \leq M_K < \infty$ for all compact $K \subset \mathbb{R}^q$.

(A2) For any compact set $K \subseteq \mathbb{R}^q$, there exists a measurable function $v_K(x)$, such that $E(v_K^2(X)) < \infty$ and for any $y_1, y_2 \in K$,

$$\max\left(|xg(x; y_1) - xg(x; y_2)|, |g(x; y_1) - g(x; y_2)|\right) \leq v_K(x)|y_1 - y_2|.$$

The conditions A1 and A2 are standard conditions for the maximal inequalities in empirical processes. They implies that the likelihood function $x \mapsto g(x; y)$ $(x \mapsto xg(x; y))$ depends on y continuously, which controls the complexity of the class of functions $x \mapsto g(x; y)$ $(x \mapsto xg(x; y))$ indexed by y and enables the use of the classical results of empirical processes. They also implies that the observation Y provides information for the whole vector of X, which precludes the degenerate situations such as $X = (X_1, X_2)^T$ and $Y = h(X_1)$. These conditions are reasonably general, including models like $g(x; y) \propto \phi(|x - y|)$ with $\phi(\cdot)$ decaying fast enough, e.g., for the Gaussian density function one can find the $v_K(x)$ is bounded by a constant. We have the following theorem whose proof is in Appendix A:

Theorem 1. *Suppose* $\left(x^{(i)}, y^{(i)}\right)$, $i = 1, \ldots, n$ *is an i.i.d sample from the joint distribution of* (X_f, Y). *Let* $x_u^{(i)}$, $i = 1, \ldots, n$, *be the updated particles given by the first order NLEAF algorithm. For any* y, *consider the empirical distribution*

$$\hat{F}_u(A|y) = \frac{1}{n}\delta_{x_u^{(i)}}(A), \quad \forall A,$$

where

$$x_u^{(i)} = \hat{m}_1(y) + x_f^{(i)} - \hat{m}_1(y^{(i)}).$$

Also let $F_u(A|y) = P(X_f \in A|y)$ *be the true conditional measure. Then, under (A0-A2) for Borel set* A *with* $\lambda(\partial A) = 0$, *we have*

$$\limsup_{n \to \infty} |\hat{F}_u(A|y) - F_u(A|y)| \leq \eta, \quad a.s.,$$

where $\lambda(\cdot)$ *is the Lebesgue measure and* $\partial A := \bar{A} \backslash A^\circ$ *is the boundary of* A, *with* $\bar{A}$ *and* A° *being the compact closure and interior of* A, *respectively.*

A by-product of the proof of Theorem 1 is the consistency of mean update:

Corollary 2. *Under (A0-A2), we have for any* y,

$$\hat{m}_1(y) \to m_1(y), \quad a.s., \quad n \to \infty.$$

Under the Gaussian linear model we have $\eta = 0$. The above results indicate the consistency of the NLEAF of order one:

Corollary 3. *Under Model (2), for any y,*

$$\hat{F}_u \xrightarrow{d} F_u, \quad a.s., \quad n \to \infty.$$

3.2.2 A second order NLEAF ($S = 2$)

Based on the estimated conditional variance in (7), one can easily develop a second order NLEAF algorithm. Now the function ξ_i is naturally chosen as

$$\xi_i(x) = \hat{m}_1(y) + (\hat{m}_2(y))^{\frac{1}{2}} \left(\hat{m}_2\left(y^{(i)}\right)\right)^{-\frac{1}{2}} \left(x - \hat{m}_1\left(y^{(i)}\right)\right). \tag{8}$$

The update formula is intuitively reasonable: Suppose $x, y \in \mathbb{R}^1$, then a large $m_2(y^{(i)})$ means that the region where $x_f^{(i)}$ lies in is highly uncertain which is possibly due to the irregular behavior of the dynamics in that region. Such a particle $x_f^{(i)}$ can provide little information on the true hidden state, therefore it is down-weighted in the transformation ξ_i, which tends to drag $x_f^{(i)}$ towards $\hat{\mu}_u = \hat{m}_1(y)$ in the updated ensemble.

It should be noted that the choice of ξ_i is apparently not unique. For example, for any orthogonal matrix U, one can define $\xi_i(x; U)$ as

$$\xi_i(x; U) = \hat{m}_1(y) + (\hat{m}_2(y))^{\frac{1}{2}} U \left(\hat{m}_2\left(y^{(i)}\right)\right)^{-\frac{1}{2}} \left(x - \hat{m}_1\left(y^{(i)}\right)\right).$$

It is easily seen that the choice of U does not change the first two moments of $\mathcal{L}\left(\xi_i(X_f; U)|y^{(i)}\right)$. The choice $U = I$ is natural in the sense that under Model (2) with $\Sigma_u = \sigma^2 I$, if $U = I$ then the second order NLEAF is asymptotically equivalent to the first order NLEAF, which is proved to be consistent (see also the discussion in [20]). In the rest of this paper, we will focus on the natural choice $U = I$.

