# Kernel embedded nonlinear observational mappings in the variational mapping particle filter

Manuel Pulido[1,2], Peter Jan vanLeeuwen[1,3] and Derek J. Posselt[4]

[1] Department of Meterology, University of Reading, UK
[2] Department of Physics, Universidad Nacional del Nordeste, Argentina
[3] Department of Atmospheric Science, Colorado State University, USA
[4] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA

**Abstract.** Recently, some studies have suggested methods to combine variational probabilistic inference with Monte Carlo sampling. One promising approach is via local optimal transport. In this approach, a gradient steepest descent method based on local optimal transport principles is formulated to deterministically transform point samples from an intermediate density to a posterior density. The local mappings that transform the intermediate densities are embedded in a reproducing kernel Hilbert space (RKHS). This variational mapping method requires the evaluation of the log-posterior density gradient and therefore the adjoint of the observational operator. In this work, we evaluate nonlinear observational mappings in the variational mapping method using two approximations that avoid the adjoint, an ensemble based approximation in which the gradient is approximated by the sample cross-covariances between the state and observational spaces the so-called ensemble space and an RKHS approximation in which the observational mapping is embedded in an RKHS and the gradient is derived there. The approximations are evaluated for highly nonlinear observational operators and in a low-dimensional chaotic dynamical system. The RKHS approximation is shown to be highly successful and superior to the ensemble approximation for non-Gaussian posterior densities.

**Keywords:** variational inference · Stein discrepancy · data assimilation.

## 1 Introduction

There is a large number of applications in which the process of interest is not directly measured, a latent process, but it is related through a map to another process which is the one observed. This problem can be framed in the classical Bayesian inference, in which the latent process is inferred from indirect noisy observations [19]. The mapping between the two processes will here be referred to as the observational mapping. Depending on the application, the observational mapping may be (partially) known through the knowledge of the physical processes involved. An example of particular interest in this work is the inference

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22747-0_11

of atmospheric state variables from satellite measurements of radiation. In other applications, the map is unknown and needs to be estimated. This is one of the central aims in machine learning applications [20].

In modeling and predicting the atmosphere, clouds play a central role. Measurements from spaceborne radars may give information on cloud properties. In this case, the observed variables are radar reflectivity and microwave radiances, while the variables of interest are cloud particle concentrations and distributions. This mapping is represented in models through parameterizations which relate cloud microphysical processes to precipitation and radiative fluxes. In several situations, the joint posterior density of model parameters and the output variables is bimodal [12]. The main factor responsible for the bimodal density is the extremely nonlinear response of model output variables to changes in microphysical parameters. The parameter prior density and observation uncertainty only play a secondary role in the resulting complexity of the posterior density.

If the latent process is governed by a time evolving stochastic dynamical system, the inference is sequential. The time evolution of the latent state is given by a Markov process–the dynamical system– while an observational mapping relates observations with the latent state. These are known as state-space models or hidden Markov models. A rather general method for inference in hidden Markov models is based on Monte Carlo sampling of the prior density, referred to as sequential Monte Carlo or particle filtering [4]. One of the major challenges in high-dimensional particle filtering is to concentrate sample points in the high-probability regions of the posterior density, the so-called typical set. In this case, they produce a non-negligible contribution to expectation estimations. Therefore, sample points are required to be located in the typical set to make the most of them.

Recent works propose to combine variational inference with Monte Carlo sampling [17]. A rigorous well-grounded framework to combine them is via local optimal transport [11, 16]. Optimal transport relates a given density with a target density trough a mapping that minimizes a risk. Hence, optimal transport concepts may be used to move sample points to locations where they maximize the amount of Shannon information they can provide. If the mapping function space is constrained to a reproducing kernel Hilbert space, the local direction that minimizes the risk, in terms of the Kullback-Leibler divergence, is well defined. This direction minimizes the Stein discrepancy [10]. An application of the variational mapping using the Stein gradient to sequential Monte Carlo methods in the framework of hidden Markov models was recently developed [16] and has been referred to as variational mapping particle filter (VMPF).

The gradient of the observation likelihood depends on the adjoint of the observational mapping. Thus, most of the approximations used for posterior inference including MAP estimation, the Kalman filter, and the stochastic and square-root ensemble Kalman filters (e.g. [1, 7]) require this adjoint of the observational mapping. However, there is a rather large number of complex observational mappings for which the adjoint is not available. In the context of the ensemble Kalman filter, an ensemble approximation of the adjoint of the

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22747-0_11

observational mapping is used [6, 7]. However, this approximation may have a detrimental effect in the inference for the highly nonlinear observational mapping of e.g. cloud parameter estimation [13–15] and in other geophysical applications [3].

A description of the VMPF in the context of observational mapping is given in Section 2. Two approximations of the adjoint of the observational mapping based on sample points evaluations of the observational operator are derived in Sections 2.1 and 2.2). Details of the experiments are given in Section 3. The VMPF with the exact gradient of the logarithm of the posterior density and the developed approximations is evaluated with nonlinear observational operators in low-dimensional spaces (Section 4). The performance of the VMPF in a chaotic dynamical system with a nonlinear observational mapping is also discussed.

## 2 Observational function with variational mappings

Suppose we want to determine a stochastic latent process $\mathbf{x}$ in $\mathbb{R}^{N_x}$, only sparse observations of another related process $\mathbf{y}$ in $\mathbb{R}^{N_y}$ are avaliable. The relationship between the processes is given through a known nonlinear observational operator $\mathcal{H}$ such that

$$\mathbf{y}_k = \mathcal{H}(\mathbf{x}_k) + \boldsymbol{\eta}_k, \tag{1}$$

where $\boldsymbol{\eta}_k$ is the random observational error which consists of realizations from a density, $p(\boldsymbol{\eta})$, that describes the measurement and representation error, $k$ denotes different realizations of the stochastic process. We assume the observational errors are unbiased, $\mathcal{E}(\boldsymbol{\eta}) = 0$.

