Particle filters, the "optimal" proposal and high-dimensional systems

Chris Snyder

*National Center for Atmospheric Research**
Boulder, Colorado 80307, United States
chriss@ucar.edu

1 Introduction

Particle filters are an appealing class of data-assimilation techniques because of their simplicity and their generality for nonlinear and non-Gaussian problems. They have seen wide use in fields outside atmospheric science and oceanography (see Doucet et al. 2001). There have also been initial efforts for applications in oceanography (van Leeuwen 2003) and land-surface modeling (Zhou et al. 2006). Further introduction to and background on particle filters can be found in Gordon et al. (1993), Doucet et al. (2001), Arulampalam et al. (2002), van Leeuwen (2009) and Boucquet et al (2010).

Particle filters are a sequential Monte-Carlo technique and can be shown to produce an ensemble of states drawn from the correct posterior distribution as the ensemble size $N_e \rightarrow \infty$. (See Doucet 1998 for a rigorous demonstration.) If there were sufficient computing to run ensemble forecasts with arbitrarily large N_e, then implementation of particle filters would be straightforward and the results general. In practice, the ensemble size is limited, and we expect that more members will be necessary as the system dimension increases. All applications of the particle filter so far, however, involve systems whose state space is either explicitly or effectively low dimensional. Thus, the effectiveness of particle filters for high-dimensional systems remains an open question.

For the simplest "bootstrap" particle filter, in which each member (or particle) is evolved under the system dynamics and assigned a weight proportional to the likelihood of the new observations given that member. When N_e is too small, one of the weights will approach unity and there will be a corresponding deterioration of the performance of the filter. Snyder et al. (2008; also Bengtsson et al. 2008, Bickel et al. 2008) demonstrate, under reasonably general assumptions, that avoiding such degeneracy requires N_e to grow exponentially with τ^2, the variance of the total log likelihood of the observations given the state. Though τ^2 need not be directly related to the system dimension, it can be expected to grow with the system dimension in the most interesting case, namely when the number of degrees of freedom with significant variance also grows with the system dimension.

Sequential importance sampling is often employed in particle filters (Doucet et al. 2000) but was not included in the analysis of Snyder et al. (2008). Rather than using evolution under the system dynamics to generate ensemble members at later times, as in the bootstrap filter, one has the freedom to chose the distribution from which later members are drawn, as long as this is accompanied by an appropriate modification of the weights assigned to each member. The distribution from which new members are drawn is known as the proposal distribution.

The present paper reviews sequential importance sampling and the so-called "optimal" proposal, which incorporates the new observations at time t_k when generating the ensemble members valid at t_k. (It is worth emphasizing from the outset the potential for confusion in the terminology "optimal proposal." The optimal proposal is is optimal only in a very restricted sense.) In the geophysical literature, several

variants of particle filters have been proposed that capitalize on the idea of using the new observations when generating members (van Leeuwen 2010, Papadakis et al. 2010, Morzfeld et al. 2011). The resulting algorithms either approximate the optimal proposal distribution (Papadakis et al. 2010), reduce to it in specific cases (Morzfeld et al. 2011) or are closely related (van Leeuwen 2010). I present a simple example showing that the optimal proposal can greatly outperform simply evolving the existing ensemble members to t_k under the system dynamics (which I will term the standard proposal). The arguments of Snyder et al. (2008) can be applied to the optimal proposal and can quantify these benefits. Nevertheless, the same arguments also demonstrate that the required ensemble size again grows exponentially for the optimal proposal, in this case as exponentially in ω^2, the variance of the log likelihood of the observations given the state at the *previous* time.

Particle-filter algorithms typically also include a resampling step, after the update of each member's weight. Given the latest members and updated weights, a new ensemble is drawn from the distribution implied by the members and their weights (or a smooth estimate of it). The result is a new ensemble with uniform weights and members replicating or clustered around locations where the updated weights were formerly large. Since they are based on the empirical distribution implied by the updated weights, resampling schemes cannot repair deficiencies in the those weights and will not be discussed further until the concluding section.

2 Background on particle filters and sequential importance sampling

2.1 Basics of particle filters

Let $\mathbf{x}_k = \mathbf{x}(t_k)$ of dimension N_x be the state of the system at time t_k. Suppose that $\mathbf{x}$ evolves according to

$$\mathbf{x}_k = \mathbf{f}(\mathbf{x}_{k-1}, \eta_{k-1}), \tag{1}$$

and that the observation $\mathbf{y}_k = \mathbf{y}(t_k)$ of dimension N_y is related to the state by

$$\mathbf{y}_k = \mathbf{h}(\mathbf{x}_k) + \varepsilon_k, \tag{2}$$

where η_k and ε_k are each i.i.d. random variables for all k and, for simplicity, are mutually independent. It will be convenient to denote with the subscript $j{:}k$ the vector obtained by concatenating the observation vectors at times $t_j, \ldots, t_k$; for example, $\mathbf{y}_{j:k}$ is the concatenation of observation vectors at times $t_j, \ldots, t_k$.

