Curse-of-dimensionality revisited: Collapse of importance sampling in very large scale systems.

Bo Li, Thomas Bengtsson, and Peter Bickel.
Department of Statistics,
University of California-Berkeley.

Abstract

It has been widely realized that Monte Carlo methods (approximation via a sample ensemble) may fail in large scale systems. This work offers some theoretical insight into this phenomenon. In the context of a particle filter (as well as in general importance samplers), we demonstrate that the maximum of the weights associated with the sample ensemble members converges to one as both sample size and system dimension tends to infinity. Under fairly weak assumptions, this convergence is shown to hold for both a Gaussian case and for a more general case with *iid* kernels. Similar singularity behavior is also shown to hold for non-Gaussian, spherically symmetric kernels (*e.g.* multivariate Cauchy distribution). In addition, in certain large scale settings, we show that the estimator of an expectation based on importance sampling converges weakly to a law, rather than the target constant. Our work is presented and discussed in the context of atmospheric data assimilation for numerical weather prediction.

1 Introduction

With ever increasing computing power and data storage capabilities very large scale scientific analyzes are feasible and necessary (*e.g.* Donoho, 2000). One important application area of high-dimensional data analysis is the atmospheric sciences, where solutions to the general (inverse) problem of combining data and model quantities are commonly required. For instance, to produce real-time weather forecasts (including hurricane and severe weather warnings), satellite radiance observations of humidity and radar backscatter of sea surface winds must be combined with time-integrated solutions of atmospheric and oceanic models. To such ends, the model-forecast/data-update cycle in numerical weather prediction has recently been formulated in a probabilistic framework (Evensen, 1994; Molteni et al., 1996; Toth and Kalnay, 1997), and much effort has been aimed at describing geophysical states through the sampling of high-dimensional probability density functions (Houtekamer and Mitchell, 1998; Houtekamer and Mitchell, 2001; Anderson, 2001; van Leeuwen, 2003). Motivated by this work, we investigate the dangers of naively using Monte Carlo approximations to estimate posterior distributions of high-dimensional systems. In particular, we show that accurate estimation of (truly) high-dimensional, non-Gaussian pdfs require sample sizes that grow exponentially with system dimension.

Much recent focus in the literature on data assimilating for numerical weather prediction has been on extending Kalman filter solutions to work efficiently in real-time in systems with degrees of freedom exceeding 10^6. One popular extension is given by the ensemble Kalman filter, a Monte-Carlo based filter version which draws samples from the posterior distribution for the state given the data and the model (Evensen and van Leeuwen, 1996; Burgers et al., 1998). However, even in the Gaussian case, the task of real-time sampling form high-dimensional distributions is conceptually non-trivial: computational resources limit sample sizes to several orders of magnitude smaller than system dimension, and the obtained samples span only a subspace of the entire system. To address the resulting problems associated with matrix rank-deficiencies and errors due to sampling variability, various filter approaches leverage sparsity constraints and localize spatial information to attenuate spurious correlations (Houtekamer and Mitchell, 2001; Hamill and Snyder, 2001; Tippett et al., 2003). Moreover, for systems with a finite number of dominant modes, moderate sample sizes are sufficient to accurately estimate posterior means and covariances (Furrer and Bengtsson, 2005).

For longer forecast lead times, the involved dynamical models may exhibit strongly nonlinear behavior and produce non-Gaussian error distributions. In these situations, optimal filtering requires the use of more fully Bayesian filtering methods to combine data and models. In the context of oceanographic data assimilation, one such approach is considered by Van Leeuwen (2003), who proposes an importance re-sampling algorithm to obtain posterior estimates of oceanic flow structures. This method falls within the set of sequential importance sampling procedures (Rubin, 1988), commonly referred to as particle filters (*e.g.* Doucet et al., 2001). Using a finite set of sample points with associated sample-weights, the particle filter seeks to propagate the probability distribution of the unknown state forward in time. Once new data is available, Bayes theorem is used to re-normalize the weights based on how "close" the associated sample points are to the data.

Although successfully applied to a variety of settings, particle filters often yield highly varying importance weights and are known to be unstable even in low-order models. Consequently, much effort has been devoted to stabilizing the filter. Remedies include re-sampling (re-normalizing) the involved empirical measure at regular time intervals (Gordon et al., 1993; Liu, 2001), marginalizing or restricting the sample space (Liu and Chen, 1998; Pitt and Shepard, 1999), and diversifying the sample (*e.g.* Gilks and Berzuini, 2001). However, these approaches do not fundamentally address slow convergence rates when Monte Carlo is applied to truly large-scale systems, but rather serve to improve filter performance in low-dimensional systems. In particular, as noted by van Leeuwen (2003), when applied to geophysical models of high-dimension, sequential importance sampling collapses after a few (or even one) observation cycles. To shed light on the effects of dimensionality on filter stability, this work describes the relationship between system dimension and required sample size. Specifically, we provide necessary sample size requirements to avoid serious filter inefficiencies encountered in truly high-dimensional problems.

This work is outlined as follows. The next section formulates the problem of using ensemble methods for approximation purposes in large scale systems, and provide motivating examples illustrating the potential difficulties of high-dimensional estimation. Our main result is presented in Section 3, where the maximum of the weights associated with the sample ensemble members is shown to converge to one for both Gaussian and general *iid* kernels. We also discuss similar behavior exhibited by non-Gaussian spherically symmetric kernels, *e.g.* multivariate Cauchy distribution. Additionally, in certain circumstances, an unexpected consequence of the singularity is illuminated; namely, that the estimator of an expectation based on importance sampling converges weakly to some law, rather than the correct constant. Our work is concluded in Section 4 with a discussion of various approaches to dimensionality reduction.

2 Setting and Motivation

2.1 Setting

The statistical context in which we motivate our work is as follows. Consider a set of n sample points $\mathbf{X} = \{X_1, \ldots, X_n\}$, where $X_i \in \Re^d$ and both the sample size n and system dimension d are "large". We assume that the sample $\mathbf{X}$ is drawn randomly from the prior (or proposal) distribution $p(X)$. New data Y is related to the state X by the conditional density $p(Y|X)$. For concreteness, a functional relationship $Y = f(X) + \varepsilon$ is assumed and ε is taken to be independent of the state. The goal is to estimate posterior expectations using the importance ratio: *e.g.*, for some function $h(\cdot)$, we want to estimate

$$E(h(X)|Y) = \int h(X) \frac{p(Y|X)p(X)}{\int p(Y|X)p(X)\mathrm{d}X}\mathrm{d}X,$$

and use

$$\hat{E}(h(X)|Y) = \sum_{i=1}^{n} h(X_i) \frac{p(Y|X_i)}{\sum_{j=1}^{n} p(Y|X_j)}.$$

The above approximation can be interpreted as an importance sampling estimator of the expectation of $E(h(X)|Y)$, where the proposal distribution is $p(X)$ and the desired distribution is $p(X|Y)$. Based on this formulation, the importance ratio

$$w_i = \frac{p(Y|X_i)}{\sum_{j=1}^{n} p(Y|X_j)}$$

is the primary object of our study.

