Bayesian data assimilation with Transport Map sampling and PGD model order reduction

P-B. Rubio, F. Louf and L. Chamoin

LMT, ENS Paris-Saclay, 61 Avenue du Président Wilson, 94235 Cachan Cedex, France
E-mail: prubio@ens-paris-saclay.fr, louf@ens-paris-saclay.fr, chamoin@ens-paris-saclay.fr

Abstract. The motivation of this work is to address real-time sequential inference of parameters with a full Bayesian formulation. The Transport Maps method allows to determine a coupling between a reference density and the posterior density. Here, the contribution is to use the Proper Generalized Decomposition method (PGD) to reduce the evaluation cost of the multi-parametric numerical model and to speed up the minimization algorithm in the Transport Map method by calculating the models derivatives with reference to the parameters in a straightforward manner with non-analytical models. A numerical example highlights the performance of the method.

Due to the constant advances in the development of numerical models, these are increasingly used to study physical systems. In this context model updating with physical observations is necessary to ensure the relevance of the simulations. The data assimilation can be performed in a single shot, or sequentially when dynamical and real-time updating of model parameters is of interest [1]. In this work we focus on the latter approach.

However, the determination of parameters from indirect noisy observations can lead to ill-posed inverse problems [2]. To solve this issue, Bayesian inference provides a natural regularization by assigning a probability weight to the set of parameter variations according to the uncertainties considered (measurement noise, model error, stochastic parameters) [5]. In this context, the parameters to be inferred are considered as random variables and the result of the inference is the posterior probability density. The posterior density is defined by the Bayes formula as the product of a prior density that does not take into account the knowledge of the measurements and a likelihood function that gives the probability that the model output coincides with the measurement for a given parameter value. This density provides global information on the parametric space and exploration methods are necessary to estimate quantities of interest (mean, variance, first order marginal,...). These estimators require the calculation of large dimension integrals. For this purpose, Monte-Carlo integration is used by generating samples by indirect sampling methods such as the Markov Chain Monte-Carlo (MCMC) method. This method involves the evaluation of the posterior density a very large number of times making its use impossible in real time [4]. Here, the idea is to consider the Bayesian formulation without simplifying hypothesis and to couple this formulation with the Transport Map method [7] and PGD model reduction methods [6].

First, the Transport Map method allows one to build a deterministic application between a probability measure (e.g.: standard normal distribution) and the posterior measure [8] [9]. This
application permits to sample the posterior density from the known sampling of the reference density. Then the PGD model reduction method is used at in two stages. The PGD is first used in the Bayesian formulation by representing the solution of the multi-parametric numerical model in an analytical way, with separated variables, calculated in an “offline” phase. Thus, the formulation of the posterior density becomes explicit with respect to the parameters and can be evaluated at a lower cost. The PGD is also used in a second stage in transport maps calculations that require the solution of a minimization problem. Indeed, thanks to the explicit formulation of the model, gradient and Hessian information can be advantageously employed in the minimization algorithm. The effectiveness of this approach is illustrated on a numerical example in the context of a welding control process.

1. Posterior sampling in Bayesian data assimilation

1.1. Bayesian formulation of inverse problems

In this paper the purpose is to characterize the Probability Density Function (PDF) of some stationary model parameters \( p \in P \) with indirect noisy measurements \( d_{\text{obs}} \). In this case, the Bayesian formulation of the inverse problem reads [3]:

\[
\pi(p|d_{\text{obs}}) = \frac{1}{C} \pi(d_{\text{obs}}|p).\pi(p)
\]

with the normalization constant \( C = \int \pi(d_{\text{obs}}|p).\pi(p)dp \). The result of the inference is the posterior density \( \pi(p|d_{\text{obs}}) \). It gives the probability of the parameters of interest \( p \) knowing the observations \( d_{\text{obs}} \). This density is proportional to the product of the prior density \( \pi(p) \) by the likelihood function \( \pi(d_{\text{obs}}|p) \). The prior density relates to the knowledge on the parameters before the assimilation of the data \( d_{\text{obs}} \). The likelihood function is linked to the probability for the model to give observations \( d_{\text{obs}} \) for a given value of the parameters \( p \). In the classical case where an additive measurement noise with density \( \pi_{\text{meas}} \) is considered, the likelihood function is the distance between the observation and the model output weighted by the measurement noise:

\[
\pi(d_{\text{obs}}|p) = \pi_{\text{meas}}(d_{\text{obs}} - M(p))
\]

In the case of sequential assimilation of measurements \( \{d_{1_{\text{obs}}, \ldots, d_{i_{\text{obs}}}}\} \) at time steps \( t_i \), \( i \in \{1, \ldots, N_t\} \) the Bayesian formulation is given by considering the prior at time \( t_i \) as the posterior at time \( t_{i-1} \):

\[
\pi(p|d_{1_{\text{obs}}, \ldots, d_{i_{\text{obs}}}}) \propto \prod_{j=1}^{i} \pi_{t_j}(d_{j_{\text{obs}}}|p).\pi(p)
\]

For a given set of measurements \( d_{j_{\text{obs}}} \), under the same hypothesis of measurement noise as previously, the likelihood function at time \( t_j \) reads:

\[
\pi_{t_j}(d_{j_{\text{obs}}}|p) = \pi_{\text{meas}} \left( d_{j_{\text{obs}}} - M(p, t_j) \right)
\]

In this formulation no assumptions is made on the densities (prior, measurement noise) nor the linearity of the model.