3.2.3 Localization for the NLEAF algorithm

As seen above, the NLEAF algorithm is similar to the EnKF in that it updates each particle explicitly instead of resampling. As a result, one may expect a similar localization procedure as described in Section 2.4 applicable to the NLEAF algorithm. Recall that the EnKF localization involves three major steps:

a) Decompose the state vector X_f into local windows $X_f(N_l)$, $l = 1, \ldots, L$, find the corresponding local observation vector $Y(N_l')$, and the local likelihood function $g_l(y(N_l'); x_f(N_l))$;

b) Update each localized ensemble;

c) Construct the whole updated ensemble by combining the local updated ensembles.

In step a), one can usually construct a local window for each coordinate of X_f, where $X_f(N_j)$ is the subset of coordinates most relevant to $X_f(j)$, $j = 1, \ldots, p$. One can either choose these coordinates by subject knowledge. For example, in geophysics each coordinate corresponds to a physical location, then one can choose the coordinates in a neighborhood of the physical location of $X_f(j)$. Or one can use data-driven variable selection procedures to determine the relevant neighborhood $X_f(N_j)$. The choice of N'_j is similar. In many cases the special structure of the observation model (the second equation in (1)) enables natural and simple solutions. For example, under the linear model $Y = HX_f + \epsilon$, if H is sparse or banded, it is possible to find a submatrix $H_j = H(N'_j, N_j)$ such that $Y(N'_j) \approx H_j X_f(N_j) + \epsilon(N'_j)$, where N'_j is a subset of $1, \ldots, q$ such that $y_{N'_j}$ is the local observation corresponding to $X_f(N_j)$.

Once step a) is done, in step b) one only needs to apply the NLEAF algorithm as described above on each of the localized ensemble. The major issue is step c). Recall that the local windows overlap with each other, therefore each coordinate might be updated simultaneously in multiple local patches. To be concrete, for any local window $N_j \subseteq \{1, \ldots, p\}$, let $N'_j \subseteq \{1, \ldots, q\}$ be the corresponding local observation window, and $g_j(x_{N_j}; y_{N'_j})$ be the local likelihood function. Define N_k, N'_k and $g_k(\cdot;\cdot)$ similarly for another local window N_k. Suppose $r \in N_j \cap N_k$, then $X_f(r)$ is updated in both of these two local windows. From now on we consider the first order and second order NLEAF separately.

In the first order NLEAF, we write the update formula for the mean in both local windows as in equation (5):

$$\hat{\mu}_u(N_j) = \hat{m}_{1,j}(y(N'_j)),$$
$$\hat{\mu}_u(N_k) = \hat{m}_{1,k}(y(N'_k)),$$

where $\hat{m}_{1,j}(\cdot)$ denotes the local estimation of $m_{1,j}(\cdot) := E(X_f(N_j)|y(N'_j))$. Recall that we denote $(N_j; r)$ the position of the index r in the vector N_j. Then $\hat{\mu}_u(N_j; r)$ and $\hat{\mu}_u(N_k; r)$ can be viewed as predictions of $X_f(r)$ given different sets of predictors, namely $Y(N'_j)$ and $Y(N'_k)$, respectively. A natural method of combining the predictions of the same variable from different models is convex combination, which is chosen either conventionally or in a data-driven manner [6, 30, 7]. In our numerical experiment we follow the conventional choice described in [25] where the combination is simply averaging the updates in a few spatially coherent local windows. It is clear that this combination procedure is also applicable to the update of each single particle for exactly the same reason.

However, the above method of combining local updates does not apply directly to the second order NLEAF because in equation (8) the left-multiplication of the matrix $(\hat{m}_2(y))^{\frac{1}{2}}\left(\hat{m}_2(y^{(i)})\right)^{-\frac{1}{2}}$ mixes the coordinates in the local window, which makes coordinates in the left hand side no longer an estimate of the corresponding coordinate of the state variable, which invalidates the convex combination.

4 Numerical experiments

We present numerical experiments on two dynamical systems, both proposed by E. Lorenz in studying the predictability of chaotic systems. These systems have been widely used as test beds for atmospheric data assimilation methods [4, 25, 2].