As discussed, the collapse of the weights to a point mass (with $max(w_i) \approx 1$) leads to disastrous behavior of the sampler. One intuition about such weight-collapses is well known, but here made precise in terms of d and n: Monte Carlo does not work if we wish to compute d-dimensional integrals with respect to product measures. For large d, the fundamental problem is that $\frac{1}{n}\sum_{i=1}^{n} p(Y|X_i)$ is a poor approximation to the constant $c(Y) = \int p(Y|X)p(X)dX$, where both $p(Y|X)$ and $p(X)$ are defined as product densities. In fact, even if the proportionality constant $c(Y)$ is known, $\frac{1}{n}\sum_{i=1}^{n} X_i p(Y|X_i)/c(Y)$ is still a poor estimate of $E(X|Y)$ unless n is grows exponentially in d. The reason for the weight collapse is that we are in a situation where the proposal distribution $p(X)$ and the desired sampling distribution are approximately mutually singular and (essentially) have disjoint support. As a consequence, the density of the desired distribution at all points of the proposed ensemble is small, but a vanishing fraction of density values predominate in relation to the others. This issue will be illuminated further at the end of Section 3.4.

Our precision of these intuitions is for situations where the proposal and desired distributions both have *iid* components, and the asymptotics apply to the case where both the ensemble size n and the dimension of the ensemble vectors d grows. Our main result, Proposition 3.1 (the Gaussian case) and Proposition 3.5 (General iid), show that if n grows only sub-exponentially in d (Gaussian case) or sub-exponentially in $d^{1/3}$ (general case), we have $max(w_i) \xrightarrow{p} 1$. These results are used in the next proposition, where we show that the usual estimate $\hat{E}$ of $E(h(X))$, where $h(\cdot)$ depends only on a small, fixed small number of components of X, is not stable. Proposition 3.6 thus establishes that $\hat{E}$ converges in law to the distribution of $h(X)$ under the proposal distribution.

2.2 Motivating examples

To illustrate the convergence of the maximum weight as n and d tend to infinity, the behavior of the importance sampler is simulated under a "null" scenario; i.e. the sampling is constructed to ensure that the importance weights are expected to be of equal magnitude. Further, to show the pervasiveness of the degeneracy for different distributions, the singularity is illustrated in both a Gaussian and a Cauchy setting. These densities are chosen to parallel the work of van Leeuwen (2003), who attempts to address filter collapse by modeling observation noise using the Cauchy distribution.

In our simulations, the observation Y is related to the state variable X through the model $Y = f(X)+\varepsilon$. We consider two error structures: first, where ε follows the multivariate normal distribution with mean 0 and covariance Σ; second, with $\varepsilon = (\varepsilon^{(1)}, \cdots, \varepsilon^{(d)})$, where each $\varepsilon^{(j)}$ is *iid* Cauchy. For the defined models, the weights can be re-expressed as

$$w_i = \frac{e^{-\varepsilon_i'\Sigma^{-1}\varepsilon_i/2}}{\sum_{\ell=1}^n e^{-\varepsilon_\ell'\Sigma^{-1}\varepsilon_\ell/2}}, \quad \text{and} \quad w_i' = \frac{\Pi_{j=1}^d 1/(1+\varepsilon_{ij}^2)}{\Sigma_{\ell=1}^n \Pi_{j=1}^d 1/(1+\varepsilon_{\ell j}^2)},$$

and these forms are used to simulate the sampling distribution of the importance weights.

Histograms of the maximum weight for the Gaussian and Cauchy simulations are displayed in the left and right panels of Figure 1. With the ensemble size fixed at $n = 1000$, the four histograms in each panel show the effects of increasing the dimension with $d = 10, 40, 100, 400$. As can be seen, for the Gaussian case, the maximum weight starts to dominate the sample for $d = 40$, while a more rapid degeneracy rate is exhibited for the Cauchy case. Each histogram is based on 400 simulation iterations.

We also examine the ensemble size needed to prevent the singularity of the maximum weight. Here, we choose $d = 100$ and a comparably large ensemble size, $n = 100000$. Results are given in Figure 2 for the Gaussian (left panel) and Cauchy cases (right). As indicated by the histogram for the Gaussian case, increasing the sample size by a factor 100 (as compared to Figure 1) ameliorates the dominating effect of the maximum effect; yet, $max(w_i)$ still

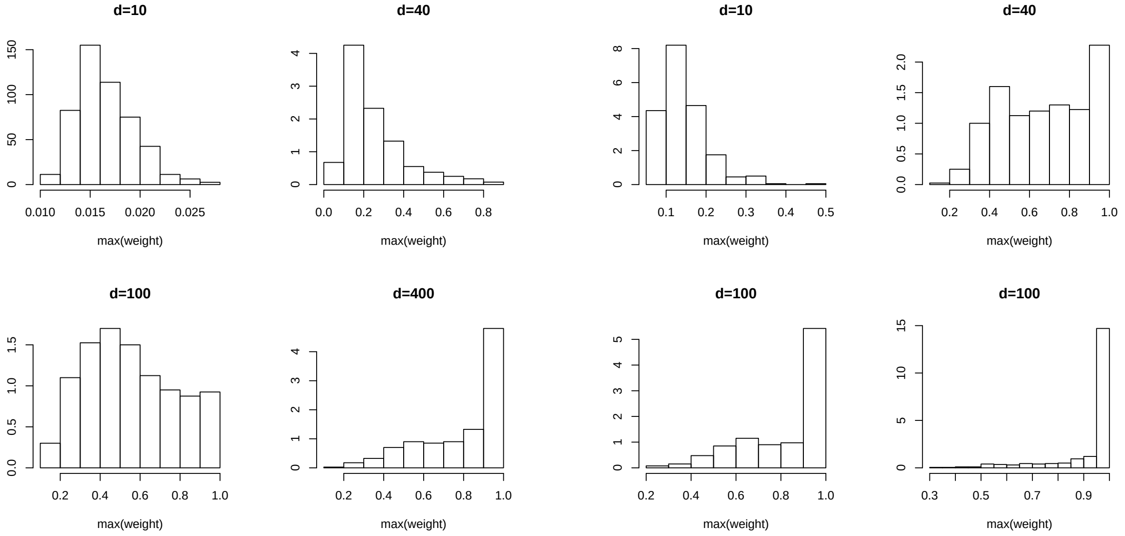

Figure 1: Histograms of $max(w_i)$ for the Gaussian case (left panel) and Cauchy case (right panel). Here, $n = 1000$, and $d = 10, 40, 100, 400$.

dominates the sample. Remarkably, for the Cauchy case, we see that increasing the sample size 100-fold barely impacts the distributions of $max(w_i)$.

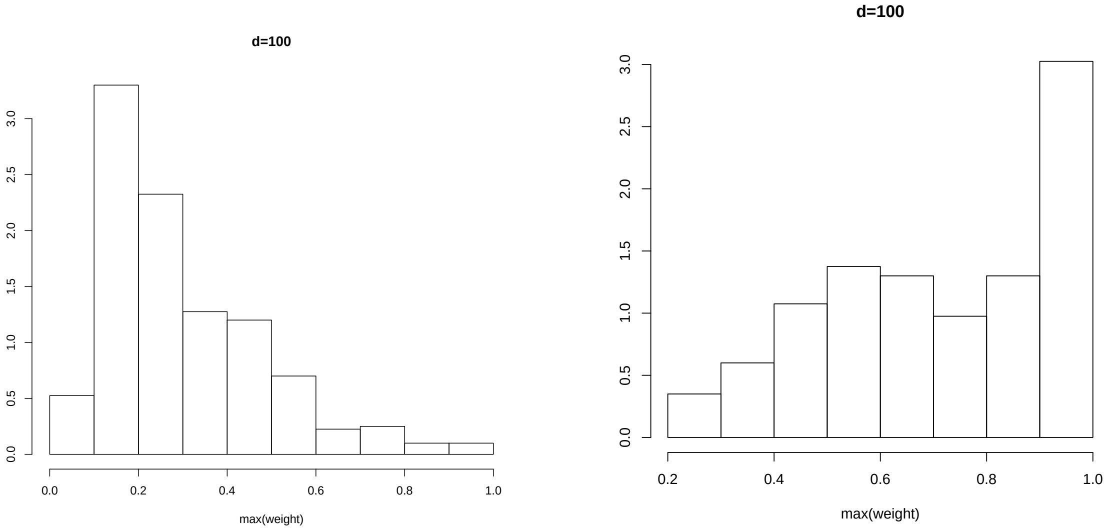

Figure 2: Histograms of $max(w_i)$ for the Gaussian case (left panel) and Cauchy case (right panel). Here, $n = 100000$, and $d = 100$.