Formally, our goal is to calculate the filtering density (that is, the conditional probability density of $\mathbf{x}_k$ given observations $\mathbf{y}_0, \ldots, \mathbf{y}_k$ up to time t_k) via Bayes rule,

$$p(\mathbf{x}_k|\mathbf{y}_{0:k}) = \frac{p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{y}_{0:k-1})}{p(\mathbf{y}_k|\mathbf{y}_{0:k-1})} \tag{3}$$

Since all pdfs in what follows will be conditioned on $\mathbf{y}_{0:k-1}$, we will omit those observations in the sequel when writing the condition statements.

Direct computation with the pdfs in (3) is not feasible unless N_x and N_y are small or the problem has other simplifying structure. Particle filters are Monte-Carlo methods that instead work with finite, weighted ensembles $\{\mathbf{x}_k^i, w_k^i; i = 1, \ldots, N_e\}$ that approximate the desired $p(\mathbf{x}_k|\mathbf{y}_k)$ in the sense that a weighted sum over the ensemble members converges to an expectation with respect to $p(\mathbf{x}_k|\mathbf{y}_k)$:

$$\sum_{i=1}^{N_e} w_k^i g(\mathbf{x}_k^i) \rightarrow \int g(\mathbf{x}_k)\, p(\mathbf{x}_k|\mathbf{y}_k)\, d\mathbf{x}_k \quad \text{as } N_e \rightarrow \infty. \tag{4}$$

2.2 Sequential importance sampling

Particle-filter algorithms may be cast in terms of sequential importance sampling. To understand importance sampling, suppose we are given two distributions $p(\mathbf{x})$ and $\pi(\mathbf{x})$, together with a random sample $\{\mathbf{x}^i, i = 1,\ldots,N_e\}$ from $\pi(\mathbf{x})$. Defining weights $w^i = p(\mathbf{x}^i)/\pi(\mathbf{x}^i)$, the weighted sample $\{\mathbf{x}^i, w^i\}$ then approximates $p(\mathbf{x})$ in the sense of (4). This is useful if $p(\mathbf{x})$ is difficult to sample from while $\pi(\mathbf{x})$ is easy. The distributions $p(\mathbf{x})$ and $\pi(\mathbf{x})$ are known as the *target* and the *proposal* respectively. See Doucet (1998) and Arulampalam et al. (2002) for further introduction sequential importance sampling.

Sequential importance sampling proceeds sequentially in time, computing at each time a weighted sample $\{\mathbf{x}^i_k, w^i_k\}$ that approximates $p(\mathbf{x}_k|\mathbf{y}_k)$. It is sufficient to consider the step from t_{k-1} to t_k, in which we start from $\{\mathbf{x}^i_{k-1}, w^i_{k-1}\}$ and compute $\{\mathbf{x}^i_k, w^i_k\}$. The idea is to apply importance sampling to the joint conditional distribution $p(\mathbf{x}_{k-1}, \mathbf{x}_k|\mathbf{y}_k)$.

We choose a proposal density of the form

$$\pi(\mathbf{x}_{k-1}, \mathbf{x}_k|\mathbf{y}_k) = \pi(\mathbf{x}_{k-1})\pi(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{y}_k), \tag{5}$$

where $\pi(\mathbf{x}_{k-1})$ is the proposal density from t_{k-1} and $\pi(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{y}_k)$ remains to be specified. Sampling from $\pi(\mathbf{x}_{k-1}, \mathbf{x}_k|\mathbf{y}_k)$, when it has the form (5), can be achieved simply by drawing $\mathbf{x}^i_k$ from $\pi(\mathbf{x}_k|\mathbf{x}^i_{k-1}, \mathbf{y}_k)$; the pair $(\mathbf{x}^i_{k-1}, \mathbf{x}^i_k)$ is a therefore random draw from $\pi(\mathbf{x}_{k-1}, \mathbf{x}_k|\mathbf{y}_k)$ since we have assumed $\mathbf{x}^i_{k-1}$ was drawn from $\pi(\mathbf{x}_{k-1})$.

The corresponding importance weight is given by

$$w^i_k \propto \frac{p(\mathbf{x}^i_{k-1}, \mathbf{x}^i_k|\mathbf{y}_k)}{\pi(\mathbf{x}^i_{k-1}, \mathbf{x}^i_k|\mathbf{y}_k)} = \frac{p(\mathbf{x}^i_{k-1}, \mathbf{x}^i_k|\mathbf{y}_k)}{\pi(\mathbf{x}^i_{k-1})\pi(\mathbf{x}^i_k|\mathbf{x}^i_{k-1}, \mathbf{y}_k)}. \tag{6}$$

The constant of proportionality is determined by requiring that $\sum w^i_k = 1$. Applying Bayes rule and the definition of a condtional density, the joint conditional density may be written as

$$p(\mathbf{x}_{k-1}, \mathbf{x}_k|\mathbf{y}_k) = p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_{k-1}, \mathbf{x}_k) = p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_{k-1})p(\mathbf{x}_{k-1}). \tag{7}$$

Substituting (7) in (6) yields

$$w^i_k \propto \frac{p(\mathbf{y}_k|\mathbf{x}^i_k)p(\mathbf{x}^i_k|\mathbf{x}^i_{k-1})}{\pi(\mathbf{x}^i_k|\mathbf{x}^i_{k-1}, \mathbf{y}_k)} w^i_{k-1}, \tag{8}$$

since $w^i_{k-1} = p(\mathbf{x}^i_{k-1})/\pi(\mathbf{x}^i_{k-1})$.