4.1 Experiments on L63

The L63 system is first introduced by [23], as one of the earliest study of chaos. This three dimensional system is determined by an ordinary differential equation

$$\frac{dx(\tau)}{d\tau} = -\sigma x + \sigma y, \tag{9}$$

$$\frac{dy(\tau)}{d\tau} = -xz + rx - y, \tag{10}$$

$$\frac{dz(\tau)}{d\tau} = xy - bz, \tag{11}$$

where τ denotes the time, $(x(\tau), y(\tau), z(\tau))^T$ is the state vector and (b, σ, r) are parameters of the system. When $b = 8/3$, $r = 28$ and $\sigma = 10$, the system is chaotic and its orbit is the well-known *Butterfly Attractor*.

In the simulation the system is discretized using the fourth order Runge-Kutta method. It is clear that the linearity of the evolution of the state vector between two successive time points depends on the length of the time interval $\Delta\tau$ between t and $t+1$ which we call the *step size*: The smaller is $\Delta\tau$, the more linear is the evolution between t and $t+1$.

In the simulation, there is a hidden true orbit $\{x_t, t \geq 0\}$. The starting point, x_0, of the true orbit is randomly chosen from the attractor. At the starting time, an ensemble of state vectors $\{x_0^{(i)}\}_{i=1}^n$, surrounding x_0 is available (e.g., perturbations of x_0 with random noise or a random sample from a small neighborhood of x_0 in the attractor). For all $t > 0$ a noisy observation $y_t = x_t + \epsilon_t$ is available with

$$\epsilon_t \overset{iid}{\sim} N(0, \sigma^2 I_3). \tag{12}$$

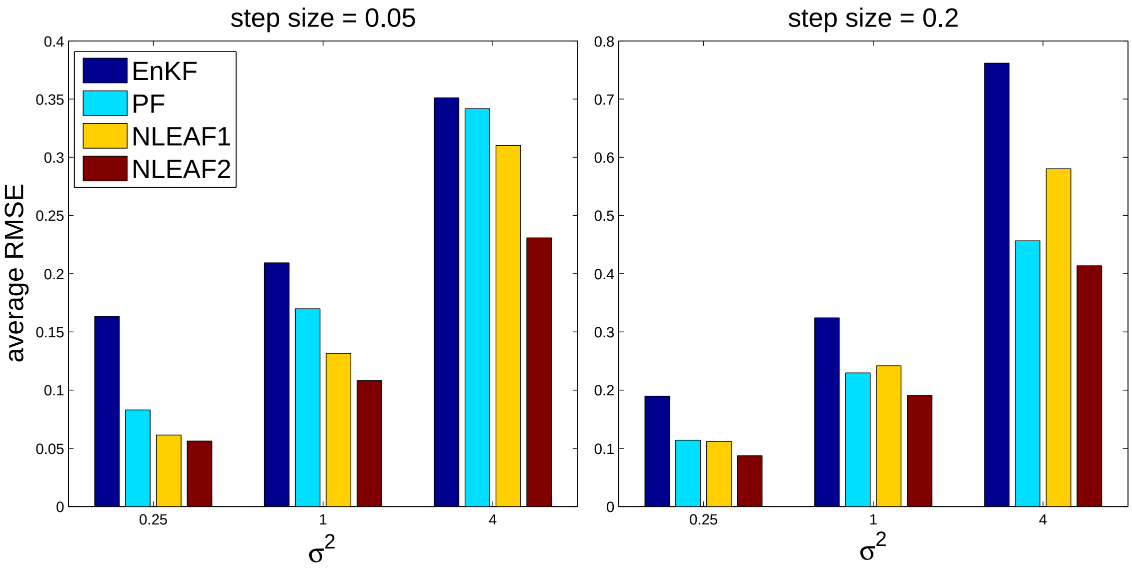

Figure 1: Average RMSE over 2000 cycles.

At each time $t \geq 1$, The updated ensemble average is used as the best single estimate of x_tTherefore, the data assimilation performance is evaluated by the root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{p}||\hat{\mu}_{u,t} - x_t||_2^2}. \tag{13}$$

We consider two time steps: 0.05 and 0.2, corresponding to the nearly linear case and the non-linear case respectively. In each case the system is propagated 2000 steps and at each time the data assimilation is performed using four different methods: the EnKF, the PF, the first order NLEAF (NLEAF1) and the second order NLEAF (NLEAF2), each with an ensemble of size 400. Also we consider three values of σ^2 in (12): 0.25, 1 and 4, corresponding to different levels of the observation accuracy. In Figure (1) we see that the EnKF gives the largest RMSE because of the non-linear dynamics. The NLEAF2 performs the best under all circumstances considered here. When step size is small, the system is nearly linear so that the NLEAF1 performs better than the PF. When the step size is large, the PF shows some advantage against the NLEAF1 which ignores the higher order moments.