As illustrated by the simulations, the convergence of the maximum importance weight holds in both the Gaussian and Cauchy setting. Further, very large sample sizes appear necessary to avoid the undesirable singularity feature. Next we present a formal study of the behavior of the maximum importance weight. Our precision about the convergence of the maximum weight is for situations where the proposal and desired distributions both have *iid* components, and the asymptotics apply to the case where both the ensemble size and the dimension of the ensemble vectors grows.

3 The singularity of the maximum weight in high-dimensional importance sampling.

We now develop the conditions under which the maximum weight $max(w_i)$ converges to unity. Specifically, under fairly weak assumptions, $max(w_i)$ is shown to approach unity as both system dimension d and sample size n tends to infinity. We treat the multivariate normal case is first, and then extend the results to a more general case. Our main results, given by Proposition 3.1 (the Gaussian case) and Proposition 3.5 (general *iid* case), show that if n grows only sub-exponentially in d (Gaussian case), or sub-exponentially in $d^{1/3}$ (general case), we have $max(x_i) \xrightarrow{p} 1$. Following these results, it is shown that the usual estimate $\hat{E}$ of $E(h(X))$, with $h(\cdot)$ depending only on a small, fixed small number of components of X, does not stabilize. Instead, Proposition 3.6 establishes that $\hat{E}$ converges in law to the distribution of $h(X)$ under the proposal distribution.

3.1 Multivariate Normal Case

As before, we assume $Y = f(X) + \varepsilon$ with $\varepsilon \sim N(0, \Sigma)$. Thus, the conditional distribution $p(Y|X)$ is multivariate normal with mean $f(X)$ and covariance Σ. Given the Gaussian assumption, the weights can be re-expressed (see Section 2.1) as

$$w_i = \frac{e^{-U_i/2}}{\sum_{\ell=1}^{n} e^{-U_\ell/2}},$$

where $U_1, \cdots, U_n$ are *iid* χ_d^2-distributed. With $U_{(1)} \leq \cdots \leq U_{(n)}$ the ordered U_i's, the maximum weight is again re-expressed,

$$w_{(n)} = \frac{1}{1 + \Sigma_{\ell=2}^{n} e^{-(U_{(\ell)} - U_{(1)})/2}}.$$

With some effort, we can establish

Proposition 3.1 *If n and d both tend to infinity, and $log(n)/d \to 0$, we have $w_{(n)} \xrightarrow{p} 1$.*

The proof of Proposition 3.1 is only sketched here and the technical details are given in the Appendix. Denote $T_{n,d} = \Sigma_{\ell=2}^{n} e^{-(U_{(\ell)} - U_{(1)})/2}$. In view of Markov's inequality, in order to show $w_{(n)} \xrightarrow{p} 1$, it suffices to show $E(T_{n,d}) \to 0$. With $F_d(z) = \left(2^{d/2}\Gamma(d/2)\right)^{-1} \int_0^z t^{d/2-1} e^{-t/2} \mathrm{d}t$ the cumulative distribution function (CDF) of the χ_d^2-distribution, and $\bar{F}_d = 1 - F_d$ the survival probability function, the expectation of $T_{n,d}$ is evaluated through the conditional expectation,

$$\begin{aligned}
E(T_{n,d}) &= E(E(V_{n,d}|U_{(1)})) \\
&= \int_0^{\infty} \frac{\left(2^{d/2}\Gamma(d/2)\right)^{-1}\int_x^{\infty}(n-1)e^{-(y-x)/2}\cdot y^{d/2-1}e^{-y/2}\mathrm{d}y}{\bar{F}_d(x)} n(\bar{F}_d(x))^{n-1}\mathrm{d}F_d(x) \\
&= 2^{-d/2}n(n-1)\int_0^{\infty} e^{x/2}\bar{F}_d(2x)(\bar{F}_d(x))^{n-2}\mathrm{d}F_d(x). \qquad (1)
\end{aligned}$$

To show $E(T_{n,d}) \to 0$ we proceed by breaking the integral (1) into several parts; see Appendix for details. The above result is sharp in the sense that if n increases sub-exponentially with respect to d, we have $w_{(n)} \xrightarrow{p} 1$.

3.2 Multivariate Cauchy Case

Again we assume the relationship $Y = f(X) + \varepsilon$, but now with ε following the multivariate Cauchy distribution. The density of a d-variate Cauchy distribution is given by $c(d)/(1+\|\varepsilon\|^2)^{\frac{1+d}{2}}$, where $\|\cdot\|$ denotes the Euclidean norm in $\Re^d$. Letting $R^2 = \|\varepsilon\|^2$, the weights can be reformulated as

$$w_i = \frac{1/(1+R_i^2)^{\frac{1+d}{2}}}{\Sigma_{\ell=1}^n 1/(1+R_\ell^2)^{\frac{1+d}{2}}}.$$

With this formulation, we will show $max(w_i) \xrightarrow{p} 1$.

Note that if $\{Z_1, \ldots, Z_d, Z_{d+1}\}$ are *iid* scalar random variables with standard normal distribution, the vector $[Z_2/Z_1, \cdots, Z_{d+1}/Z_1]'$ follows the d-variate Cauchy distribution. Now, for large d, the distribution of R^2/d may be approximated by the distribution of $1/Z_1^2$, and the density of R^2 is (approximately) given by $f_{R^2}(v) = (2\pi)^{-1/2}d^{1/2}v^{-3/2}e^{-d/2v}$. Moreover, from extreme value theory, with $R_{(1)} \leq \ldots \leq R_{(n)}$ the ordered norms associated with the sample, we have $R_{(1)}^2 \sim \frac{d}{2\text{log}n}$. As before, we express the maximum weight by

$$w_{(n)} = \frac{1}{1+T_{n,d}},$$

where $T_{n,d} = \Sigma_{\ell=2}^n\left[(1+R_{(1)}^2)/(1+R_{(\ell)}^2)\right]^{\frac{1+d}{2}}$. Then, with $(\log n)^2/d = O(1)$, we have $ET_{n,d} \to 0$, and can state the following result.

Proposition 3.2 *If n and d both tend to infinity, and $(\log n)^2/d = O(1)$, we have $w_{(n)} \xrightarrow{p} 1$.*

A proof of Proposition 3.2 is given in the Appendix.

Thus, we arrive at the same conclusion as for the multivariate Gaussian case, but under slightly different assumptions. Next we relax the distributional assumptions of $p(Y|X)$ to include more general product densities.

3.3 Generalization to *iid* Kernels

We now verify that the singular phenomenon of the maximum importance weight sustains in a more general distributional case. We first introduce a lemma that gives a normal approximation to the distribution of *iid* sums which appear in a reformulation of the importance weights used to establish the convergence. The lemma is a combination of Theorems 7.1.1 and 8.1.1 of Ibragimov-Linnik, 1971, and is valid for moderately large deviations.