Sequential importance sampling at t_k thus proceeds in two steps: First, drawing a random sample $\{\mathbf{x}^i_{k-1}\}$ from the chosen proposal distribution and, second, updating the weight w^i_{k-1} associated with each member, or *particle*, via (8). The update of the weights is sequential in the sense that it uses only $\mathbf{y}_k$, $\mathbf{x}^i_{k-1}$ and w^i_{k-1}, and no information from times earlier than t_{k-1}. This simplification follows from the assumptions that the dynamics (1) is Markovian (that is, $\mathbf{x}_k$ depends only on $\mathbf{x}_{k-1}$) and the observation equation (2) at t_k depends only on $\mathbf{x}_k$.

The performance of importance sampling depends strongly on the choice of the proposal distribution. Frequently, the transition distribution for the dynamics is used,

$$\pi(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{y}_k) = p(\mathbf{x}_k|\mathbf{x}_{k-1}). \tag{9}$$

The particles $\mathbf{x}^i_k$ are a random draw from the prior distribution $p(\mathbf{x}_k)$ and can be generated simply by evolving $\mathbf{x}^i_{k-1}$ forward under the system dynamics. The weights are then updated by multiplying by the observation likelihood for each member,

$$w^i_k = p(\mathbf{y}_k|\mathbf{x}^i_k)w^i_{k-1}. \tag{10}$$

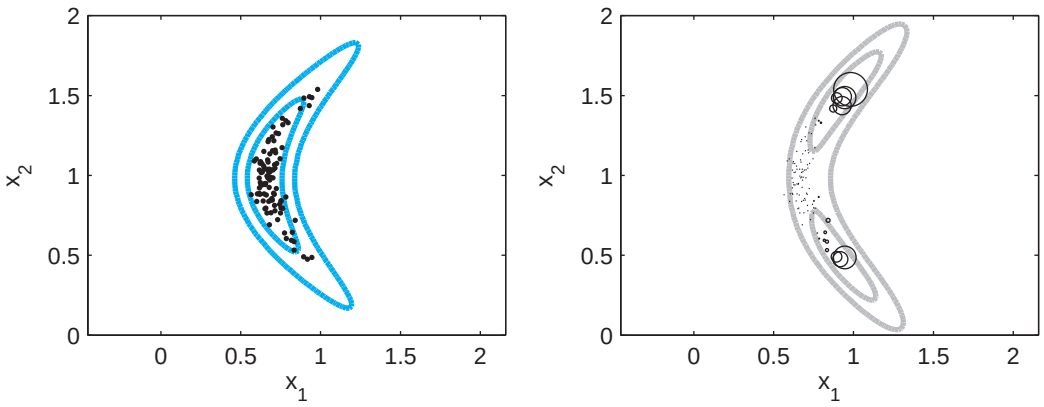

Figure 1: A two-dimensional example of importance sampling using the proposal (10). (a) The prior pdf (the proposal; contours) and a sample from that pdf (dots). (b) The posterior pdf (contours) given an observation $y = x_1 + \varepsilon = 1.1$, where $\varepsilon \sim N(0,1)$, and the weighted particles, shown as circles with radius proportional to $p(y|\mathbf{x})$.

This is the case considered by Snyder et al. (2008). I will term (9) the *standard* proposal.

Figure 1 illustrates how importance sampling works in a two-dimensional example and using the standard proposal. The prior pdf (the proposal) and a sample of size 100 from that proposal are shown in Fig. 1a. (In a slight abuse of previous notation, x_1 and x_2 are the two components of the state vector $\mathbf{x}$.) We are given a observation $y = x_1 + \varepsilon = 1.1$ of the first component contaminated by Gaussian noise of zero mean and standard deviation 1. The posterior pdf is shown in Fig. 1b, together with circles at the locations of the particles and having radius proportional to $p(y|\mathbf{x})$. Although the ensemble was drawn from the proposal distribution, it represents the posterior well after weighting by $p(y|\mathbf{x})$, since the posterior is proportional to the product of the proposal (9) and $p(y|\mathbf{x})$.

Another possible choice for the proposal distribution is

$$\pi(\mathbf{x}_k|\mathbf{x}_{k-1},\mathbf{y}_k) = p(\mathbf{x}_k|\mathbf{x}_{k-1},\mathbf{y}_k), \tag{11}$$

with weights updated according to

$$w_k^i = p(\mathbf{y}_k|\mathbf{x}_{k-1}^i)w_{k-1}^i. \tag{12}$$

Equation (12) is derived by using Bayes rule and the fact that $\mathbf{y}_k|\mathbf{x}_k$ is independent of $\mathbf{x}_{k-1}$ to write

$$p(\mathbf{x}_k|\mathbf{x}_{k-1},\mathbf{y}_k) = p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_{k-1})/p(\mathbf{y}_k|\mathbf{x}_{k-1}),$$

and substituting this result into (8). Unlike the standard proposal (9), this proposal depends on the new observations $\mathbf{y}_k$; generating samples from (11) is therefore more closely related to data assimilation than to ensemble forecasting, which (9) mimics. In addition, the weights at time t_k are independent of the sample $\{\mathbf{x}_k^i\}$ drawn from the proposal; instead, they depend on the particle $\mathbf{x}_{k-1}^i$ from the previous time.

