Advances in Water Resources 86 (2015) 231–239

Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres

Iterative ensemble smoothers in the annealed importance sampling framework

Andreas S. Stordal [a,*], Ahmed H. Elsheikh [b]

[a] IRIS, P.O. Box 8046, Stavanger 4068, Norway
[b] Heriot-Watt University, Institute of Petroleum Engineering, Edinburgh EH14 4AS, United Kingdom

ARTICLE INFO

Article history:
Received 16 May 2015
Revised 25 September 2015
Accepted 30 September 2015
Available online 19 October 2015

Keywords:
Data assimilation
Iterative ensemble methods
Annealed importance sampling

ABSTRACT

Iterative ensemble techniques for solving inverse problems has recently gained a lot of interest in many geophysical communities. This popularity is attributed to the simplicity of implementation, general reliability and the ability to deal with the forward model as a black box without requiring the implementation of analytical gradients. Although several variants exist, we focus on the ensemble smoother with multiple data assimilation. This study highlights the similarity between the ensemble smoother and other existing techniques such as particle flow and annealed importance sampling. It is shown how a sequential Monte Carlo sampler can be used in combination with an annealing process to weight-correct the sampling procedure used in the ensemble smoother. Two different approximations in high dimensions, where the curse of dimensionality is unavoidable, are also presented. The methods proposed are compared with an MCMC run on a synthetic reservoir model.

1. Introduction

Inverse problems arise in many geosciences such as reservoir engineering, hydrology, atmospheric chemistry and oceanography. Common for all these areas is that the numerical model describing the dynamics of the physical models are defined on a very large scale. The order of unknown parameters and/or variables is usually at least $O(10^2)$ and could reach $O(10^{12})$. Even for the moderate case, the dimension of the parameter space and/or state space is too large for standard Bayesian inversion/filtering utilizing variants of MCMC [30] or sequential Monte Carlo methods [13]. A relatively new MCMC approach was proposed in [10]. The proposed method has the advantage over standard Metropolis Hastings implementation in the sense that only the likelihood and not the product of the prior and the likelihood has to be evaluated in the accept-reject step. This can be a huge advantage if the prior is only known implicitly or if measurements are of much lower dimension than the parameters/state vector. Although it was demonstrated that it converges faster than old algorithms, there is still a consensus in the geoscience community that MCMC methods are too time consuming for large scale applications [27] unless some proxy model or dimension reduction technique is used. Sequential Monte Carlo methods, also known as particle filter [13], suffer from the curse of dimensionality [2]. Even if a recent analysis shows that stability of a particle filter may be obtained in $O(Nd^3)$ operations [4], where N is the number of particles and d is the dimension of the state space, it is still far too computationally expensive for large scale models where d is at least $O(10^2)$ and each numerical function evaluation may take several minutes.

Our focus here is therefore on the more practical sequential Monte Carlo methods that are applied in many large scale real world problems. The most famous one being arguably the ensemble Kalman filter (EnKF) in different variants [17]. Although EnKF is easy to implement and do not suffer from curse of dimensionality in the same way as other Monte Carlo methods, there is always an asymptotic bias when applied to nonlinear problems [24]. This observation is of course also true for the smoother version of EnKF (EnKS, [34]) that assimilates all data at ones. Here we focus on parameter estimation, hence the smoother is equivalent to an off-line batch updating algorithm.

A new EnKS approach with multiple linear update steps has recently been suggested [3,9,15]. We show how this iterative ensemble smoother method, denoted ESMDA in the following, can be weight corrected in the nonlinear case. The idea is similar to Gaussian particle flow [5], where a Gaussian distribution evolves in pseudo time and takes into account partial information from the measurements sequentially. The mean and covariance of this Gaussian flow are obtained from a linearization of the initial state. The implementation of the ESMDA avoids linearization of the model in the nonlinear case which is desirable in large scale models where gradients are often

* Corresponding author. Tel.: +4755543849.
E-mail addresses: Andreas.S.Stordal@iris.no (A.S. Stordal), Ahmed.Elsheikh@pet.hw.ac.uk (A.H. Elsheikh).