Lemma 3.3 *Suppose $Y_1, \cdots, Y_d$ are iid satisfying the Cramer condition $E(e^{a|Y_1|}) < \infty$ for some positive a. Let μ and σ denote the mean and standard deviation of Y_1; g_d and G_d denote the density function and CDF of $S_p = (Y_1 + \cdots + Y_d - d\mu)/\sigma\sqrt{d}$; and ϕ and Φ denote the density function and CDF of the standard normal distribution. Then, as $d \to \infty$, for $s \geq 1$ and $s = o(d^{1/2})$, we have*

$$\begin{aligned} g_d(s) &= \phi(s) exp\big(\lambda(s/\sqrt{d})s^3/\sqrt{d}\big)(1 + O(s/\sqrt{d})), \\ g_d(-s) &= \phi(s) exp\big(-\lambda(-s/\sqrt{d})s^3/\sqrt{d}\big)(1 + O(s/\sqrt{d})), \\ \bar{G}_d(s) &= \bar{\Phi}(s) exp\big(\lambda(s/\sqrt{d})s^3/\sqrt{d}\big)(1 + O(s/\sqrt{d})), \\ G_d(-s) &= \Phi(-s) exp\big(-\lambda(-s/\sqrt{d})s^3/\sqrt{d}\big)(1 + O(s/\sqrt{d})). \end{aligned}$$

Here, $\lambda(s)$ is Cramer's power series, convergent for $|s| \leq \epsilon(a)$, with $\epsilon(a)$ depending only on a.

An immediate corollary is given below. It gives a sharp normal approximation for narrower deviations, which is needed in our developments.

Corollary 3.4 *For the setting described in Lemma 3.2, as $d \to \infty$, for $s \geq 1$ and $s = o(d^{1/6})$, we have*

$$\begin{aligned} g_d(s) = (1 + o(1))\phi(s),\ g_d(-s) = (1 + o(1))\phi(s), \\ \bar{G}_d(s) = (1 + o(1))\bar{\Phi}(s),\ G_d(-s) = (1 + o(1))\Phi(-s). \end{aligned}$$

Here, $o(1)$ is in the sense of uniformity.

We are now ready to treat the general *iid* case. With $X = (X^{(1)}, \cdots, X^{(d)}), Y = (Y^{(1)}, \cdots, Y^{(d)})$, we require the following decomposition of the observational density

$$p(Y|X) = \prod_{j=1}^{d} p_1(Y^{(j)}|X^{(j)}),$$

for some density function p_1. Thus, given the state X, the components of the observation vector are *iid*, and the j-th component of the observation depends only on the j-th state variable. For convenience, we set $X_{ij} = X_i^{(j)}, Y_{ij} = Y_i^{(j)}$, let $X_i = (X_{i1}, \cdots, X_{id}), Y_i = (Y_{i1}, \cdots, Y_{id})$, $1 \leq i \leq d$, and denote the *iid* state-observation pairs by $(X_1, Y_1), \cdots, (X_n, Y_n)$.

Now, with $U_i = -\Sigma_{j=1}^d \log(p(Y_{ij}|X_{ij}))$ and $U_i = d\mu + \sigma\sqrt{d}S_i$, using the observation density decomposition, we rewrite the weights as

$$w_i = \frac{e^{\sigma\sqrt{d}S_i}}{\Sigma_{l=1}^n e^{\sigma\sqrt{d}S_l}},$$

where $\mu = -E\log(p(Y^{(1)}|X^{(1)}))$ and $\sigma^2 = Var\big(\log(p(Y^{(1)}|X^{(1)}))\big)$. Hence, analogous to the normal case, the maximum weight can be expressed as

$$w_{(n)} = \frac{1}{1 + \Sigma_{l=2}^{n} e^{-\sigma\sqrt{d}(S_{(l)} - S_{(1)})}}.$$

With $T_{n,d} = \Sigma_{\ell=2}^{n} e^{-\sigma\sqrt{d}(S_{(\ell)} - S_{(1)})}$, we show in the Appendix that $E(T_{n,d}) \sim \frac{1}{\sigma\sqrt{d} - \sqrt{2\text{log}n}}$. We are now ready to state our next proposition.

Proposition 3.5 *Assume* $E\big[(p(Y^{(1)}|X^{(1)}))^a\big] < \infty$*, for some* $a > 0$*. As* n *and* d *both tend to infinity, and* $(\log(n))^3/d \to 0$*, we have*

$$w_{(n)} \sim \frac{1}{1 + \frac{1}{\sigma\sqrt{d} - \sqrt{2logn}}},$$

and, certainly, $w_{(n)} \xrightarrow{p} 1$.

The above result will be utilized next to demonstrate a strange limiting behavior of a importance sampling estimator of an expectation.

3.4 Failure of Importance Sampling with increasing dimension

So far our developments have been focused on likelihood-based update mechanisms in the particle filter context. As a natural next step, we now investigate the performance of general importance sampling procedures for large scale systems. Not surprisingly, even in a general context, the importance weights still behave singularly. Here, we give a concrete example of such degeneracy.

Let $\{X_1, \cdots, X_n\}$ be an *iid* sample with reference density $q(x)$. We again assume that each sample point is a d-dimensional vector, $X_i = (X_{i1}, \ldots, X_{id})$. Letting $E_p(\cdot)$ denote expectation with respect to the target density $p(X)$, we approximate $E_p(h(X))$ by $\Sigma_{i=1}^{n} w_i h(X_i)$, where

$$w_i = \frac{p(X_i)/q(X_i)}{\Sigma_{\ell=1}^{n} p(X_\ell)/q(X_\ell)}.$$

We shall assume $p(X_i) = \Pi_{j=1}^{d} p_1(X_{ij})$ and $q(X_i) = \Pi_{j=1}^{d} q_1(X_{ij})$, and further that $h(X)$ in fact depends on a small, fixed number of components of X.

Using Proposition 3.5, the proposed estimator $\Sigma_{i=1}^{n} w_i h(X_i)$ will be shown to converge in law to a random quantity; as opposed to our target quantity $E_p(h(X))$. For ease of notation, but without loss of generality, we assume $h(X) = h(X^{(1)})$, where $X^{(1)} = X_{11}$ is the first component of X. We have the following result.

Proposition 3.6 *Assume* h *is uniformly bounded over the sample space of* X_{11}*, and* $E_{q_1}\big[(p_1(X^{(1)})/q_1(X^{(1)}))^a\big] < \infty$*, for some* $a > 0$*. As* n *and* d *both tend to infinity, and* $(\log n)^3/d \to 0$*, we have*

$$\Sigma_{i=1}^{n} w_i h(X_i) \xrightarrow{d} h(X_{11}),$$

where X_{11} *obeys the law* $q_1(\cdot)$ *inherited from the sampling distribution* $q(\cdot)$*.*

To verify the result we let $V_{ij} = -\log\big(p_1(X_{ij})/q_1(X_{ij})\big)$, and rewrite the weights as $w_i = \frac{e^{-(V_{i1}+\cdots+V_{id})}}{\Sigma_{i=1}^n e^{-(V_{i1}+\cdots+V_{id})}}$. In light of Proposition 3.5, we see that $max(w_i) \xrightarrow{p} 1$. Now define the random index

$$I = I_{n,p} = \{k : 1 \le k \le n, V_{k1} + \cdots + V_{kd} = \min_{1\le i\le n}(V_{i1} + \cdots + V_{id})\}.$$

Thus, I denotes the index for the maximum weight.