4.2 Experiments on L96

The L96 system is introduced in [24] in the study of predictability of high dimensional chaotic systems. The state vector is 40 dimensional, and the dynamics is given by an ODE as follows:

$$\frac{dx_j(t)}{dt} = (x_{j+1} - x_{j-2})x_{j-1} - x_j + 8, \quad \text{for } j = 1, \ldots, 40, \tag{14}$$

where $x_0 = x_{40}$, $x_{-1} = x_{39}$ and $x_{41} = x_1$. This system mimics the evolution of some meteorological quantity at 40 equally spaced grid points along a latitude circle. The system is discretized with a time step of $\Delta\tau = 0.05$, which is analogous to a 6 hour in the real world.

Although the dimensionality of the L96 system is still far from the reality, it has been challenging for many standard data assimilation methods including the Kalman filter variants. Among the vast literature, we mention only two previous works: [25] considered the localized ensemble Kalman filter in an approximately linear case ($\delta\tau = 0.05$) and a complete observation, that is

$$Y_t = X_t + \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} N(0, I_{40}). \tag{15}$$

We call this set-up the *easy case.* On the other hand, [4] studied a localized Gaussian mixture filter in a highly non-linear case ($\delta\tau = 0.4$) and an incomplete observation: for $j = 1, \ldots, 20$,

$$Y_t(j) = X_t(2j-1) + \epsilon_t(j), \quad \epsilon_t \overset{iid}{\sim} N(0, I_{20}/2). \tag{16}$$

We call this set-up the *hard case.*

The major criterion is still the RMSE defined in (13). Moreover, because of its dimensionality and resemblance to real atmospheric data, we do care about the computation, where the main restriction is the ensemble size.

We consider both the easy case and the hard case. The system is propagated 2000 steps from a random starting point with data assimilation performed at each step. Because of the localization, we do not use the second order NLEAF. Instead, we use a variant of NLEAF1, namely NLEAF1q, with the letter “q” for “quadratic”, in which the function $m_1(\cdot) = E(X_f|Y = \cdot)$ is estimated using a quadratic regression of X_f on Y. To be concrete, in the NLEAF1q algorithm $\hat{m}_1(\cdot)$ is the minimizer over all quadratic functions $m(\cdot)$ of the square loss:

$$\sum_{i=1}^{n} \left(m(y^{(i)}) - x_f^{(i)} \right)^2 .$$

We consider the NLEAF1q algorithm because we believe sometimes $g(\cdot, \cdot)$ may not be available explicitly and the y's are generated by a black-box function of x. We emphasize that in the NLEAF1q algorithm, the function $g(\cdot, \cdot)$ is pretended to be unknown and not used.

In both NLEAF1 and NLEAF1q, the localization is as described in Section 2.2, which is also essentially the same as in [25]: Let l be a pre-chosen window size. For each $j = 1, \ldots, 40$, let $N_j = (j - l, \ldots, j, \ldots, j + l)$ be the local window centered at j. The corresponding local observation window N_j' is the local observations of $X(N_j)$. For example, if $l = 2$, then $N_1 = (39, 40, 1, 2, 3)$. In the easy case, $N_1' = (39, 40, 1, 2, 3)$ since the observation is complete (eq. (15)); In the hard case the observation is incomplete (eq. (16)) and we have

Table 2: The RMSE over 2000 assimilation cycles in the hard case. Ensemble size = 400.

NLEAF			NLEAFq			EnKF			XEnsF		
mean	med	std	mean	med	std	mean	med	std	mean	med	std
0.65	0.63	0.20	0.71	0.67	0.22	0.83	0.75	0.31	0.92	0.85	0.31

$N'_1 = (20, 1, 2)$.For each j, the coordinate $X(j)$ of the state variable X is updated in $2l + 1$ local windows. In the first order NLEAF algorithm, for $k \in N_j$, $X(j)$ is updated in the local window N_k using the conditional expectation given $y_{N'_k}$ (or $y^{(i)}_{N'_k}$). Similar to the scheme proposed in [25], we combine the local updates of $X_f(j)$ from N_{j-1}, N_j and N_{j+1} by simply averaging them. One can also use a data-driven method at a higher computational cost ([6, 30, 7]).