In the particle-filtering literature, (11) has come to be known as the "optimal" proposal. This terminology can be confusing, as the optimality does *not* refer to the performance of the resulting particle filter. Instead, (11) is optimal in the sense that it achieves the minimum variance of w_k^i over different random draws of $\mathbf{x}_k^i$, since by (12) w_k^i is independent of $\mathbf{x}_k^i$ and so that variance is zero.

2.3 Sampling from the optimal proposal

In contrast to the standard proposal, sampling from $p(\mathbf{x}_k|\mathbf{x}_{k-1},\mathbf{y}_k)$ is nontrivial in general, because of the conditioning on $\mathbf{y}_k$. Morzfeld et al. (2011) present an approach that relies on having an efficient means

to find the mode of $p(\mathbf{x}_k|\mathbf{x}_{k-1},\mathbf{y}_k)$. For the purposes of this paper, it suffices to examine a reasonably general setting in which analytic expressions for $p(\mathbf{x}_k|\mathbf{x}_{k-1},\mathbf{y}_k)$ are available.

Analytic progress is possible when the system and observation noise are additive and Gaussian, and the observation operator is linear. Let the system be

$$\mathbf{x}_k = M(\mathbf{x}_{k-1}) + \eta_k, \quad \mathbf{y}_k = \mathbf{H}\mathbf{x}_k + \varepsilon_k, \tag{13}$$

with $\eta_k \sim N(0,\mathbf{Q})$ and $\varepsilon_k \sim N(0,\mathbf{R})$.

In that case, as discussed in Doucet et al. (2000),

$$\mathbf{x}_k|\mathbf{x}_{k-1},\mathbf{y}_k \sim N(\bar{\mathbf{x}}_k,\mathbf{P}), \tag{14}$$

where

$$\bar{\mathbf{x}}_k = (\mathbf{I}-\mathbf{K}\mathbf{H})\mathbf{x}_{k-1} + \mathbf{K}\mathbf{y}_k, \quad \mathbf{P} = (\mathbf{I}-\mathbf{K}\mathbf{H})\mathbf{Q} \tag{15}$$

with $\mathbf{K} = \mathbf{Q}\mathbf{H}^T(\mathbf{H}\mathbf{Q}\mathbf{H}^T+\mathbf{R})^{-1}$. To see this, note that the system dynamics imply that $\mathbf{x}_k|\mathbf{x}_{k-1} \sim N(M(\mathbf{x}_{k-1}),\mathbf{Q})$; conditioning on $\mathbf{y}_k$ can then be achieved by applying Bayes rule using the standard Kalman-filter update for a prior with covariance $\mathbf{Q}$. Thus, the optimal proposal becomes a Gaussian with mean and covariance given by (15).

The weights also have an analytic expression, since (13) immediately implies that

$$\mathbf{y}_k|\mathbf{x}_{k-1} \sim N\left(\mathbf{H}M(\mathbf{x}_{k-1}),\mathbf{H}\mathbf{Q}\mathbf{H}^T+\mathbf{R}\right). \tag{16}$$

The updated weights w_k^i, given by (12), are thus obtained by evaluating at $\mathbf{y}_k$ the pdf for a Gaussian of mean $\mathbf{H}M(\mathbf{x}_{k-1}^i)$ and covariance $\mathbf{H}\mathbf{Q}\mathbf{H}^T+\mathbf{R}$.

3 Behavior of the weights in a simple example

A common issue in particle filtering is the tendency for one or a few of the weights to be much larger than the rest. This phenomenon is known as *degeneracy*. It can occur spuriously, owing to sampling variability in the algorithm, and in that case the weighted sample will be a poor approximation to the posterior pdf. Much of the subtlety of particle-filter algorithms, including the choice of proposal, centers on avoiding this degeneracy. This section illustrates the problem for both the standard proposal and the optimal proposal.

Consider the simple system in which

$$\mathbf{x}_k = a\mathbf{x}_{k-1} + \eta_{k-1}, \quad \mathbf{y}_k = \mathbf{x}_k + \varepsilon_k, \tag{17}$$

with $a > 0$ a scalar, $\eta_{k-1} \sim N(0,q^2\mathbf{I})$ and $\varepsilon_k \sim N(0,\mathbf{I})$. Each element of the state vector evolves and is observed independently and both the system dynamics and the observation equation are linear with additive Gaussian noise. Taking the observation-error variance to be unity, there are two parameters: a, which sets the change of variance of each element of $\mathbf{x}$ under the deterministic dynamics, and q, the standard deviation of the system noise.