We are going to show $\Sigma_{i=1}^n w_i h(X_{1i}) \xrightarrow{d} h(X^{(1)})$. Since $w_I \xrightarrow{p} 1$, which implies $\Sigma_{i\neq I} w_i \xrightarrow{p} 0$, assisted by the assumption that h is bounded, we have

$$|\Sigma_{i=1}^n w_i h(X_{1i}) - h(X_{1I})| = |(w_I - 1)h(X_{1I}) + \Sigma_{i\neq I} w_i h(X_{1i})| \xrightarrow{d} 0.$$

Thus, we need to show $X_{1I} \xrightarrow{d} X_{11}$. The remaining details are left to the Appendix.

The preceding proposition implies that the estimator of the expectation based on importance sampling may not be consistent in high dimensional circumstances. A more "traditional" prospective, with fixed dimensionality, provides some insight into this phenomena. A simple variance calculation, which neglects the denominator (or assumes the denominator is good estimator of 1), of $\hat{E}(h(X)) = \frac{\Sigma_{i=1}^n h(X_i)p(X_i)/q(X_i)}{\Sigma_{i=1}^n p(X_i)/q(X_i)}$, leads to the notion of Effective Sample Size (ESS) (*e.g.* see Liu, 2001). The ESS is defined as ESS $= \frac{n}{1+Var(p(X)/q(X))}$, and under our assumption, we find

$$1 + Var_q(p(X)/q(X)) = \left[E_{q_1}\big(p_1(X^{(1)})/q_1(X^{(1)})\big)^2\right]^d.$$

Now, by the Cauchy-Schwartz inequality, as long as p_1 and q_1 share support, we know that

$$E_{q_1}\big(p_1(X^{(1)})/q_1(X^{(1)})\big)^2 > 1.$$

Thus, for large d, ESS is small, and the conclusion is similar to ours. However, in our case, since under $(\log n)^3/d \to 0$, both the numerator and denominator in $\hat{E}h(X)$ explode, and the calculation that yields the variance approximation is not valid in our case.

4 Discussion

The developments in this paper demonstrate that brute-force only implementations of Monte Carlo methods to describe high-dimensional probability distributions will fail. Of course, this finding is not new - nor particularly profound; rather, our work makes explicit the rates at which sample sizes must grow (wrt system dimension) to avoid singularities and degeneracies. In particular, we give necessary bounds on n to avoid convergence to unity of the maximum importance weight, and, naturally, accurate estimation will require even larger sample sizes than those implied by our results. Not surprisingly, degeneracies have been observed in geophysical systems of moderate dimension (Anderson and Anderson, 1999; Bengtsson et al., 2003; also, C.Snyder/NCAR & T. Hamill/NOAA, personal communication, 2001). The usual manifestation of this degeneracy are Monte Carlo samples that are too "close" to the data, quickly producing singular probability measures, in particular as the filter is cycled over time.

The obvious remedy to this phenomenon is to achieve some form of dimensionality reduction, and the high-dimensional form in which the data are presented is typically open to such reduction with subsequent effective analysis. For instance, in the case of the ensemble Kalman filter, by imposing sparsity constraints through spatial localization (Houtekamer and Mitchell, 2001, Hamill Whitaker and Snyder, 2001; also, Furrer and Bengtsson, 2005). Be that is may, as shown in this work, for fully Bayesian analyzes of high-dimensional systems, reduction becomes absolutely essential lest spurious sample variability is to dominate the analyzes.

In the context of numerical weather prediction, one approach to dimension reduction may be to condition sample draws on a larger information set. One idea is given by Berliner (2001), who constructs proposal distributions by incorporating dynamic information in a low-order model. Other examples of geophysically constrained sampling schemes are given by Bayesian Hierarchical Models (*e.g.* Wikle et al., 2001; Hoar et al., 2003), but require computationally heavy, chain-based sampling and thus do not extend in any obvious manner to real-time applications. Another possibility is to break the system into lower-dimensional sets and sequentially perform the sampling (*e.g.* Bengtsson et al., 2003). Ideally, this approach involves identification of independent, low-dimensional manifolds.

Appendix

Proof of Proposition 3.1

From Lemma 3 and Lemma 4 of Bickel and Levina (2004), we have, for $z \in [d, (1 + (1 - \epsilon)/\epsilon^2)d]$,

$$\bar{F}_d(z) \leq exp\big(- \frac{\epsilon^2 (z-d)^2}{2d} \big).$$

Here, $\bar{F}_d(z)$ denotes the survival probability function (see Section 3.1). We set $\epsilon = 0.61$. With this choice of ϵ the above inequality holds for $z \in [d, 2.04d]$, and this bound is invoked several times in our proof. Another needed approximation is obtained by application of Sterling's formula:

$$\big(2^{d/2}\Gamma(d/2)\big)^{-1} \preceq C' e^{d/2} d^{-(d+1)/2}.$$

Now let $I(x) = 2^{-d/2} n(n-1) e^{x/2} \bar{F}_d(2x)(\bar{F}_d(x))^{n-2}$. Following the developments of Section 3.1, we need to show $\int_0^\infty I(x) \mathrm{d}F_d(x) \to 0$. We evaluate the integral over three regions: (i) $[0, 0.69d]$, (ii) $[0.69d, 1.005d]$, and (iii) $[1.005d, \infty]$:

(i)

$$\begin{aligned} &\int_0^{0.69d} I(x) \mathrm{d}F_d(x) \\ &< 2^{-d/2} n(n-1) \int_0^{0.69d} e^{0.69d/2} (\bar{F}_d(x))^{n-2} \mathrm{d}F_d(x) \\ &< e^{-(\log(2) - 0.69)d/2} n < n e^{-0.001d} \quad \to \quad 0 \end{aligned}$$

(ii)

$$\int_{\cup_{k=1}^8 J_k} I(x) \mathrm{d}F_d(x) < \sum_{k=1}^8 Cn(n-1)e^{-c_k d} \to 0$$

(iii)

$$\begin{aligned} \int_{1.005d}^\infty I(x) \mathrm{d}F_d(x) &< n(n-1) \int_{1.005d}^\infty (\bar{F}_d(x))^{n-2} \mathrm{d}F_d(x) \\ = \quad n\big(\bar{F}_d(1.005d)\big)^{n-1} &< n e^{-(0.005\epsilon)^2 d(n-1)/2} \to 0 \end{aligned}$$