4.2.1 The hard case

In the hard case we compare four methods: the NLEAF1; the NLEAF1q; the mixture ensemble filter (XEnsF [4]); the EnKF without localization. Following the set-up in [4], the ensemble size is fixed to be 400. We compare the performance of NLEAF1 and NLEAF1q directly with those reported in [4], summarized in Table 2, where we see similar results as in the L63 experiment: The NLEAF1 gives much smaller RMSE than both the XEnsF and the EnKF. This is the first time the authors see the average RMSE goes below 0.7 in this set-up.

4.2.2 The easy case

In the easy case we compare three methods: the NLEAF1, the NLEAF1q and the ensemble transform Kalman filter (LETKF) proposed in [25], which achieves the best known performance in this set-up, with an average RMSE of about 0.2 using an ensemble as small as 10. It is reported that enlarging the ensemble size does not improve the accuracy of EnKF (LETKF) while the NLEAF is expected to work better for larger ensembles. Here we consider different ensemble sizes ranging from 10 to 400. The result is summarized in Figure 2 where only the mean of the average RMSE is plotted. The median and the variance are qualitatively similar to those presented in the hard case and are omitted here. We see that the LETKF still gives the best performance especially for small ensemble sizes. The NLEAF1 becomes competitive when the ensemble size is moderately large. From the plot it is also reasonable to expect even smaller RMSE of NLEAF1 given even larger ensembles.

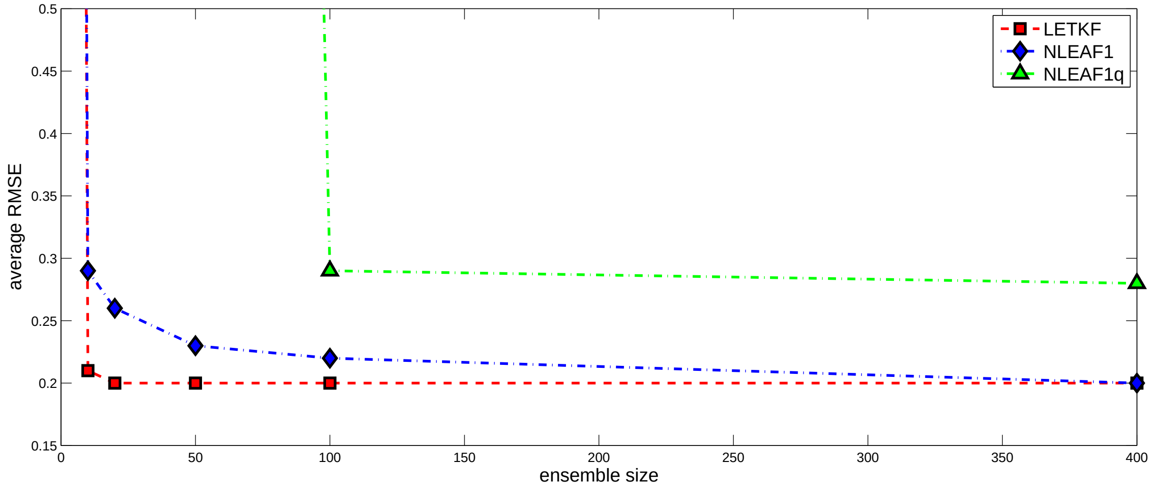

Figure 2: Average RMSE over 2000 cycles in the easy case of L96 system, ensemble size = 400.

The performance of NLEAF1q is not as good as the other two methods but we believe it is of practical interest since it requires much less *a priori* knowledge on the observation mechanism.

4.2.3 An intermediate case

So far both the easy and the hard cases are of practical interests: The easy case is analogous to the 6-hour operational data assimilation; The hard case challenges forecast in the presence of high nonlinearity and incomplete observation which is often the case in practice. As a result, it would be interesting to consider an *intermediate case* where the time step is still short as in the easy case but the observation is incomplete as in the hard case, with a larger observation noise:

$$Y_t(j) = X_t(2j-1) + \epsilon_t(j), \quad \epsilon_t \overset{iid}{\sim} N(0, 2I_{20}). \tag{17}$$

Again we let the ensemble size vary from 10 to 400. The results are summarized in Figure 3. Now the NLEAF1 and NLEAF1q gives much better relative results than in the easy case. The NLEAF1 is competitive for a ensemble as large as 100. Here again we see the potentiality of improvement for the NLEAF1 when the ensemble gets large. The NLEAF1q algorithm does a decent job for large ensembles too.