Using Bayes rule, the density of the latent process conditioned on the realizations of the observed process is

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}). \tag{2}$$

Let us assume the prior knowledge of $\mathbf{x}$ is through a sample $\{\mathbf{x}^j, j = 1, \cdots, N_p\} \triangleq \mathbf{x}^{1:N_p}$. A standard importance sampling technique for Bayesian inference assumes that the prior density $p(\mathbf{x})$ is a proposal density of $p(\mathbf{x}|\mathbf{y})$ so that this posterior distributions is written as the sample points of the prior density with weigths given by the likelihood of the sample at the points[4]. A better proposal density may be considered assuming knowledge of the observation. In this case, weights are expected to be more equally distributed within sample points so that the variance of the weights is smaller.

Our aim is to find a sequence of mappings, $\mathbf{x}_i = T_i(\mathbf{x}_{i-1})$ that transforms from sample points of $p(\mathbf{x})$ to sample points of $p(\mathbf{x}|\mathbf{y})$ . Considering these mappings, the relationship between the transformed density after the mappings and the initial density is

$$q(\mathbf{x}_I(\mathbf{x}_0)) = \prod_{i=1}^{I} |\nabla T_i| \, q(\mathbf{x}_0), \tag{3}$$

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22747-0_11

where the initial density $q(\mathbf{x}_0)$ is in principle the prior density, while the target density of the transformations is the posterior density $p(\mathbf{x}_0|\mathbf{y})$ and $|\nabla T_i|$ are the Jacobians of the transformations.

Therefore, in order to get equally-weighted sample points that optimally represent the posterior density, we have to find a series of maps $T$ that transforms the prior into the posterior density. In terms of the particles, the goal is to drive them from the prior density to the posterior density. In this work, sample points will also be referred to as particles interchangeably. This process, of driving the particles from one to other density, could be framed as maximizing the likelihood of the particles. Alternatively, it can be formulated as a Kullback-Leibler divergence (KLD) optimization given the well-known equivalence between marginal likelihood maximization and KLD minimization. The KLD between the intermediate density and the target density is

$$\mathcal{D}_{KL}(q_T\|p) = \int q_T(\mathbf{x}) \log \left[\frac{q_T(\mathbf{x})}{p(\mathbf{x}|\mathbf{y})}\right] \mathrm{d}\mathbf{x} \tag{4}$$

The aim is to determine the *local* transformation $T$ that produces the deepest descent in KLD. The derivation of the steepest descent gradient has been already given in previous works [10, 16]. Assuming the transformation $T$ is in a reproducing kernel Hilbert space (RKHS), the gradient of the KLD is given by

$$\nabla \mathcal{D}_{KL}(\mathbf{x}) = -\int \left[K(\mathbf{x}', \mathbf{x}) \nabla \log p(\mathbf{x}'|\mathbf{y}) + \nabla_{\mathbf{x}'} K(\mathbf{x}', \mathbf{x})\right] \mathrm{d}\mathbf{x}' \tag{5}$$

where $K(\mathbf{x}', \mathbf{x})$ is the reproducing kernel and the gradient is at $\mathbf{x}$ where the local transformation is produced.

Each of the particles is moved along the steepest descent direction $\mathbf{v}(\mathbf{x}) = -\nabla \mathcal{D}_{KL}$,

$$\mathbf{x}_{i+1}^j = T_{i+1}(\mathbf{x}_i^j) = \mathbf{x}_i^j + \epsilon \mathbf{v}(\mathbf{x}_i^j). \tag{6}$$

The particles are tracers in a flow given by the KLD gradient. In essence, the objective is to determine the direction of steepest descent at each sample point and to move them along these directions. The pseudo-time step $\epsilon$ in (6) should be small enough so that the particle trajectories do not intersect and therefore the smoothness of the flow is conserved. The overall performance of the variational mapping in a sequential Monte Carlo algorithm is evaluated in [16] and is termed as the variational mapping particle filter (VMPF).

To obtain the gradient of the Kulback-Leibler divergence at a sample point (5), we require an analytical expression of the log-posterior gradient, which can be expressed in terms of the prior density and the observation likelihood using (2),

$$\nabla \log p(\mathbf{x}|\mathbf{y}) = \nabla \log p(\mathbf{x}) + \nabla \log p(\mathbf{y}|\mathbf{x}) \tag{7}$$

Assuming Gaussian observational errors, $p(\boldsymbol{\eta}) \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$, and using (2), the log-observation likelihood gradient is

$$\nabla \log p(\mathbf{y}|\mathbf{x}) = (\nabla \mathcal{H}(\mathbf{x}))^\top \mathbf{R}^{-1}(\mathbf{y} - \mathcal{H}(\mathbf{x})). \tag{8}$$

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22747-0_11

The observational operator has a major role in (8). For a linear observational mapping, a linear log observation likelihood gradient results. On the other hand, a nonlinear mapping produces a nonlinear likelihood gradient. Therefore, it induces a non-Gaussian posterior distribution. In the case of a non-injective mapping, more than one root of (8) are expected, which results in a multimodal posterior density. This rather common feature in the observational mapping is examined exhaustively in the experiments.