To establish the result in (ii), we split the integral into 8 regions: i.e. $[0.69d, 1.005d] = \cup_{k=1}^8 J_k$, where $J_1 = [0.69d, 0.77d]$, $J_2 = [0.77d, 0.82d]$, $J_3 = [0.82d, 0.84d]$, $J_4 = [0.84d, 0.87d]$, $J_5 = [0.87d, 0.90d]$, $J_6 = [0.90d, 0.93d]$, $J_7 = [0.93d, 0.96d]$, and $J_8 = [0.96d, 1.005d]$. Now, for each J_k there exists a positive number c_k such that

$$\int_{J_k} I(x) \mathrm{d}F_d(x) < Cn(n-1)e^{-c_k d} \to 0.$$

For instance, for the interval J_1, we have

$$\begin{aligned}
\int_{0.69d}^{0.77d} I(x)\mathrm{d}F_d(x) &\leq 2^{-d/2}n(n-1)\int_{0.69d}^{0.77d} e^{-x/2}\bar{F}_d(2x)\mathrm{d}F_d(x) \\
< \;& 2^{-d/2}n(n-1)\int_{0.69d}^{0.78d} e^{-x/2}e^{-\frac{\epsilon^2(2x-d)^2}{2d}}\mathrm{d}F_d(x) \\
< \;& 2^{-d/2}n(n-1)e^{-\frac{\epsilon^2(2\times 0.69d-d)^2}{2d}}\int_{0.69d}^{0.77d} e^{-x/2}\mathrm{d}F_d(x) \\
= \;& 2^{-d/2}n(n-1)e^{-\frac{\epsilon^2(2\times 0.69d-d)^2}{2d}}\left(2^{d/2}\Gamma(d/2)\right)^{-1}(0.77d)^{d/2} \\
\preceq \;& C2^{-d/2}n(n-1)e^{-\frac{\epsilon^2(2\times 0.69d-d)^2}{2d}}\left(e^{d/2}d^{-d/2}(\pi d)^{-1/2}\right)(0.77d)^{d/2} \\
= \;& Cn(n-1)(\pi d)^{-1/2}e^{-(log(2)+(\epsilon(2*0.69-1))^2+log(1/0.77)-1)d/2} \\
< \;& Cn(n-1)e^{-0.008d} \to 0.
\end{aligned}$$

The remaining regions in (ii) are evaluated similarly.

Proof of Proposition 3.2

With the preparation given earlier, we can carry out the following approximate calculation.

$$\begin{aligned}
ET_{n,d} &\approx (2\pi)^{-1/2}(n-1)d^{1/2}\int_{\frac{d}{2\log n}}^{\infty}\left[(1+\frac{d}{2\log n})/(1+v)\right]^{\frac{1+d}{2}}v^{-3/2}e^{-d/2v}\mathrm{d}v \\
&\approx (2\pi)^{-1/2}(n-1)d^{1/2}(\frac{d}{2\log n})^{\frac{1+d}{2}}\int_{\frac{d}{2\log n}}^{\infty} v^{-(d+4)/2}e^{-d/2v}\mathrm{d}v \\
&\overset{v=d/(2u\log n)}{=} (2\pi)^{-1/2}(2\log n)^{1/2}(n-1)\int_0^1 u^{\frac{d}{2}}/n^u\mathrm{d}u \qquad (2)
\end{aligned}$$

Pick $\epsilon_n = \frac{\log\log n}{4\log n}$, we have

$$\begin{aligned}
(2\log n)^{1/2}(n-1)\int_0^{1-\epsilon_n} u^{\frac{d}{2}}/n^u\mathrm{d}u &< \frac{(2\log n)^{1/2}(n-1)(1-\epsilon_n)^{\frac{d}{2}+1}}{\frac{d}{2}+1} \\
&= \frac{(2\log n)^{1/2}e^{\log(n-1)-\epsilon_n(\frac{d}{2}+1)(1+o(1))}}{\frac{d}{2}+1} \\
&\to 0 \qquad (3)
\end{aligned}$$

and applying the inequality $1-(1-t)^n \leq nt$ $(0<t<1)$ yields

$$\begin{aligned}
(2\log n)^{1/2}(n-1)\int_{1-\epsilon_n}^{1} u^{\frac{d}{2}}/n^u\mathrm{d}u &< \frac{(2\log n)^{1/2}n^{\epsilon_n}(1-(1-\epsilon_n)^{\frac{d}{2}+1})}{\frac{d}{2}+1} \\
&< (2\log n)^{1/2}n^{\epsilon_n}\epsilon_n \\
&= \frac{1}{2(4\log n)^{1/4}} \to 0 \qquad (4)
\end{aligned}$$

(3),(4) and (5) together show $ET_{n,d} \to 0$.

Proof of Proposition 3.5

Let G_d denote the CDF of S_i. It follows from lemma 2.2 that $G_d(s)$ can be well approximated by the CDF of a standard normal distribution $\Phi(s)$ as $s = o(d^{1/6})$. In what follows, we shall invoke this approximation over the interval $[-(d\mathrm{log}n)^{1/8}, (d\mathrm{log}n)^{1/8}]$ (note $(d\mathrm{log}n)^{1/8} = o(d^{1/6})$ by $(\mathrm{log}n)^3/d \to 0$) and the inequality $1-(1-t)^n \le nt$ $(0<t<1)$ somewhere.

$$\begin{aligned}
& ET_{n,d} \\
= \;& (E(T_{n,d}1(S_{(1)} < -(d\mathrm{log}n)^{1/8})) \\
& +(n-1)E\Big[\frac{(\int_{(d\mathrm{log}n)^{1/8}}^{\infty} + \int_{S_{(1)}}^{(d\mathrm{log}n)^{1/8}})e^{-\sigma\sqrt{d}(z-S_{(1)})}g_d(z)dz}{\bar{G}_d(S_{(1)})}1(S_{(1)} > -(d\mathrm{log}n)^{1/8})\Big] \\
\le \;& (n-1)P(|S_{(1)}| > (d\mathrm{log}n)^{1/8})) + (n-1)(\bar{\Phi}((d\mathrm{log}n)^{1/8}) + \bar{G}_d((d\mathrm{log}n)^{1/8}))E\Big[\frac{1}{\bar{G}_d(S_{(1)})}\Big] \\
& +(n-1)e^{\sigma^2 d/2}(1+o(1))E\Big[\frac{e^{\sigma\sqrt{d}S_{(1)}}\bar{\Phi}(\sigma\sqrt{d}+S_{(1)})}{\bar{\Phi}_d(S_{(1)})}\Big] \\
\le \;& (n-1)\big[1-\big(1-G_d(-(d\mathrm{log}n)^{1/8})\big)^n\big] + n\big(\bar{\Phi}((d\mathrm{log}n)^{1/8}) + \bar{G}_d((d\mathrm{log}n)^{1/8})\big) \\
& +(n-1)e^{\sigma^2 d/2}(1+o(1))E\Big[\frac{e^{\sigma\sqrt{d}S_{(1)}}\bar{\Phi}(\sigma\sqrt{d}+S_{(1)})}{\bar{\Phi}(S_{(1)})}\Big] \\
\le \;& n(n-1)G_p(-(d\mathrm{log}n)^{1/8}) + n\big(\bar{\Phi}((d\mathrm{log}n)^{1/8}) + \bar{G}_d((p\mathrm{log}n)^{1/8})\big) \\
& +(n-1)e^{\sigma^2 d/2}(1+o(1))E\Big[\frac{e^{\sigma\sqrt{d}S_{(1)}}\bar{\Phi}(\sigma\sqrt{d}+S_{(1)})}{\bar{\Phi}(S_{(1)})}\Big]
\end{aligned}$$

Hence we derive that

$$ET_{n,p} = (n-1)e^{\sigma^2 d/2}(1+o(1))E\Big[\frac{e^{\sigma\sqrt{d}S_{(1)}}\bar{\Phi}(\sigma\sqrt{d}+S_{(1)})}{\bar{\Phi}(S_{(1)})}\Big] + o(\frac{1}{d}) \tag{5}$$

since

$$dn(n-1)G_d(-(d\mathrm{log}n)^{1/8}) = n(n-1)\Phi(-(d\mathrm{log}n)^{1/8}) \sim dn(n-1)\frac{e^{-(d\mathrm{log}n)^{1/4}/2}}{(d\mathrm{log})^{1/8}} \to 0$$