It should be noted that the LETKF tends to lose accuracy when the ensemble size gets beyond 20. There are two possible reasons for this phenomenon: first, the method of combining updates in different local windows might not be optimal for this set-up in varying ensemble sizes; second, the the mis-specification of the linear model assumed by the ensemble

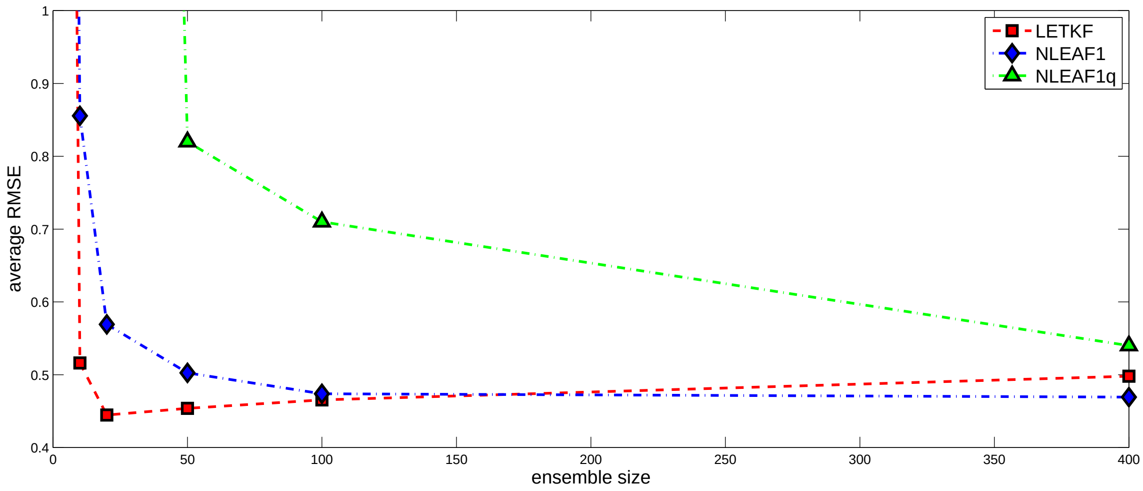

Figure 3: Average RMSE over 2000 cycles in the intermediate case of L96 system, ensemble size = 400.

Kalman filter incurs a larger bias when the ensemble size gets large.

5 Conclusion

As the increasing availability of both sophisticated climate models and massive sequential data, scientific applications such as numerical weather forecasting pose new challenges on statistical inference on high-dimensional nonlinear state space model. The proposed NLEAF algorithm is a combination of the traditional ensemble Kalman filter and particle filter which is adaptive to the nonlinearity of the dynamics and also easily scalable to high-dimensional situations. In two classical test beds for atmospheric data assimilation, very simple NLEAF algorithms give reasonably good performances. They outperforms the state-of-art methods in the nonlinear set-up, while still being competitive even in the linear situation where the EnKF is expected to be nearly optimal. We also observe that the NLEAF algorithm has the potential to improve its accuracy for larger ensembles, while the EnKF does not. Furthermore, the NLEAF algorithm is flexible and allows the observation model to be unknown and estimated from the data, which makes itself more applicable for many real world problems where the observation error can hardly be specified *a priori*.

There are still issues to be addressed. For example, the localization for NLEAF of order two or higher will be useful since we observed a substantial improvement of accuracy by NLEAF2 in the L63 system. A further question is that whether the NLEAF algorithm can be used in combination with other dimension reduction methods such as manifold learning and regularization.

A Proofs

A.1 Proof of Theorem 1

Suppose $(x_f^{(i)}, y^{(i)})$, $i = 1, \ldots, n$ is an i.i.d sample from the joint distribution of (X_f, Y), For any y, consider the empirical distribution

$$F_u^*(A|y) = \frac{1}{n}\delta_{x_u^{*(i)}}(A), \quad \forall A,$$

with

$$x_u^{*(i)} = m_1(y) + x_f^{(i)} - m_1(y^{(i)}).$$

Note that the NLEAF update in equation (5) uses $\hat{m}_1(\cdot)$ instead of $m_1(\cdot)$. The rough idea is that if $\hat{m}_1(\cdot)$ approximates $m_1(\cdot)$ well enough, one might expect $x_u^{(i)} \approx x_u^{*(i)}$ and the result follows from Hoeffding's inequality. To show that $x_u^{(i)}$ does approximates $x_u^{*(i)}$ we use the empirical process theory. The maximal inequality of the empirical process requires the majority of $y^{(i)}$ lies in a compact set, which is of high probability if the compact set is large enough.