### 2.1 Observational mapping in the RKHS

For many applications in geophysical systems, the observational operator is a rather complex mapping that involves physical processes, for instance as mentioned the transformation from cloud properties to observed radar reflectivity. Even when the observational operator is available, the tangent-linear and adjoint operators of the observational mapping are often not available and their development and use could be costly in terms of human resources and computationally demanding in its evaluation. In this work, we derive a Monte Carlo approximation of the term $(\nabla\mathcal{H}(\mathbf{x}))^\top$ in (8). In coherence with the formulation of the variational mapping, we now also assume that the process $\mathcal{H}(\mathbf{x})$ is in the reproducing kernel Hilbert space (RKHS). This assumption is similar to that in support vector machines, where the mapping is also assumed to lie in an RKHS [20]. In that case, we can use the reproducing property for $\mathcal{H}(\mathbf{x})$,

$$\mathcal{H}(\mathbf{x}) = \langle \mathcal{H}(\mathbf{x}')|\, K(\mathbf{x}, \mathbf{x}')\rangle\,. \tag{9}$$

where $\langle\cdot|\,\cdot\rangle$ is the RKHS inner product. Using the $N_p$ particles $\mathbf{x}^{1:N_p}$ to generate a finite Hilbert space, the Monte Carlo approximation of the gradient of (9) is

$$\nabla\mathcal{H}(\mathbf{x}) \approx \frac{1}{N_p}\sum_{j=1}^{N_p}\mathcal{H}(\mathbf{x}^j)\nabla_{\mathbf{x}}K(\mathbf{x}, \mathbf{x}^j). \tag{10}$$

We have now an expression of the gradient of the observational operator that only depends on its evaluation at the particle positions. From the RKHS theory, we know that the approximated value in (10) will converge towards the exact one when $N_p \to \infty$ assuming $\mathcal{H}(\mathbf{x})$ is sufficiently smooth. Convergence of the gradient in the RKHS has been examined in [21].

The expression of the gradient of the Kullback-Leibler divergence of the VMPF (5) using a Monte Carlo integration is

$$\nabla\mathcal{D}_{KL}(\mathbf{x}) = -\frac{1}{N_p}\sum_{l=1}^{N_p}\left[K(\mathbf{x}^l, \mathbf{x})\nabla\log p(\mathbf{x}^l|\mathbf{y}) + \nabla_{\mathbf{x}^l}K(\mathbf{x}^l, \mathbf{x})\right]. \tag{11}$$

using (7) and (10) in (11), the gradient becomes

$$\nabla\mathcal{D}_{KL}(\mathbf{x}) = -\frac{1}{N_p}\sum_{l=1}^{N_p}\left\{K(\mathbf{x}^l, \mathbf{x})\left[\nabla\log p(\mathbf{x}^l) + \left(\frac{1}{N_p}\Sigma_j\mathcal{H}(\mathbf{x}^j)\nabla_{\mathbf{x}^l}K(\mathbf{x}^l, \mathbf{x}^j)\right)^\top \mathbf{R}^{-1}(\mathbf{y} - \mathcal{H}(\mathbf{x}))\right] + \nabla_{\mathbf{x}^l}K(\mathbf{x}^l, \mathbf{x})\right\}. \tag{12}$$

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22747-0_11

This expression depends only on the evaluation of the observational operator at the sample points. Therefore, the number of evaluations of the observational operator in (12) is $N_p$ at each mapping iteration. No extra evaluations from the original variational mapping are required. The Gram matrix and the gradient of the kernels are already available since they are required in the second right-hand side term of (11). In conclusion, the main complexity of the algorithm is still of order $N_p^2$ as in the original VMPF.

There is a problem for the RKHS approximation of the observational mapping (10) in the regions where the sample points are sparse. An experiment to illustrate this drawback is shown in Section 3. This problem appears because the kernel values between those sparse points and the rest of the sample points have only few points (the closest ones to the one in consideration) with non-negligible contributions and all the other kernel values are (close to) 0. Note that the square of the bandwidth is chosen to be smaller than the trace of the sample covariance to allow for more detailed structures in the density of $\mathcal{H}(\mathbf{x})$. One way to solve this problem could be using an adaptive kernel bandwidth based on the distance to the k-nearest neighbors. A simpler solution is to normalize the contributions of the kernels

$$\mathcal{H}(\mathbf{x}) \approx \frac{\sum_{j=1}^{N_p} \mathcal{H}(\mathbf{x}^j) K(\mathbf{x}, \mathbf{x}^j)}{\sum_{l=1}^{N_p} K(\mathbf{x}, \mathbf{x}^l)} \tag{13}$$

In this way, the contribution of the kernel functions evaluated at each sample point is normalized. This approximation to the gradient of the observational mapping is evaluated in the experiments.

### 2.2 Observational operator in the ensemble space

Instead of constraining the observational operator to the RKHS, it can be expressed in the ensemble perturbation space. This type of approximations is common in ensemble Kalman filtering. Indeed, the whole estimation problem may be transformed and determined in the ensemble perturbation space [7]. Here, we derive an approximate expression for the tangent-linear model, i.e. the gradient of the observational mapping, and its adjoint model based on the perturbations of the particles (ensemble members) to the mean.