Next, applying lemma 2.2 and Mill's ratio $\bar{\Phi}(x) \sim \frac{\phi(x)}{x}$ (as $x \to 0$), we have

$$\begin{aligned}
P(S_{(1)} \le -\sqrt{2\mathrm{log}n} - \epsilon) &= 1 - (\bar{G}_d(-\sqrt{2\mathrm{log}n} - \epsilon))^n \\
&= 1 - (1-(1+o(1))\bar{\Phi}(\sqrt{2\mathrm{log}n} + \epsilon))^n \\
&\to 0
\end{aligned}$$

since $n\bar{\Phi}(\sqrt{2\mathrm{log}n} + \epsilon) \sim \frac{n\phi(\sqrt{2\mathrm{log}n}+\epsilon)}{\sqrt{2\mathrm{log}n}+\epsilon} \to 0$. Analogously,

$$\begin{aligned}
P(S_{(1)} \ge -\sqrt{2\mathrm{log}n} + \epsilon) &= (\bar{G}_d(-\sqrt{2\mathrm{log}n} + \epsilon))^n \\
&= (1-(1+o(1))\bar{\Phi}(\sqrt{2\mathrm{log}n} - \epsilon))^n \\
&\to 0
\end{aligned}$$

since $n\bar{\Phi}(\sqrt{2\text{log}n}-\epsilon) \sim \frac{n\phi(\sqrt{2\text{log}n}-\epsilon)}{\sqrt{2\text{log}n}-\epsilon} \to \infty$. Combining the last two facts yields $S_{(1)} + \sqrt{2\text{log}n} \xrightarrow{P} 0$. Then looking back to (5) and applying bounded convergence theorem gives $EY_{n,d} \sim \frac{1}{\sigma\sqrt{d}-\sqrt{2\text{log}n}}$.

Proof of Proposition 3.6

Following the previous discussion, it suffices to show, for every measurable set $\mathcal{A} \subset \Re$, $\big|P(X_{1I} \in \mathcal{A}) - P(X_{11} \in \mathcal{A})\big| \to 0$. Note $V_{ij} = -\text{log}\big(p_1(X_{ij})/q_1(X_{ij})\big)$, we simply need to show, for every measurable set $\mathcal{B} \subset \Re$, $\big|P(V_{1I} \in \mathcal{B}) - P(V_{11} \in \mathcal{B})\big| \to 0$. Notice

$$\begin{aligned} & \big|P(V_{1I} \in \mathcal{B}) - P(V_{11} \in \mathcal{B})\big| \\ = & \big|\Sigma_{k=1}^n P(V_{1I} \in \mathcal{B}, I = k) - P(V_{11} \in \mathcal{B})\big| \\ \leq & \big|\Sigma_{k=1}^n E\big((P(I = k|V_{1k}) - \frac{1}{n})1(V_{1k} \in \mathcal{B})\big)\big| \\ \leq & \Sigma_{k=1}^n E\big|P(I = k|V_{1k}) - \frac{1}{n}\big| \\ = & nE\big|(P(I = 1|V_{11}) - \frac{1}{n})\big| \end{aligned}$$

It suffices to show that

$$nE\big|P(I = 1|V_{11}) - \frac{1}{n}\big| \to 0 \tag{6}$$

Let $Z_{jk} = (V_{jk} - \mu)/\sqrt{2}\sigma$. Note that

$$\begin{aligned} P(I = 1|V_{11}) &= P(\frac{Z_{11} + \cdots + Z_{1d}}{\sqrt{d}} = \min_{1\leq i\leq n}\frac{Z_{i1} + \cdots + Z_{id}}{\sqrt{d}}|Z_{11}) \\ &= E(\bar{G}_d^{n-1}(\frac{Z_{11} + \cdots + Z_{1d}}{\sqrt{d}})|Z_{11}) \end{aligned}$$

and $E(\bar{G}_d^{n-1}((\frac{Z_{11}+\cdots+Z_{n1}}{\sqrt{d}}) = \frac{1}{n}$. Let $Z_1, \cdots, Z_n$ and Z_1' are *iid* copies of Z_{11} and denote $S_d = \frac{Z_1+Z_2+\cdots+Z_d}{\sqrt{d}}, S_d' = \frac{Z_1'+Z_2+\cdots+Z_d}{\sqrt{d}}$. It is equivalent to show

$$nE\big|E\big[\bar{G}_d^{n-1}(S_d) - \bar{G}_d^{n-1}(S_d')|Z_1\big]\big| \to 0 \tag{7}$$

Firstly, for each small $\epsilon > 0$, we can choose $C > 0$, such that $P(|Z_1 - Z_1'| > C) < \epsilon/2$. Then

$$\begin{aligned} & nE\big|E\big[\bar{G}_d^{n-1}(S_d) - \bar{G}_d^{n-1}(S_d')|Z_1\big]1(|Z_1 - Z_1'| > C)\big| \\ \leq & nE\big|E\big[\bar{G}_d^{n-1}(S_d) - \bar{G}_d^{n-1}(S_d')|Z_1\big]\big|P(|Z_1 - Z_1'| > C) \\ \leq & nE\big(E\big[\bar{G}_d^{n-1}(S_d) + \bar{G}_d^{n-1}(S_d')|Z_1\big]\big)P(|Z_1 - Z_1'| > C) \\ = & 2P(|(Z_1 - Z_1')| > C) \\ \leq & \epsilon \end{aligned}$$

Secondly, let $\tau_{n,d} = \sqrt{2\text{log}n - \frac{1}{4}\text{log}(d\text{log}n)}$. We proceed as

$$\begin{aligned}
& nE\big|E\big[\bar{G}_d^{n-1}(S_d) - \bar{G}_d^{n-1}(S'_d)|Z_1\big]1(|Z_1 - Z'_1| \le C, S'_d > -\tau_{n,d})\big| \\
\le \;& nE\big[\bar{G}_d^{n-1}(S_d)1(S_d > -\tau_{n,d} + \frac{C}{\sqrt{d}})\big] + nE\big[\bar{G}_d^{n-1}(S'_d)1(S'_d > -\tau_{n,d}\big] \\
= \;& \bar{G}_d^{n}(-\tau_{n,d} + \frac{C}{\sqrt{p}}) + \bar{G}_d^{n}(-\tau_{n,d}) \\
= \;& \big[1 - (1+o(1))\bar{\Phi}(\tau_{n,d} - \frac{C}{\sqrt{d}})\big]^n + \big[1 - (1+o(1))\bar{\Phi}(\tau_{n,d})\big]^n \\
\to \;& 0
\end{aligned}$$

since $n\bar{\Phi}(\tau_{n,d} - \frac{C}{\sqrt{d}}) \sim \frac{n\phi(\tau_{n,d} - \frac{C}{\sqrt{d}})}{\tau_{n,d} - \frac{C}{\sqrt{d}}} = \frac{(d\text{log}(n))^{1/8}}{\tau_{n,d} - \frac{C}{\sqrt{d}}} \to \infty$ and $n\bar{\Phi}(\tau_{n,d}) \to \infty$ similarly.