The results of section 2c can now be applied. If we make the further assumption that $\mathbf{x}_{k-1} \sim N(0,\mathbf{I})$, then for the standard proposal the distributions needed for sampling $\mathbf{x}_k$ and for updating the weights are, respectively,

$$\mathbf{x}_k|\mathbf{x}_{k-1} \sim N(a\mathbf{x}_{k-1},q^2\mathbf{I}), \quad \mathbf{y}_k|\mathbf{x}_k \sim N(\mathbf{x}_k,\mathbf{I}), \tag{18}$$

while those needed for the optimal proposal follow (14) and (16) and are given by

$$\mathbf{x}_k|\mathbf{x}_{k-1},\mathbf{y}_k \sim N\left(\frac{a\mathbf{x}_{k-1}+q^2\mathbf{y}_k}{1+q^2},\frac{a^2+q^2}{1+a^2+q^2}\mathbf{I}\right), \quad \mathbf{y}_k|\mathbf{x}_{k-1} \sim N\left(a\mathbf{x}_{k-1},\left(1+q^2\right)\mathbf{I}\right). \tag{19}$$

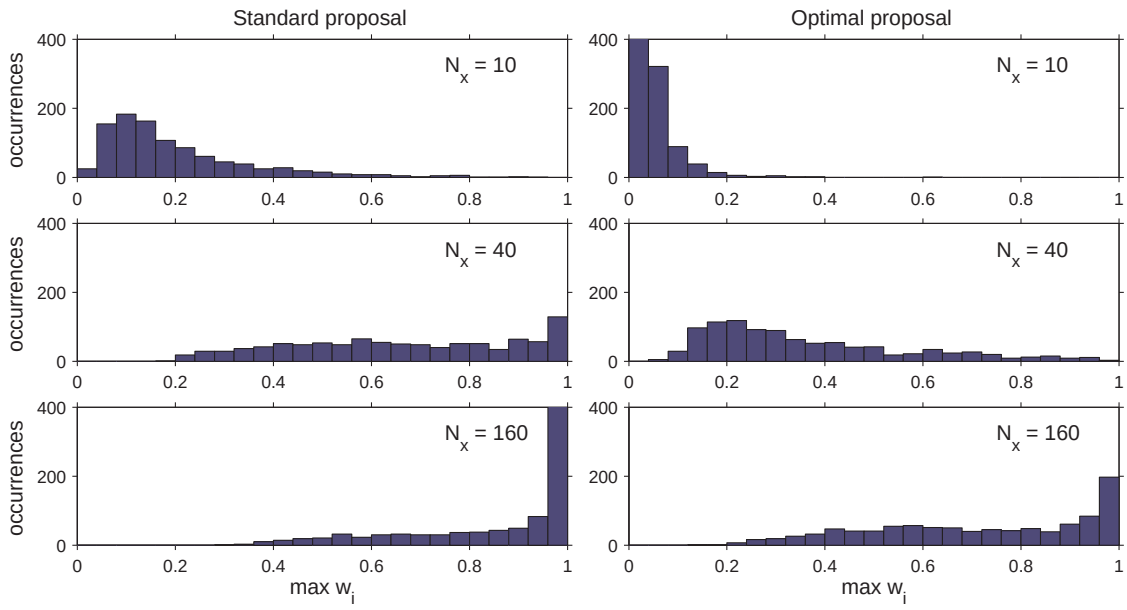

Figure 2: Histograms of the maximum weight from 10^3 simulations of the system (17), using either the standard proposals (left column) or the optimal proposal (right column). The weights come from a single update step for observations $\mathbf{y}_k$, with $\mathbf{x}^i_{k-1}$ drawn from $N(0,\mathbf{I})$. In the case of the optimal proposal with $N_x = 10$ and the standard proposal witht $N_x = 160$, the first and last bin, respectively, have greater than 400 occurrences.

The optimal proposal distribution involves a Kalman-filter update of the forecast from $\mathbf{x}_{k-1}$ given observations $\mathbf{y}_k$.

The new weights satisfy

$$w^i_k \propto \exp\left(-\frac{1}{2}|\mathbf{y}_k - \mathbf{x}^i_k|^2\right), \tag{20}$$

for the standard proposal and

$$w^i_k \propto \exp\left(-\frac{|\mathbf{y}_k - a\mathbf{x}^i_{k-1}|^2}{2\,(1+q^2)}\right). \tag{21}$$

The arguments of the exponentials in (20) and (21) have expected values that grow linearly with N_y. Thus, when N_y is large, a unit change in the arguments will produce increasingly dramatic changes in w^i_k, leading to a situation in which one or a few realizations of $\mathbf{x}^i_k$ (or $\mathbf{x}^i_{k-1}$ in the case of the optimal proposal) produce weights that are much larger than all others. For specified N_y, however, the argument of the exponential in (21) will be less than that in (20) with high probability: the denominator is always larger and $a\mathbf{x}^i_{k-1}$ will usually be closer to $\mathbf{y}_k$ than $\mathbf{x}^i_k$ is, since $\mathbf{x}^i_k$ is affected by the system noise. This suggests, correctly, that the optimal proposal will reduce the problem of degeneracy.

These points are illustrated in Fig. 2, which shows histograms of the maximum weight from simulations of a single update step with each of the proposals. The ensemble size is fixed, $N_e = 10^3$, and the dimension of the state and observation vectors varies, $N_x = 10$, 40, 160. As N_x increases, maximum weights close to unity become more frequent with either proposal – this is the degeneracy problem or, in the terminology of Snyder et al. (2008), the collapse of the weights. At each N_x, however, the degeneracy is less pronounced when using the optimal proposal.