The increments are approximated with a first-order Taylor series around the mean $\overline{\mathbf{x}}$

$$\mathbf{y} - \mathcal{H}(\mathbf{x}) \approx \mathbf{y} - \mathcal{H}(\overline{\mathbf{x}}) - \mathbf{H}(\mathbf{x} - \overline{\mathbf{x}}), \tag{14}$$

where $\mathbf{H}$ is the tangent-linear operator of $\mathcal{H}$ at $\overline{\mathbf{x}}$. The perturbation matrices in the state and observational spaces are composed by the differences between the ensemble members and the mean, namely

$$\mathbf{X} = \frac{1}{\sqrt{N_p - 1}} \left( \mathbf{x}^{(1)} - \overline{\mathbf{x}}, \mathbf{x}^{(2)} - \overline{\mathbf{x}}, \cdots, \mathbf{x}^{(N_p)} - \overline{\mathbf{x}} \right), \tag{15}$$

$$\mathbf{Y} = \frac{1}{\sqrt{N_p - 1}} \left( \mathcal{H}(\mathbf{x}^{(1)}) - \overline{\mathcal{H}(\mathbf{x})}, \cdots, \mathcal{H}(\mathbf{x}^{(N_p)}) - \overline{\mathcal{H}(\mathbf{x})} \right). \tag{16}$$

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22747-0_11

where $\mathbf{X}$ is an $N_x \times N_p$ matrix and $\mathbf{Y}$ is $N_y \times N_p$. The included normalization factor is $\sqrt{N_p - 1}$ to avoid bias in the sample covariance. Thus, the prior sample covariance is $\mathbf{P} = \mathbf{X}\mathbf{X}^\top$.

The approximated tangent-linear operator of $\mathcal{H}$ at the ensemble space is then given by

$$\nabla\mathcal{H}(\overline{\mathbf{x}}) = \mathbf{H} \approx \mathbf{Y}\mathbf{X}^\dagger, \tag{17}$$

where $\mathbf{P}_{yx} \triangleq \mathbf{Y}\mathbf{X}^\top$. For the adjoint approximation, the transpose of non-Gaussianity in the posterior density for the inverse model is the nonlinear- ity in the observational operator. The prior density is N (0.5, 1). The observation corresponds to a true state of 3 with an observational error of R = 0.5. The approximations are evaluated with two nonlinear observation opera- tors. We use a quadratic relationship, H(x) = x2 , which is expected to lead to a bimodal posterior density because of its non-injectivity. Non-injective ob- servational operators associate an observation with more than one state. The observation likelihood function then contains several maxima. Therefore, if the prior density is non-null for these states associated to the likelihood maxima, they result in a multimodal posterior density. For a challenging evaluation of the of (17) is used,

$$\nabla\mathcal{H}(\overline{\mathbf{x}})^\top \approx \mathbf{H}^\top = (\mathbf{X}^\dagger)^\top\mathbf{Y}^\top. \tag{18}$$

## 3 Experiments

In the observational mapping experiments, an iid sample from the prior density is given, which for simplicity is assumed in general to be normally distributed. Furthermore, observational errors are assumed Gaussian. Thus, the only source of non-Gaussianity in the posterior density for the inverse model is the nonlinearity in the observational operator. The prior density is $\mathcal{N}(0.5, 1)$. The observation corresponds to a true state of 3 with an observational error of $R = 0.5$.

The approximations are evaluated with two nonlinear observation operators. We use a quadratic relationship, $\mathcal{H}(x) = x^2$, which is expected to lead to a bimodal posterior density because of its non-injectivity. Non-injective observational operators associate an observation with more than one state. The observation likelihood function then contains several maxima. Therefore, if the prior density is non-null for these states associated to the likelihood maxima, they result in modes of the posterior density. For a challenging evaluation of the gradient approximations of the observational operator, we also use the absolute value $y = |x|$ which contains a discontinuity in its derivative. Representations of the observational mapping through a small number of basis functions is expected to give an inaccurate approximation of this derivative. Although these observational operators are only motivated in evaluating the approximations, they are indeed found in several applications. In particular, the absolute value is a frequent operator that appears when measuring wind and current speeds with instruments that are not able to distinguish flow direction.

Note that the prior density we use is not symmetric around 0, while the chosen observational operators are. Besides, the true state is in a region of very small

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22747-0_11

prior density. These choices have been taken so that the gradient approximation from sample points represents a challenge.

**Algorithm 1** Variational mapping algorithm

**Input:** Given $\mathbf{x}_0^{(1:N_p)}$, $\mathbf{y}$, $\mathcal{H}(\cdot)$, and $p(\boldsymbol{\eta})$
  **repeat** ▷ Mapping iterations
    **for** $j = 1, N_p$ **do**
      $\mathbf{x}_i^{(j)} \leftarrow \mathbf{x}_{i-1}^{(j)} - \epsilon \nabla \mathcal{D}_{KL}(\mathbf{x}_{i-1}^{(j)})$ ▷ $\nabla \mathcal{D}_{KL}$ using different particle approximations (11), (12). ▷ $\epsilon$ obtained with ADAM.
    **end for**
    i ← i+1
  **until** Stopping criterion met $|\nabla \mathcal{D}_{KL}|/|\nabla \mathcal{D}_{KL0}| < \delta$
**Output:** $\mathbf{x}_i^{(1:N_p)}$