Thirdly, observe $g_d(z) = \phi(z)(1 + o(1))$ over the interval $[-(d\text{log}n)^{1/8} - \frac{C}{\sqrt{d}}, -\tau_{n,d} + \frac{C}{\sqrt{d}}]$ by Lemma 2.2 and the assumption $(\text{log}n)^3/d \to 0$.

$$\begin{aligned}
& nE\big|E\big[\bar{G}_d^{n-1}(S_d) - \bar{G}_d^{n-1}(S'_d)|Z_1\big]1(|Z_1 - Z'_1| \le C, -(d\text{log}n)^{1/8} \le S'_d \le -\tau_{n,d})\big| \\
\le \;& \frac{n(n-1)}{\sqrt{d}}E\Big|E\big[(Z_1 - Z'_1)g_d(\lambda S_d + (1-\lambda)S'_d)\bar{G}_d^{n-2}(\lambda S_d + (1-\lambda)S'_d)|Z_1\big] \\
& \times 1(|Z_1 - Z'_1| \le C, -(d\text{logn})^{1/8} - \frac{C}{\sqrt{d}} \le S_d \le -\tau_{n,d} + \frac{C}{\sqrt{d}}, -(d\text{log}n)^{1/8} \le S'_d \le -\tau_{n,d})\Big| \\
\le \;& \frac{Cn(n-1)}{\sqrt{p}}\phi(-\tau_{n,p} + \frac{C}{\sqrt{d}})E\Big(\bar{G}_d^{n-2}(S_d)1(-(d\text{log}n)^{1/8} - \frac{C}{\sqrt{d}} \le S_d \le -\tau_{n,d} + \frac{C}{\sqrt{d}}) \\
& + \bar{G}_d^{n-2}(S'_d)1(-(d\text{logn})^{1/8} \le S'_d \le -\tau_{n,d})\Big)(1 + o(1)) \\
= \;& \frac{Cn}{\sqrt{d}}\phi(-\tau_{n,d} + \frac{C}{\sqrt{d}})(1 + o(1))\big(2 - \bar{G}_d^{n-1}(-\tau_{n,d} + \frac{C}{\sqrt{d}}) - \bar{G}_d^{n-1}(-\tau_{n,d})\big) \\
\le \;& \frac{2Cn}{\sqrt{d}}\phi(-\tau_{n,d} + \frac{C}{\sqrt{d}})(1 + o(1)) \\
\to \;& 0
\end{aligned}$$

since $\frac{2Cn}{\sqrt{d}}\phi(-\tau_{n,d} + \frac{C}{\sqrt{d}}) \sim \frac{2C(2d\text{log}(n))^{1/4}(1+o(1))}{\sqrt{d}} \to 0$.

Fourthly, as in the second case, we have (We shall apply the basic inequality $1-(1-t)^n \le nt$ $(0 < t < 1)$ again.)

$$\begin{aligned}
& nE\big|E\big[\bar{G}_d^{n-1}(S_d) - \bar{G}_d^{n-1}(S_d')|Z_1\big]1(|Z_1 - Z_1'| \le C, S_d' < -(d\mathrm{log}n)^{1/8})\big| \\
\le\ & nE(\bar{G}_d^{n-1}(S_d)1(S_d < -(d\mathrm{log}n)^{1/8} + \frac{C}{\sqrt{d}})) + nE(\bar{G}_d^{n-1}(S_d')1(S_d' < -(d\mathrm{log}n)^{1/8}) \\
=\ & \big[1 - (1 - \bar{G}_d((d\mathrm{log}n)^{1/8} - \frac{C}{\sqrt{d}}))^n\big] + \big[1 - (1 - \bar{G}_d^n((d\mathrm{log}n)^{1/8})^n\big] \\
\le\ & n\bar{G}_d((d\mathrm{log}n)^{1/8} - \frac{C}{\sqrt{d}}) + n\bar{G}_d((d\mathrm{log}n)^{1/8}) \\
=\ & n\big[\bar{\Phi}((d\mathrm{log}n)^{1/8} - \frac{C}{\sqrt{d}}) + \bar{\Phi}((d\mathrm{log}n)^{1/8})\big](1 + o(1)) \\
\to\ & 0
\end{aligned}$$

since $n\bar{\Phi}((d\mathrm{log}n)^{1/8} - \frac{C}{\sqrt{d}}) \sim \frac{n\phi((d\mathrm{log}n)^{1/8} - \frac{C}{\sqrt{d}})}{(d\mathrm{log}n)^{1/8} - \frac{C}{\sqrt{d}}} \to 0$. Similarly for the second term. Hence (7) is proved via the four-step procedure. The claim in the proposition follows.

References

Anderson, J. (2001). An ensemble adjustment filter for data assimilation. *Monthly Weather Review 129*, 2884–2903.

Anderson, J. and S. Anderson (1999). A monte-carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review 127*, 2741–2758.

Bengtsson, T., C. Snyder, and D. Nychka (2003). Toward a nonlinear ensemble filter for high-dimensional systems. *Journal of Geophysical Research-Atmospheres 108*(D24), 8775.

Berliner, L. M. (2001). Monte Carlo based ensemble forecasting. *Stat. Comput. 11*(3), 269–275.

Bickel, P. J. and E. Levina (2004). Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli 10*(6), 989–1010.

Burgers, G., P. J., P. van Leeuwen, and G. Evensen (1998). Analysis scheme in the ensemble kalman filter. *Monthly Weather Review 126*, 1719–1724.

Donoho, D. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *Aide-Memoire of a Lecture at AMS conference on Math Challenges of 21st Centuary.*

Doucet, A., N. Freitas, and N. Gordon (Eds.) (2001). *Sequential Monte Carlo Methods in Practice.* Springer-Verlag.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research 99*(10), 143–162.

Evensen, G. and P. J. van Leeuwen (1996). Assimilation of geostat altimeter data for the Agulhas Current using the ensemble Kalman Filter. *Monthly Weather Review 124*, 85–96.

Furrer, R. and T. Bengtsson (2005). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis-Revised*.

Gilks, W. and C. Berzuini (2001). Following a moving target—Monte Carlo inference for dynamic Bayesian models. *J. R. Stat. Soc. Ser. B Stat. Methodol. 63*(1), 127–146.

Gordon, N., D. Salmon, and A. Smith (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F 140*(2), 107–113.

Hamill, T. M., J. S. W. and C. Snyder (2001). Distance-dependent filtering of background error covariance estimates in an ensemble kalman filter. *Monthly Weather Review 129*, 2776–2790.

Hoar, T., R. Milliff, D. Nychka, C. Wikle, and L. Berliner (2003). Winds from a Bayesian hierarchical model: Computation for atmosphere-ocean research. *Journal of Computational and Graphical Statistics 4*(781-807).

Houtekamer, P. and H. Mitchell (2001). A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review 129*, 123–137.

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

Liu, J. (2001). *Monte Carlo strategies in scientific computing.* Springer Series in Statistics. New York: Springer-Verlag.

Liu, J. and R. Chen (1998). Sequential Monte Carlo methods for dynamic systems. *J. Amer. Statist. Assoc. 93*(443), 1032–1044.

Molteni, F., R. Buizza, T. Palmer, and T. Petroliagis (1996). The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal fo the Royal Meteorological Society 122*, 73–119.

Pitt, M. and N. Shepard (1999). Filtering via simulation: Auxilliary particle filters. *Journal of American Statistical Association 94*(446), 590–599.

Rubin, D. (1988). *Bayesian Statistics 3*, Chapter Using the SIR algorithm to simulate posterior distributions, pp. 395–402. Oxford University Press.

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

Toth, Z. and E. Kalnay (1997). Ensemble forecasting at NCEP and the breeding method. *Monthly Weather Review 125*, 3297–3319.

van Leeuwen, P. (2003). A variance-minimizing filter for large scale applications. *Monthly Weather Reviews 131*, 2071–2084.

Wikle, C., R. Milliff, D. Nychka, and L. Berliner (2001). Spatiotemporal hierarchical bayesian modeling: Tropical ocean surface winds. *Journal of American Statistical Association 96*, 382–397.