For any $0 < \epsilon < 1$, one can find a compact set $K(\epsilon)$ such that $P(Y \in K) \geq 1 - \epsilon$. Define the set J as

$$J = \{i : y^{(i)} \in K(\epsilon)\}.$$

Consider the event

$$E_1 = \left\{\frac{|J|}{n} \geq 1 - 2\epsilon\right\},$$

then we have, by Hoeffding's inequality,

$$P(E_1) \geq 1 - \exp\left(-2n\epsilon^2\right). \tag{18}$$

Let $B(\epsilon) = \inf_{y \in K(\epsilon)} \int g(y; x) f(x) dx > 0$. Consider the events

$$E_2 = \left\{\sup_{y \in K(\epsilon)} \left|\frac{1}{n}\sum_{i=1}^{n} g(y; x_f^{(i)}) - \int g(y; x) f(x) dx\right| \leq \min\left(\frac{B(\epsilon)}{2}, \frac{B^2(\epsilon)}{8M(\epsilon)}\right)\right\},$$

where $M(\epsilon) = M_{K(\epsilon)}$ as defined in Assumption A1, and

$$E_3 = \left\{\sup_{y \in K(\epsilon)} \left|\frac{1}{n}\sum_{i=1}^{n} x_f^{(i)} g(y; x_f^{(i)}) - \int x g(y; x) f(x) dx\right| \leq \frac{B(\epsilon)\epsilon}{8}\right\}.$$

By assumption A1 and A2 and the maximal inequality of empirical process [28, 26], there exist functions $c_i(\epsilon)$, $i = 1, 2$, such that

$$P(E_2^C) \leq c_1(\epsilon) n^{q-1} \exp\left(-n c_2(\epsilon)\right),$$

and

$$P(E_3^C) \leq c_1(\epsilon) n^{q-1} \exp\left(-n c_2(\epsilon)\right).$$

Note that on $E_2 \bigcap E_3$, we have $|\hat{m}_1(y) - m_1(y)| \leq \epsilon/2$, for all $y \in K(\epsilon)$. As a result, on $E_2 \bigcap E_3$, we have,

$$|x_u^{(i)} - x_u^{*(i)}| \leq \epsilon, \quad \forall i \in J.$$

Then we have, on $E_1 \bigcap E_2 \bigcap E_3$,

$$\begin{aligned} \hat{F}_u(A|y) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_A(x_u^{(i)}) &\geq \frac{1}{n}\sum_{i\in J} \mathbb{1}_A(x_u^{(i)}) \\ &\geq \frac{1}{n}\sum_{i\in J} \mathbb{1}_{A_\epsilon^-}(x_u^{*(i)}) \\ &\geq \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{A_\epsilon^-}(x_u^{*(i)}) - \frac{|J^C|}{n} \\ &\geq \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{A_\epsilon^-}(x_u^{*(i)}) - 2\epsilon, \end{aligned}$$

where the set A_ϵ^- is defined as

$$A_\epsilon^- = \{x \in A : D(x, \epsilon) \subseteq A\},$$

with $D(x, \epsilon)$ being the ϵ-open ball centering at x.

Consider event E_4:

$$E_4 = \left\{ \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{A_\epsilon^-}(x_u^{*(i)}) \geq F_u(A_\epsilon^-|y) - \eta - \epsilon \right\}.$$

Again, note that $\mathbb{1}_{A_\epsilon^-}(x_u^{*(i)})$ are independent Bernoulli random variables with probability at least $F_u(A_\epsilon^-|y) - \eta$, by Hoeffding's inequality, we have

$$P(E_4) \geq 1 - \exp(-2n\epsilon^2).$$

Then on $\bigcap_{k=1}^4 E_k$, we have

$$\begin{aligned} \hat{F}_u(A|y) - F(A|y) &\geq F(A_\epsilon^-|y) - F(A|y) - \eta - 3\epsilon \\ &= -\eta - \rho^-(\epsilon) - 3\epsilon, \end{aligned}$$

where $\rho^-(\epsilon) = F(A|y) - F(A_\epsilon^-|y)$ is a continuous non-decreasing function of ϵ with $\rho^-(0) = 0$ because $\lambda(\partial A) = 0$. As a result, there exists functions $C_1(\epsilon) > 0$, $C_2(\epsilon) > 0$ independent of n, such that

$$P\left(\hat{F}_u(A|y) - F_u(A|y) \geq -\eta - \epsilon\right) \geq 1 - C_1(\epsilon)n^{q-1}\exp\left(-C_2(\epsilon)n\right).$$

A similar bound for the other direction can be obtained using the same argument. By the Borel-Cantelli lemma we have,

$$\left|\hat{F}_u(A|y) - F_u(A|y)\right| \leq \eta + \epsilon, \quad \text{a.s.}$$

Note that the above convergence is for any $\epsilon > 0$, therefore we have

$$\left|\hat{F}_u(A|y) - F_u(A|y)\right| \leq \eta, \quad \text{a.s.}$$

References

[1] J. Anderson. An ensemble adjustment kalman filter for data assimilation. *Monthly Weather Review*, 129:2884–2903, 2001.