4 Behavior of the weights

As shown by the preceding example, a key question for particle filters is how the required ensemble size increases as the dimension of the state increases. Bengtsson et al. (2008), Bickel et al. (2008) and Snyder et al. (2008) analyze the collapse of the weights in detail for the standard proposal. This section reviews their asymptotic arguments and results, and outlines how those results extend to the optimal proposal (11) and weights updated by (12).

Consider the update (8) for the weights at t_k and suppose that the weights at t_{k-1} are uniform. The latter condition means that we examine only the degeneracy that can occur over a single update. Let

$$V(\mathbf{x}_k, \mathbf{x}_{k-1}, \mathbf{y}_k) \equiv -\log(w_k/w_{k-1}) = \log\left(p(\mathbf{y}_k|\mathbf{x}_k)p(\mathbf{x}_k|\mathbf{x}_{k-1})/\pi(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{y}_k)\right). \tag{22}$$

The negative logarithm is included in the definition of V for convenience in the manipulations that follow. For the standard and optimal proposals, V is given by

$$V(\mathbf{x}_k, \mathbf{x}_{k-1}, \mathbf{y}_k) = \begin{cases} -\log p(\mathbf{y}_k|\mathbf{x}_k) & \text{for standard proposal distribution} \\ -\log p(\mathbf{y}_k|\mathbf{x}_{k-1}) & \text{for optimal proposal distribution} \end{cases}$$

We are interested in V as a random variable with $\mathbf{y}_k$ given, $\mathbf{x}_{k-1}$ distributed according to $p(\mathbf{x}_{k-1})$ and $\mathbf{x}_k$ distributed as $\pi(\mathbf{x}_k|\mathbf{x}_{k-1}, \mathbf{y}_k)$. Using an expectation over that distribution, we also define

$$\tau^2 = \mathrm{var}(V).$$

Since the maximum weight corresponds to the minimum V in a given sample, the left-hand tail of the distribution of V in particular governs the degeneracy of the weights.

The example of section 3 provides a particularly tractable situation. If $y_{k,j}$ and $x_{k,j}$ are the jth components of $\mathbf{y}_k$ and $\mathbf{x}_k$, respectively, then (18) and (19) imply

$$2V(\mathbf{x}_k, \mathbf{x}_{k-1}, \mathbf{y}_k) = \begin{cases} \sum_{j=1}^{N_y}(y_{k,j} - x_{k,j})^2 & \text{for standard proposal distribution} \\ \left(1+q^2\right)^{-1}\sum_{j=1}^{N_y}(y_{k,j} - a x_{k-1,j})^2 & \text{for optimal proposal distribution} \end{cases} \tag{24}$$

Under the assumptions of section 3, each term in the sums in (23) is independent and identically distributed (iid). The distribution of V therefore approaches a Gaussian when N_y, the number of observations and the number of terms in the sum, is large.

If V has an approximately Gaussian distribution when N_y is large, asymptotic results for the sample minimum of a Gaussian and for the tails of the Gaussian density and cumulative distribution function can be brought to bear (Bengtsson et al. 2008; Snyder et al. (2008), section 4b). This yields the relation

$$E(1/w^{(N_e)}) \sim 1 + \frac{\sqrt{2\log N_e}}{\tau}, \tag{25}$$

which is valid when $\sqrt{\log N_e}/\tau \ll 1$ and where the superscript (N_e) indicates the maximum of a sample of size N_e. Thus, the maximum weight $w^{(N_e)}$ approaches 1 as $\sqrt{2\log N_e}/\tau \to 0$ and, if one considers a system with larger τ, N_e must increase as $\exp\left(\tau^2/2\right)$ in order to keep $E(1/w^{(N_e)})$ constant. The exponential dependence of N_e on τ occurs for either proposal distribution, though for a given system and observational network, τ will differ between the two proposals.

The asymptotic approach of V to a Gaussian as $N_y \to \infty$ can be shown under more general conditions than requiring each degree of freedom to be iid and independently observed (Bengtsson et al. 2008, Bickel et al. 2008). For the standard proposal, Bengtsson et al. (2008) show that V is asymptotically Gaussian when $y_{k,j}|\mathbf{x}_k$ and $y_{k,l}|\mathbf{x}_k$ are independent for $j \neq l$ and the likelihoods $p(y_{k,j}|\mathbf{x}_k)$, considered as functions of $\mathbf{x}_k$ with $\mathbf{y}_k$ fixed, have sufficiently similar distributions and are only weakly dependent

as j varies. Those arguments do not extend directly to optimal proposal, as $p(\mathbf{y}_k|\mathbf{x}_{k-1})$ need not factor into a product over likelihoods for individual components $y_{k,j}$ even when the observations errors are independent. Stronger results hold for linear, Gaussian systems (i.e. those of the form (13) but with $M(\mathbf{x}_k) = \mathbf{M}\mathbf{x}_k$ linear; Bengtsson et al. 2008, Bickel et al. 2008, and section 5 of Snyder et al. 2008) and for such systems the results can be extended to the optimal proposal. The derivation begins by applying a linear transformation to the observation variables so that, in terms of the transformed variables, the observation errors have identity covariance and V can be written as a sum over independent terms,