**Algorithm 2** VMPF algorithm

**Input:** Given $\mathbf{x}_{k-1}^{(1:N_p)}$, $\mathbf{y}_k$, , $\mathcal{H}(\cdot)$, $\mathcal{M}(\cdot)$, and $p(\boldsymbol{\eta})$
  **for** $j = 1, N_p$ **do**
    $\mathbf{x}_{k,0}^{(j)} \leftarrow \mathcal{M}(\mathbf{x}_{k-1}^{(j)}, \boldsymbol{\eta}_k)$ ▷ Forecast stage
  **end for**
  **repeat** ▷ Mapping iterations
    **for** $j = 1, N_p$ **do**
      $\mathbf{x}_{k,i}^{(j)} \leftarrow \mathbf{x}_{k,i-1}^{(j)} - \epsilon \nabla \mathcal{D}_{KL}(\mathbf{x}_{k,i-1}^{(j)})$ ▷ $\nabla \mathcal{D}_{KL}$ using different particle approximations (11), (12). ▷ $\epsilon$ obtained with ADAM.
    **end for**
    i ← i+1
  **until** Stopping criterion met $|\nabla \mathcal{D}_{KL}|/|\nabla \mathcal{D}_{KL0}| < \delta$
**Output:** $\mathbf{x}_{k,i}^{(1:N_p)}$

In a last set of experiments, we evaluate the use of a nonlinear observation operator in a chaotic dynamical system. The state of the 3-variable Lorenz-63 dynamical system corresponds to the latent process. Observations are obtained with the absolute observational mapping from the latent state and a Gaussian noise. The results from the VMPF using 100 particles are compared with the SIR particle filter [4] using 1000 and 10000 particles.

The pseudo-code of the variational mapping methodology for the observational mapping is shown in Algorithm 1. A single posterior density is estimated through the mapping iterations. The particles are moved along the steepest descent direction as in traditional multidimensional optimization. However, multiple points of the cost function, i.e. KL divergence, are followed at the same time. Furthermore, the distribution of these sample points defines the gradient of the cost function in each iteration. In other words, the particles –sample points– interact during the optimization. The termination criterion of the optimization is based on the mean value of the modulus of the KLD gradient (considering all the points). The pseudo-code of the variational mapping particle filter includes the estimation of a posterior density for each cycle–each time observations are available. The posterior density at one cycle is propagated through the set of particles using the dynamical model to obtain the sample prediction density at the next observation time. Algorithm 2 shows the variational mapping particle filter pseudo-code. A detailed description of the VMPF may be found in [16].

The optimization in the VMPF is conducted through ADAM [9], a second-moment optimization method. The tuning parameters are set to the recommended values, first moment parameter $\beta_1 = 0.9$ and second moment parameter $\beta_2 = 0.99$. The learning rate was set to 0.03. The maximum number of opti-

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22747-0_11

mization iterations is set to 500 (this is not reached in any of the experiments), and the criterion for termination is based on the mean value of $|\nabla \mathcal{D}_{KL}|$, the required threshold is $|\nabla \mathcal{D}_{KL}|/|\nabla \mathcal{D}_{KL1}| < 0.01$ where $|\nabla \mathcal{D}_{KL1}|$ corresponds to the first iteration. The required number of iterations under these settings is about $100 - 150$ in the observational mapping experiments and about 50 iterations in the dynamical experiments, however a few cycles may require more than 200 iterations. For more computationally consuming experiments, a higher learning rate may be more efficient. However, we have priorized the smoothness of the mappings in these proof-of-concept experiments.

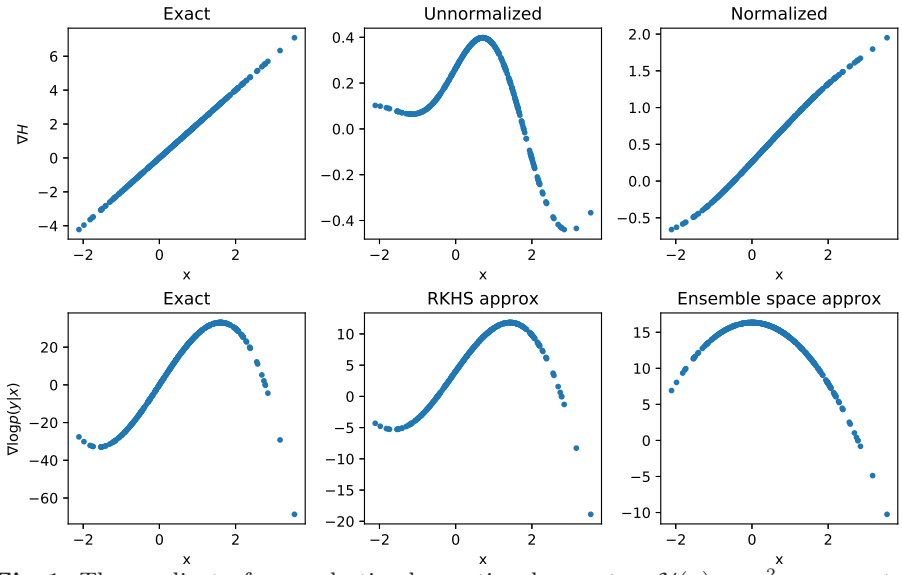


**Fig. 1.** The gradient of a quadratic observational operator, $\mathcal{H}(x) = x^2$, represented by the samples of the prior density for the exact calculation (left upper panel), the approximation using unnormalized kernels (middle upper panel) and using normalized kernels (right upper panel). The gradient of the log-observation likelihood, (8) with exact gradient (left lower panel), normalized RKHS approximation (middle lower panel) and ensemble approximation (right lower panel).

For all the experiments, radial basis functions are used as kernels. A Mahalanobis distance is taken, $K(\mathbf{x}, \mathbf{x}') = \exp\left(-\|\mathbf{x} - \mathbf{x}'\|^2_{\mathbf{A}}\right)$. The Mahalanobis matrix $\mathbf{A}$, hereinafter referred to as kernel covariance, is chosen proportional to the prior sample covariance in the observational mapping experiments and the model error covariance in the dynamical system experiment. The proportionality factor, which could be interpreted as the bandwidth of an isotropic kernel, is determined with the Scott rule. However, some extra manual tuning of it was required for some of the experiments.