[2] J. L. Anderson. Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter. *Physica D*, 230:99–111, 2007.

[3] T. Bengtsson, P. Bickel, and B. Li. Curse of dimensionality revisited: the collapse of importance sampling in very large scale systems. *IMS Collections: Probability and Statistics*, 2:316–334, 2008.

[4] T. Bengtsson, C. Snyder, and D. Nychka. Toward a nonlinear ensemble filter for high-dimensional systems : Application of recent advances in space-time statistics to atmospheric data. *J. Geophys. Res.*, 108(D24):STS2.1–STS2.10, 2003.

[5] C. H. Bishop, B. Etherton, and S. J. Majumdar. Adaptive sampling with the ensemble transformation kalman filter. part i: theoretical aspects. *Monthly Weather Review*, 129:420–436, 2001.

[6] L. Breiman. Stacked regressions. *Machine Learning*, 24:49–64, 1996.

[7] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

[8] A. Chorin and X. Tu. Non-Bayesian particle filters. 2009.

[9] G. Evensen. Sequential data assimilation with a non-linear quasi-geostrophic model using monte carlo methods to forecast error statistics. *J. Geophys. Res.*, 99(C5):10143–10162, 1994.

[10] G. Evensen. The ensemble kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53:343–367, 2003.

[11] G. Evensen. *Data assimilation: the ensemble Kalman filter*. Springer, 2007.

[12] J. Fan and Q. Yao. *Nonlinear time series: nonparametric and parametric methods.* Springer, 2003.

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

[14] N. Gordon, D. Salmon, and A. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F*, 140:107–113, 1993.

[15] P. L. Houtekamer and H. L. Mitchell. Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, 126:796–811, 1998.

[16] S. J. Julier and J. K. Uhlmann. A new extension of the Kalman filter to nonlinear systems. In *Proc. of AeroSense: The 11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls*, volume Multi Sensor Fusion, Tracking and Resource Management II, Orlando, Florida, 1997.

[17] H. R. Künsch. State space and hidden Markov models. In O. E. Barndorff-Nielsen, D. R. Cox, and C. Klüppelberg, editors, *Complex Stochastic Systems*, pages 109–173. Chapman and Hall, 2001.

[18] H. R. Künsch. Recursive Monte Carlo filters: algorithms and theoretical analysis. *The Annals of Statistics*, 33:1983–2021, 2005.

[19] F. Le Gland, V. Monbet, and V. Tran. Large sample asymptotics for the ensemble Kalman filter. 2009.

[20] J. Lei, P. Bickel, and C. Snyder. Comparison of ensemble Kalman filters under non-Gaussianity. *submitted to Monthly Weather Review*, 2009.

[21] J. Liu. *Monte Carlo strategies in scientific computing.* Springer, 2001.

[22] J. Liu and R. Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044, 1998.

[23] E. N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20:130–141, 1963.

[24] E. N. Lorenz. Predictability: a problem partly solved. In *Proc. Seminar on Predictability*, volume 1, Shinfield Park, Reading, Berkshire, United Kingdom, 1996. European Centre for Medium-Range Weather Forecast.

[25] E. Ott, B. Hunt, I. Szunyogh, A. Zimin, E. Kostelich, M. Corazza, E. Kalnay, D. Patil, and J. Yorke. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus*, 56A:415–428, 2004.

[26] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *The Annals of Probability*, 22:28–76, 1994.

[27] M. K. Tippett, J. L. Anderson, C. H. Bishop, T. M. Hamill, and J. S. Whitaker. Ensemble square root filters. *Monthly Weather Review*, 131:1485–1490, 2003.

[28] A. W. van der Vaart. *Asymptotic Statistics*, chapter 19. Cambridge University Press, 2001.

[29] J. S. Whitaker and T. M. Hamill. Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, 130:1913–1924, 2002.

[30] Y. Yang. Adaptive regression by mixing. *Journal of the American Statistical Association*, 96(454):574–588, 2001.