http://dx.doi.org/10.1016/j.advwatres.2015.09.030
0309-1708/

hard to come by due to the frequent use of commercial software and/or lack of an adjoint code. The proposed methodology can be derived from annealed importance sampling [25]. Annealed Importance sampling (AIS) is an iterative importance sampling algorithm where the target distribution is tempered at each iteration. The method is a special case of the sequential Monte Carlo samplers defined by [14]. More details will be provided in Section 2.1. We discuss robustness of the proposed method and introduce a weight variance reduction technique [33] for implementation in large scale models. Finally, we extend this approach to Gaussian mixtures [8,20] in order to develop an iterative Gaussian mixture method along the lines of [31]. A Gaussian mixture smoother is a hybrid between importance sampling and the EnKS. It consists of a linear Kalman type update of the parameters and a weighting step where the weights are computed from an approximate likelihood function. The proposed methods are demonstrated on a canonical problem, simply to verify the theory presented, and on a synthetic 2-dimensional petroleum reservoir model. The paper is concluded with a summary section.

2. Bayesian inversion and annealed importance sampling

In the Bayesian framework we are concerned with posterior inference of a random variable $X \in \mathcal{R}^d$ given some prior information, usually defined via a prior probability density function, and some random measurements $Y \in \mathcal{R}^q$. X is typically either a time series $X = (X_1, \ldots, X_T)$ with prior information given by the joint probability density $p_X(x) = p_{X_1,\ldots,X_T}(x_1, \ldots, x_t) = p_{x_1}(x_1) \prod_{t=2}^{T} p_{X_t|X_{t-1}}(x_t|x_{t-1})$ or a vector of parameters with a prior probability density $p_X(x)$. In the rest of the manuscript we omit the subscript and simply denote $p(\cdot)$ to be the probability density of its argument as long as there is no room for confusion.

We assume that the measurements are linked with the variable X through a likelihood function $\ell(x) \overset{\text{def}}{=} p(y|x)$. The measurements are typically modeled as a (nonlinear) function of X with additive random noise. That is

$$Y = \mathcal{H}(X) + \epsilon, \tag{2.1}$$

where $\mathcal{H} : \mathcal{R}^d \to \mathcal{R}^q$ and the random variable ϵ is independent of X. Posterior inference (mean, mode, probability regions etc) can all be computed via the posterior probability density function

$$p(x|y) = p(x)\ell(x)C^{-1}, \tag{2.2}$$

where $C = \int \ell(x)p(x)\,dx$ is the normalizing constant. Unless stated otherwise, we assume in the following that ϵ in (2.1) is a zero mean Gaussian random variable with a covariance matrix $\mathbb{R}$. Note that Y is a vector where the different entries may consist of measurements collected at different time instances.

An annealing process [22] in the Bayesian framework typically consists of tempering either the likelihood function (or the posterior) by introducing a (discrete) pseudo time index $j = 0, \ldots, J$ and then take advantage of the fact that one may rewrite the likelihood as

$$\ell(x) = \prod_{j=0}^{J} \ell(x)^{\alpha_j}, \tag{2.3}$$

with the constraints that $\sum_{j=0}^{J} \alpha_j = 1$ and for all j, $0 \le \alpha_j \le 1$. The idea of annealing in Monte Carlo sampling is to slowly change the distribution of the samples from the prior to the posterior. These slow transitions are typically less variance prone than methods that targets the posterior directly from the prior. The name of course comes from its physical analog of heating material slowly in order to increase a material's ductility.