## 4 Results

Figure 1 shows the results of the derivative of the quadratic observational operator (upper panels) represented by using the sample points of the prior density.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22747-0_11

The exact calculation is shown in left upper panel of Figure 1. The approximation in the RHKS using unnormalized kernel functions in the finite space, (10), is in the middle panel and the one using normalized kernels, (13), in the right upper panel. The approximation to the mapping gives small values in isolated sample points because of only a few points contribute to the kernel integration in sparse areas. The normalization factor incorporates weights according to the density of points around the samples, producing a better estimate of the functional representation of the mapping and its derivative. However, note that some smoothing is still found in the extremes which results in approximated derivative values smaller that the true ones. This effect in the sparse sample points should manifest in strong convex functions as the one used in the evaluation. The normalized kernel approximation–apart from the amplitude– gives a rather good functional dependence. There are some minor deviations in the functional dependence mainly produced by the asymmetry introduced between the sampling and the observational mapping.

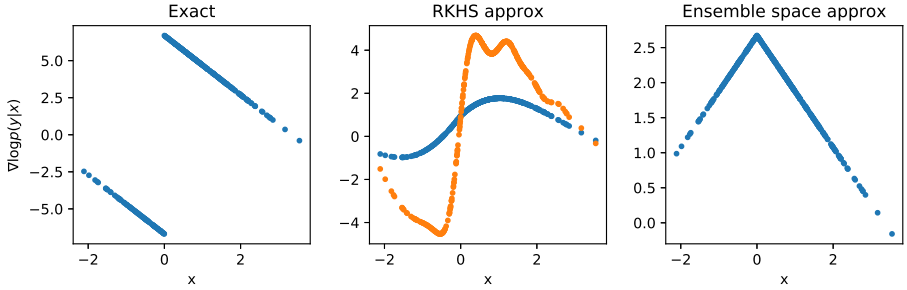


**Fig. 2.** Gradient of the observation mapping $\mathcal{H}(x) = |x|$ for the exact calculation, the approximation using kernels and the approximation using perturbations in the ensemble space. Two kernel bandwidths, $\gamma = 1$ (blue dots) and $\gamma = 0.3$ (orange dots) are shown for the RKHS approximation.

The impact of approximating $\nabla\mathcal{H}$ on the gradient of the observation likelihood is shown in the lower panels of Fig. 1. The overall structure using the RKHS approximation is recovered. However, the amplitude of the gradient is underestimated. The ensemble space approximation gives a constant gradient of the observational mapping independent of the sample points, (17), which is expected to give the mean gradient of the mapping. In terms of the gradient of the observation likelihood, it results in a quadratic function, because of the increment term in (8) between observations and the particles. This would only be a good approximation of the true observation likelihood gradient (left panel) close to the observation. For methods that only give the maximum a posteriori solution, a relatively coarse representation of the gradient of the log-likelihood function may be enough to give a good estimation. Thus, they only require a precise gradient of the observational operator close to the observation. On the other hand, an accurate representation in a larger region is required if the inference problem also deals with uncertainty quantification.

Results for the absolute observational operator are shown in Fig. 2. Gaussian kernels act as smoothers (e.g. [20]), so that the approximation with Gaussian

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22747-0_11

kernels to the absolute function is a smooth function and so the derivative is similar to a tanh-function with a smooth transition between the positive and negative values. The transition can be more abrupt if the kernel bandwidth is reduced from $\gamma = 1$ to 0.3 (middle panel in Fig. 2). However, the sampling noise is increased in that case. Note also that the amplitude of the function approximation is closer to the true one for the narrower kernel bandwidth. A narrower kernel bandwidth uses less sample points to approximate the mapping. Hence, it diminishes the smoothing. The ensemble space average produces a correct gradient of the log-likelihood close to the observations in the positive state values, but a wrong one for negative state values (lower right panel). Because the amplitude in $\nabla\mathcal{H}$ results from an average of all the sample points, it is underestimated in the absolute mapping and so in the gradient of the log-likelihood.

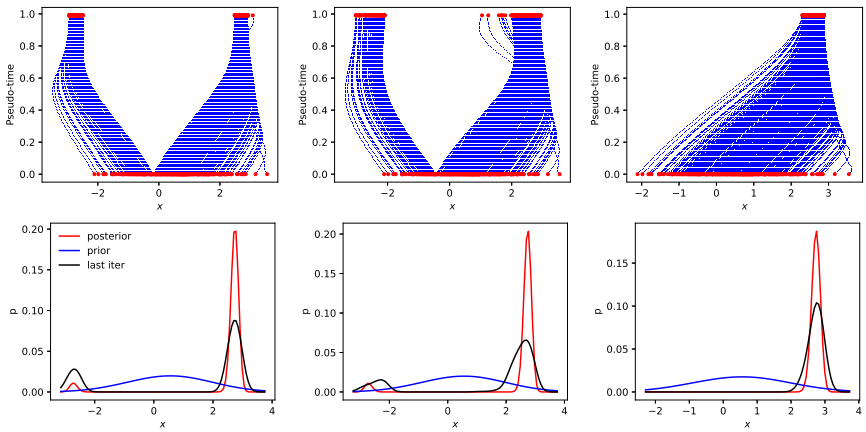


**Fig. 3.** Evolution in pseudo-time of the sample points for a quadratic observational mapping (upper panels) for the experiment with exact gradient (left panel), RKHS approximation (middle panel) and ensemble approximation (right panel). Posterior density (lower panels) for the exact quadratic observational mapping (red line) and the one obtained with VMPF (black line)