$$V = \sum_{j=1}^{N_y} V_j = \frac{1}{2}\sum_{j=1}^{N_y} (y'_{k,j} - \mathbf{G}\mathbf{x}_{k-1})^2,$$

where the prime denotes a transformed observation variable. If $\mathrm{cov}\left((\mathbf{HQH}^T + \mathbf{R})^{-1/2}\mathbf{HM}\mathbf{x}_{k-1}\right)$ has the eigenvalue-eigenvector representation $\mathbf{E}\Lambda\mathbf{E}^T$, with $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{N_y})$, the required transformation is

$$\mathbf{y}' = \mathbf{E}^T(\mathbf{HQH}^T + \mathbf{R})^{-1/2}\mathbf{y},$$

which implies $\mathbf{G} = \mathbf{E}^T(\mathbf{HQH}^T + \mathbf{R})^{-1/2}\mathbf{HM}$. Following Bickel et al. (2008), the distribution of V approaches a Gaussian as $N_y \to \infty$ if and only if

$$\sum_{j=1}^{N_y} \lambda_j^2 \to \infty. \tag{27}$$

Following this somewhat technical exposition, it is worth recalling the main question, which is how the ensemble size required by the particle filter increases as the dimension of the state increases. Unfortunately, (25) relates (an expectation of) the maximum weight to N_e and τ^2, but not to the state dimension N_x. Returning again to the simple example, where each degree of freedom is iid and independently observed, $\tau^2 \propto N_y = N_x$. In general, the relation of τ^2 and N_x is less straightforward, because components of $\mathbf{x}_k$ may be dependent. When components are more dependent, the tendency for collapse of the weights is reduced, both because the most likely value of V is smaller and, more important, the left-hand tail of the distribution of V is compressed; in effect, the distribution of V behaves as though N_y were smaller.

Finally, we consider the quantitative dependence of τ, and thus the degeneracy of the weights, on the choice of proposal, again using the system of section 3 as an example. Using (24) and the relation between variance and kurtosis for a Gaussian distribution,

$$\tau^2 = \begin{cases} N_y(a^2+q^2)\left(\frac{3}{2}a^2 + \frac{3}{2}q^2 + 1\right) & \text{for standard proposal distribution} \\ N_y(q^2+1)^{-2}a^2\left(\frac{3}{2}a^2 + q^2 + 1\right) & \text{for optimal proposal distribution} \end{cases} \tag{28}$$

Consistent with the qualitative argument in section 3, τ^2 is always greater for the standard proposal than for the optimal proposal. The two proposals give the same τ^2 in the limit that system dynamics has no noise, $q^2 = 0$. As q increases (or a decreases, also increasing the relative importance of the system noise), the differences in τ^2 between the two proposals increases. For $a = q = 1/2$, which makes the prior variance of $\mathbf{x}_k$ equal to the observation-error variance, τ^2 from the standard proposal is 5 times that from the optimal proposal.

Since the ensemble size necessary to avoid degeneracy grows exponentially with τ^2, the optimal proposal can be effective with dramatically smaller ensembles in any given problem. Figure 3 compares results from the two proposals and shows the minimum N_e for which $E(1/w^{(N_e)}) < 1/0.9$ as a function of N_x, using $a = q = 1/2$. For both proposals, the necessary N_e increases exponentially with τ^2, and thus the state dimension N_x, as predicted by (25). At a given N_x, however, the optimal proposal needs orders of magnitude fewer ensemble members. The ratio of the slopes of the best-fit lines for $\log(N_e)$ versus N_x is 4.6, in reasonable agreement with the ratio of 5 predicted by (28).

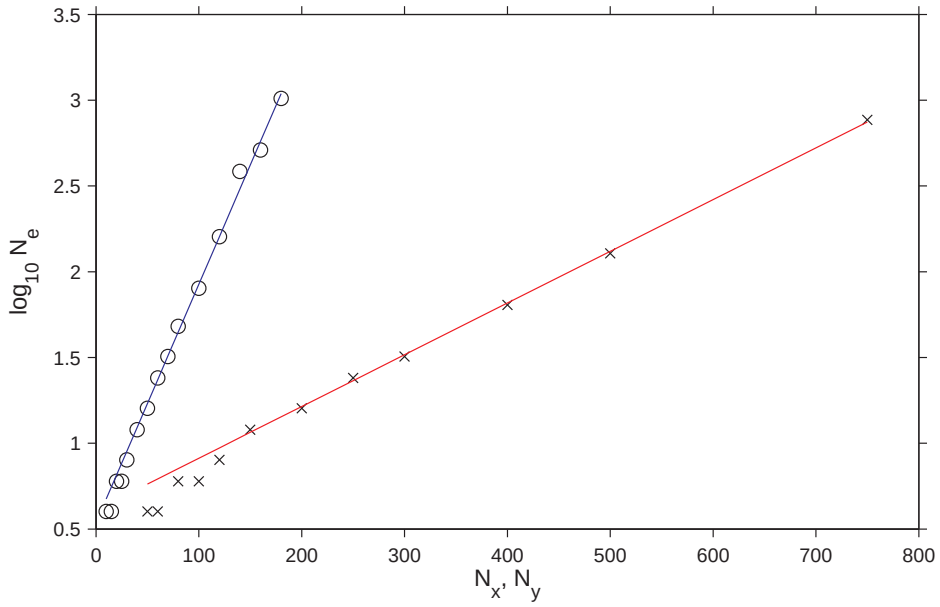

Figure 3: The minimum N_e such that $E(1/w^{(N_e)}) < 1/0.9$ for various values of N_x in the system (17). Results are shown for the standard proposal distribution(circles) and the optimal proposal (crosses), together with best-fit lines for each proposal that omit the data for the four smallest values of N_x. The expectation of $1/w^{(N_e)}$ is computed over 10^3 realizations.