In the ESMDA the measurement error covariance matrix $\mathbb{R}$ is inflated to $\alpha_j^{-1}\mathbb{R}$ (note that each $\alpha_j^{-1} \ge 1$) at the iteration j with the same sum constraint on $\{\alpha_j\}_{j=0}^{J}$. In [15] linear algebra was used to show the equivalence with the standard EnKS for linear models (note that with our definition α_j corresponds to α_j^{-1} in [15]). Similarly, the update equations may be derived from (2.3). Hence the ESMDA can be formulated as an annealed importance sampling algorithm [25] in the Gauss-linear case. However, the method cannot be generalized to the nonlinear case in this framework without either modifying the ESMDA algorithm or modifying the annealed importance sampling algorithm [14]. Furthermore, the algorithm can be formulated as a population Monte Carlo [6] approach in the adaptive importance sampling framework (see e.g. [7,18,26]). However, for complex posterior distributions and high dimensional system, the annealing process is usually superior [14] and thus remains the focus here. The details of the extended algorithm are given in Section 2.1.

The annealing part of the annealed importance sampling algorithm enters through a sequence of probability densities, defined by a sequence of functions, $\{f_j\}_{j=1}^{J}$. Here we use the following definitions

$$f_j(x) = p(x)\ell(x)^{\beta_j}, \tag{2.4}$$

for a sequence $0 = \beta_0 < \beta_1 < \cdots < \beta_J = 1$, although other alternatives for f_j exist [25]. Note that $f_j(x)$ also depends on y through $\ell(x)^{\beta_j} = p(y|x)^{\beta_j}$. Since we are not guaranteed that f_j integrates to 1 for each j, a target density $p_j(x|y)$ can be defined by normalizing (2.4). That is, for each j the target density is defined by

$$p_j(x|y) = \frac{f_j(x)}{\int f_j(x)\,dx}. \tag{2.5}$$

The relation to (2.3) is given by $\beta_j = \sum_{k=0}^{j} \alpha_k$. Furthermore we notice that $p_0(x|y) = p(x)$ is the prior and $p_J(x|y) = p(x|y)$ is the posterior. Instead of targeting one density, the annealing process targets multiple densities sequentially in (pseudo) time.

In many applications of Bayesian inversion, it is usually impossible to draw random samples directly from the target density. An alternative method, at least for low dimensional systems, is to use importance sampling. A standard importance sampling approach consists of selecting an importance function (a density) $q(x)$ which is easy to sample from and with at least the same support as the target $p(x|y)$ (that is $q(x) > 0$ whenever $p(x|y) > 0$). Then N i.i.d. samples are drawn from $q(x)$ and for any measurable function $h(x)$, $\mathbf{E}[h(X)] = \int h(x)p(x|y)\,dx$ can be estimated from the weighted sum

$$I_N(h) = \sum_{i=1}^{N} h(X_i)w(X_i), \tag{2.6}$$

where the importance weight w is defined by $w(x) = p(x|y)/q(x)$. The function h can be used to answer questions of interest, for example if $h(x) = x$ then $I_N(h)$ estimates the mean of the target distribution and if $h(x) = 1_A(x)$ then $I_N(h)$ estimates the probability of $X \in A$ for some measurable set A with respect to the target density. Also note that since $p(x|y)$ is known only up to a normalizing constant, the importance weights have to be normalized. For complex target distributions, the naive estimator (2.6) is relatively variance prone requiring a large sample size, N, to produce an accurate estimate. The annealed importance sampling algorithm [25] tries to overcome this by sequentially targeting $p_j(x|y)$, starting with the prior at $j = 0$ and ending up with the original target at $j = J$. As the targets are slowly changing from the prior towards the posterior this algorithm is typically more stable (smaller variance) than standard importance sampling methods [14]. Since this is also a sequential Monte Carlo method [13] in pseudo time, a re-sampling step may be performed between each time step if the weights are varying too much [23]. Large variation of the weights is linked to large variance of the estimator in (2.6). In principle, at pseudo time j, one could sample from an importance density $q_j(x_j)$ and evaluate the weights $\{w(X_j^i) = p_j(X_j^i|y)/q_j(X_j^i)\}_{i=1}^{N}$. However, since the idea is to slowly move towards the target, $q_j(x_j)$ has to depend on x_{j-1}, otherwise the information from $p_{j-1}(x|y)$ is lost. Hence it is necessary to specify a transition kernel (a conditional