Figure 3 exhibits the trajectories of the sample points as a function of pseudo-time between the initial iteration of the filter (representing the prior density) and up to the convergence criterion is met which is based on the module of the gradient of the Kullback-Leibler divergence. The experiment corresponds to the quadratic mapping. In both the exact and the RKHS approximation, samples are attracted to two different positive/negative regions which represent a bimodal posterior density. Because of the asymmetry in the prior density (whose the mean is 0.5) more particles are attracted to the positive region. The particles finish more disperse in the RKHS approximation than in the exact gradient calculation. The ensemble space approximation for the gradient of the observational mapping removes the bimodality of the posterior density and the particles are only attracted by the dominant mode (right panel). Lower panels in Fig. 3 compare the analytical posterior density with the one obtained with the VMPF, representing the final sample with kernel density estimation in coherence with the RKHS used in the mappings. The VMPF using the exact observational map-

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22747-0_11

ping is shown in the left panel of Fig. 3, the one using the RKHS approximation (middle panel) and the ensemble approximation (right panel). The exact case shows some smoothing of the main mode mainly because the observation is at a low density region of the likelihood. Tests with a narrower kernel bandwidth diminish the effect but it does not disappear. In the case of the RKHS approximation, there is some spread of the sample points toward lower values. This effect may be linked to the lower values of the gradient of the likelihood in this approximation. The ensemble approximation removes the smaller mode and only represents the main one. The slightly wider representation of uncertainty around the main mode is mainly controlled by the kernel bandwidth.

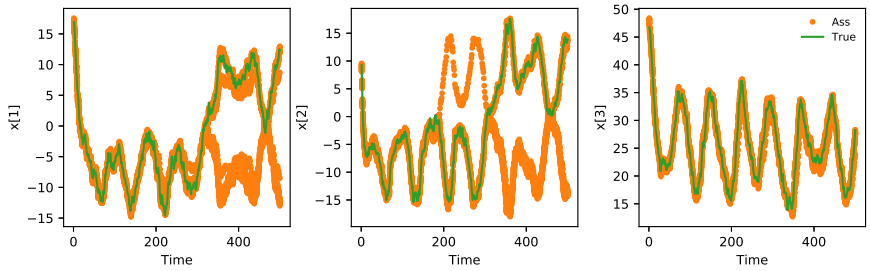


**Fig. 4.** The temporal evolution of the true state variables of the stochastic Lorenz-63 dynamical system (green line) and trajectories (40) of the particles resulting from the VMPF (orange dots), namely $\mathbf{x}_{1:K}^{(1:40)}$ as resulting from the output of Algorithm 2. Panels show each variable of the Lorenz 63 system. Time units are cycles.

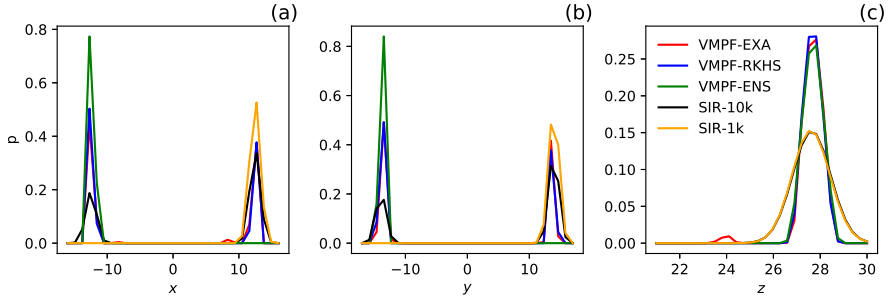


**Fig. 5.** Marginalized sequential posterior density represented through kernel density estimation, obtained with the VMPF using the exact (VPMF-EXA), RKHS (VMPF-RKHS) and ensemble calculations (VMPF-ENS) of the adjoint. The densities from SIR particle filter with 1000 (SIR-1k) and 10000 particles (SIR-10k) are also shown. Panels show marginalized density as a function of each variable for the Lorenz 63 system at the 500 cycle.

Figure 4 shows the evolution of the three variables of the Lorenz-63 system for a selected set of particles from VMPF (orange dots) and the true trajectory (green line). Because the apriori density at the initial time is prescribed as Gaussian, the posterior density evolves as unimodal until the true state changes of attractor. This occurs at about the cycle 300. From that time, trajectories of the VMPF particles are located in both attractors because the absolute observations cannot distinguish in which attractor the system is. In other words,

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22747-0_11

the subsequent posterior density evolution from $t = 300$ undergoes a transition to a bimodal density. Figure 5 shows the marginal posterior densities in each variable for the VMPF at cycle 500 . Both the exact calculation and the RKHS approximation in the observation likelihood gradient in the VMPF are able to capture the bimodality of the posterior density using 100 particles. On the other hand, the ensemble approximation only gives an unimodal density. For comparison we also show in Fig. 5 the corresponding marginalized posterior density of the SIR particle filter with 1000 and 10,000 particles. The SIR filter requires 10,000 particles to capture the bimodal structure of the posterior density while VMPF only requires 100 particles.

## 5 Conclusions

This work evaluates the use of a nonlinear observational operator in the variational mapping particle filter. Non-injectivity of the observational mappings leads to multimodal posterior densities which is known to represent a challenge for inference methods. The variational mapping particle filter is able to capture multimodes in the density in offline and online experiments. Particles are attracted to the modes in coherence with the gradient of the posterior density and local optimal transport principles.