5 Summary and conclusions

While particle filters have been successful on low-dimensional systems in a variety of fields, there is little if any experience with the very high-dimensional systems that are central to most geophysical applications. Even in low-dimensional systems, particle filters have a tendency for the weights assigned to each member to become extremely uneven, so that the ensemble becomes degenerate in the sense that one or a few members receive almost all the weight. Snyder et al. (2008; also Bengtsson et al. 2008, Bickel et al. 2008) demonstrate, under reasonably general assumptions, that avoiding such degeneracy requires the ensemble size to grow exponentially with the variance of the total log likelihood of the observations given the state.

Sequential importance sampling underlies the particle filter but was not considered in Snyder et al. (2008). In SIS, one chooses a distribution, called the proposal distribution, from which the particles are drawn and then computes appropriate weights for each particles so that the weighted sample approximates the correct posterior distribution. This paper has reviewed sequential importance sampling, with an emphasis on the so-called optimal proposal distribution, $p(\mathbf{x}_k|\mathbf{x}_{k-1},\mathbf{y}_k)$, which utilizes the latest observations when drawing the new ensemble members, and comparison against the standard proposal, which simply evolves the members to the next observation time using the system dynamics.

A simple example in which the necessary distributions may all be handled analytically shows that degeneracy, while still present for sufficiently high-dimensional systems, is reduced with the optimal proposal relative to the standard proposal. The previous asymptotic arguments can also be extended to the optimal proposal. They demonstrate that it does not avoid the exponential increase of the required ensemble size as the system dimension grows. They also provide a quantitative measure of how much the optimal proposal improves over the standard proposal. In essence, use of the optimal proposal reduces the factor in the exponent in the relation between the ensemble size and the state dimension and so can dramatically reduce the required ensemble size. It seems clear that the optimal proposal will facilitate the use of particle filters for systems of moderate dimension (a few tens or hundreds), even if it does not

immediately provide a path to a truly high-dimensional particle filter.

References

Arulampalam, M. S., S. Maskell, N. Gordon and T. Clapp, 2002: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Processing*, **50**, 174–188.

Bengtsson T., C. Snyder, and D. Nychka, 2003: Toward a nonlinear ensemble filter for high-dimensional systems. *J. Geophys. Res.*, **108(D24)**, 8775–8785.

Bengtsson, T., P. Bickel and B. Li, 2008: Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. *IMS Collections*, **2**, 316–334. doi: 10.1214/193940307000000518.

Bickel, P., B. Li and T. Bengtsson, 2007: Sharp failure rates for the bootstrap filter in high dimensions. *IMS Collections*, **3**, 318–329. doi: 10.1214/074921708000000228

Bocquet, M., C. A. Pires and L. Wu, 2010: Beyond Gaussian statistical modeling in geophysical data assimilation. *Mon. Wea. Rev.*, **138**, 2997-3023.

Doucet A. 1998. On sequential simulation-based methods for Bayesian ltering. Technical Report, University of Cambridge, Dept. of Engineering, CUED-F-ENG-TR310, 26 p.

Doucet, A., S. Godsill, and C. Andrieu, 2000: On sequential Monte Carlo sampling methods for Bayesian filtering. *Statist. Comput.*, **10**, 197-208.

Doucet A., N. de Freitas, N. Gordon, Eds., 2001: *Sequential Monte Carlo Methods in Practice.* Springer-Verlag.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith, 1993: Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc.*, **140**, 107-113.

van Leeuwen, P. J.. 2003: A variance-minimizing filter for large-scale applications. *Mon. Wea. Rev.*, **131**, 2071–2084.

van Leeuwen, P. J., 2009: Particle filtering in geophysical systems. *Mon. Wea. Rev.*, **137**, 40894114.

van Leeuwen, P.J., 2010. Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Quart. J. Roy. Meteor. Soc.*, **136**, 1991-1999.

Morzfeld, M., X. Tu, E. Atkins and A. J. Chorin, 2011: A random map implementation of implicit filters. *J. Comput. Phys.*, **231**, 2049-2066.

Papadakis, N., E. Mémin, A. Cuzol and N. Gengembre, 2010: Data assimilation with the weighted ensemble Kalman filter. *Tellus*, **62A**, 673-697.

Snyder, C., T. Bengtsson, P. Bickel and J. L. Anderson, 2008: Obstacles to high-dimensional particle filtering. *Mon. Wea. Rev.*, **136**, 4629–4640.