Two approximations of the gradient of the observation mapping are evaluated. The representation of the observational mapping in the RKHS which overall exhibits a good performance in non-Gaussian densities. Because of the smoothing associated with this representation, it may slightly shift the modes in multi-modal densities for cases in which observations are in low-density regions of the prior density. The evaluation with the Lorenz-63 shows that the impact of this smoothing in a sequential scheme is negligible even for the absolute value observational mapping–a discontinuous gradient. The ensemble approximation of the gradient as expected does not capture multimodality, but it gives a good approximation of the main mode of the posterior density and its uncertainty.

We have not considered here other approximations which could require further evaluations of the observational mapping apart from the sample points to estimate the gradient of the mapping. One of these possibilities is the evaluation of the gradient at each sample point from finite differences. For applications of moderate dimensions and complex observational mapping the computational cost of these further evaluations required at each iteration of the variational inference algorithm and at each sample is prohibitive. The RKHS approximation of the observational operator is expected to be affected by the curse of dimensionality for high-dimensional states. A potential way to circunvent this limitation could be to divide the state space in the variables which are close to linear dependence from those state variables with a significant nonlinear observational function response. In this case, the partial derivatives of quasi-linear variables may be approximated with the ensemble approximation while the derivatives of highly nonlinear variables may be obtained through the RKHS approximation in the lower-dimensional subspace.

ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22747-0_11

In all the experiments, radial basis functions are used as kernels. We took this choice because the structure of errors was assumed Gaussian. On the other hand, the kernel covariance and in particular the bandwidth are key hyperparameters for a good performance of the inference. In the proof-of-concept experiments an expensive trial-and-error methodology is used to manually tune the hyperparameters. Adaptative estimates of the hyperparameters are highly required. Standard adaptative bandwidth selection[18] does not appear a good option for non-injective observational mappings.

## References


1. Burgers, G., Jan van Leeuwen, P. and Evensen, G.: Analysis scheme in the ensemble Kalman filter. *Monthly weather review*, **126**, 1719-1724 (1998).
2. Daum, F. and Huang, J.: Nonlinear filters with log-homotopy. *In Signal and Data Processing of Small Targets 2007*. **6699**, p. 669918 (2007).
3. Evensen, G.: Analysis of iterative ensemble smoothers for solving inverse problems. *Computational Geosciences*, **22**, 885-908 (2018).
4. Gordon, N.J., Salmond, D.J. and Smith, A.F.: Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F-radar and signal processing*, **140**, 107-113 (1993).
5. Hoffman, M.D., Blei, D.M., Wang, C. and Paisley, J.: Stochastic variational inference. *The Journal of Machine Learning Research*, **14**, 1303-1347 (2013).
6. Houtekamer, P.L. and Mitchell, H.L.: A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, **129**, 123-137 (2001).
7. Hunt, B.R., Kostelich, E.J. and Szunyogh, I.: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D*, **230**, 112-126 (2007).
8. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S. and Saul, L.K.: An introduction to variational methods for graphical models. *Machine learning*, **37**, 183-233 (1999).
9. Kingma, D. and Ba, J.: Adam: A method for stochastic optimization. *In Int. Conf. on Learning Repres. (ICLR)* arXiv preprint arXiv:1412.6980 (2015).
10. Liu, Q. and Wang, D.: Stein variational gradient descent: A general purpose bayesian inference algorithm. *In Advances In Neural Information Processing Systems*, 2378-2386 (2016).
11. Marzouk Y, T Moselhy, M Parno, A Spantini (2017) An introduction to sampling via measure transport. To appear in *Handbook of Uncertainty Quantification*; R. Ghanem, D. Higdon, and H. Owhadi, editors; Springer. arXiv:1602.05023
12. Posselt, D.J. and Vukicevic, T.: Robust characterization of model physics uncertainty for simulations of deep moist convection. *Monthly Weather Review*, **138**, 1513-1535 (2010).
13. Posselt, D.J. and Bishop, C.H.: Nonlinear parameter estimation: Comparison of an ensemble Kalman smoother with a Markov chain Monte Carlo algorithm. *Monthly Weather Review*, **140**, 1957-1974 (2012).
14. Posselt, D. J., D. Hodyss, and C. H. Bishop: Errors in Ensemble Kalman Smoother Estimates of Cloud Microphysical Parameters, *Mon. Wea. Rev.*, **142**, 1631-1654 (2014).
15. Posselt, D. J.: A Bayesian Examination of Deep Convective Squall Line Sensitivity to Changes in Cloud Microphysical Parameters. *J. Atmos. Sci.*, **73**, 637-665 (2016).
16. Pulido M., and P. J. vanLeeuwen: Kernel embedding of maps for Bayesian inference: The variational mapping particle filter. Submitted. https://arxiv.org/pdf/1805.11380 (2018).
17. Saeedi, A., Kulkarni, T.D., Mansinghka, V.K. and Gershman, S.J.: Variational particle approximations. *The Journal of Machine Learning Research*, **18**, 2328-2356 (2017).
18. Scholkopf, B. and Smola, A.J.: *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press (2002).
19. Tarantola, A.: *Inverse problem theory and methods for model parameter estimation* (Vol. 89). SIAM (2005).
20. Vapnik, V.: *The nature of statistical learning theory*. Springer science & Business Media (2013).
21. Zhou, D.X.: Derivative reproducing properties for kernel methods in learning theory. *Journal of computational and Applied Mathematics*, **220**, 456-463 (2008).


ICCS Camera Ready Version 2019
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-22747-0_11