1.2. Transport Maps sampling

The principle of the Transport Maps strategy is to build a deterministic coupling \( M \) between the reference probability measure \( \nu_\rho \) and the posterior measure \( \nu_\pi \) derived from the Bayesian formulation. The purpose is to find the change of variables such that:

\[
\int gd\nu_\pi = \int g \circ M d\nu_\rho
\]
In this framework, it is possible to transport samples drawn according to the reference density in order to become samples drawn according the target density (Figure 1). In the considered inference problems the target density will be the posterior density $\pi(p|d_{\text{obs}})$ and the standard normal Gaussian density will be chosen as the reference density.

A preliminary work about optimal transports has been developed in [12]. Recently, this work was adapted to Bayesian inference [9] with effective computation tools (see http://transportmaps.mit.edu). From the reference density $\rho$, the purpose is to build the map $M : \mathbb{R}^d \rightarrow \mathbb{R}^d$ such as:

$$\nu_\pi \approx M_* \nu_\rho = \rho \circ M^{-1} |\det \nabla M^{-1}|$$

To quantify the difference between the two distributions $\nu_\pi$ et $M_* \nu_\rho$ ($_*$ denotes the push forward operator), the Kullback-Leibler divergence $D_{KL}$ is used:

$$D_{KL}(M_* \nu_\rho || \nu_\pi) = \mathbb{E}_\rho \left[ \ln \frac{\nu_\rho}{M_* \nu_\pi} \right]$$

Maps $M$ are searched among Knothe-Rosenblatt rearrangement. This particular choice of structure is motivated by the properties of unique minimizer of (7), optimality regarding the weighted quadratic cost and computational feasibility. The maps $M$ are parameterized as:

$$M(p) = \begin{bmatrix} M^1(p_1) \\ M^2(p_1, p_2) \\ \vdots \\ M^d(p_1, p_2, \ldots, p_d) \end{bmatrix}$$

with: $M^k(a_c^k, a_e^k, p) = \Phi_c(p) a_c^k + \int_0^{p_k} (\Phi_e(p_1, \ldots, p_{k-1}, \theta) a_e^k)^2 d\theta$. Functions $\Phi_c$ et $\Phi_e$ are Hermite polynomials with coefficients $a_c$ et $a_e$. Eventually, with this parameterization the map $M$ is found by minimizing the K-L divergence [7]. By using a quadrature rule $(\omega_i, p_i)_{i=1}^N$ for $\rho$ (Monte-Carlo, Gauss) the minimization problem finally reads:

$$\min_{a_c^1, \ldots, a_c^d, a_e^1, \ldots, a_e^d} \sum_{i=1}^N \omega_i \left[ -\log(\tilde{\pi} \circ M(a_c^1, \ldots, a_c^d, a_e^1, \ldots, a_e^d, p_i)) - \log(|\det \nabla M(a_c^1, \ldots, a_c^d, a_e^1, \ldots, a_e^d, p_i)|) \right]$$

Figure 1. Transport Maps principle

3
The target density is replaced by its non-normalized version $\bar{\pi}$ without modifying the solution. In the case of sequential inference, the Transport Map method exploits the Markov structure of the posterior density. Indeed, instead of being computed fully, the map between the reference density $\rho$ and the posterior density at time $t_i$ is obtained by composition of low-order maps:

$$(M_1 \circ \ldots \circ M_i)_\sharp \rho(p) = (M_i \circ \ldots \circ M_1)_\sharp \rho(p) \approx \pi(p|d_{obs}^i, \ldots, d_{obs}^1)$$

(10)

The map $M_1$ represents the coupling between the density $\rho(p)$ and the first posterior density $\pi^i(p|d_{obs}^1) \propto \pi_t^1(d_{obs}^1|p)\pi(p)$. Then, each map $M_i$, $i \in \{2, \ldots, N_t\}$ is computed between $\rho$ and the density $\pi^*_i$ defined by:

$$\pi^*_i(p) = \pi_{t_i}(d_{obs}^i|M_{i-1}(p))\rho(p)$$

(11)

Once a map $M$ is computed, the quality of the approximation $M_\sharp \nu_\rho$ of the measure $\nu_\pi$ can be estimated by the convergence criteria $\epsilon_\sigma$ defined in [9] as:

$$\epsilon_\sigma = \frac{1}{2} \text{Var}_\rho \left[ \ln \frac{\nu_\rho}{M_\sharp^{-1} \nu_\pi} \right]$$

(12)

The computation cost of this criterion is very low as the integral is computed with the same quadrature rule used in the computation of the K-L divergence. Thus, an adaptability strategy regarding the order of the maps can be used to derive an automatic algorithm to deal with sequential inference problems.

2. Using PGD model order reduction in Bayesian inference

In the formulation (3), the posterior density can be explicitly expressed as a function of the parameters $p$ if the model is also explicit regarding the parameters. However, in complex engineering applications, the models are derived from the numerical solution of Partial Differential Equations (PDEs). Those PDEs can depend on parameters $p$, time $t$ and space $x$. Thus, a direct multi-parametric solution is not available for real-time applications. Hence the use of reduced order models is necessary. In the classical version, the PGD model is built directly from the weak formulation of the considered PDE. Then, the global solution $u_m$ of the PDE in terms of time, parameters and space is computed in a separated form [6]:

$$u_m(x, t, p) = \sum_{k=1}^{m} \Lambda_k(x)\lambda_k(t) \prod_{i=1}^{d} \alpha_{ik}(p_i)$$

(13)

The computation of the PGD model can be performed in an offline phase and it can then be evaluated for all time, parameters and space by evaluation of products and sums of one dimensional functions (modes).

2.1. Transport Map sampling with PGD models

Once the PGD approximation $u_m(x, t, p)$ is built, an explicit formulation of the non-normalized posterior density can be derived. By the observation operator $O$ we extract the output $d_m(p, t) = O(u_m(x, t, p))$ from the field $u_m(x, t, p)$. In this case, the non-normalized posterior density $\bar{\pi}$ reads:

$$\bar{\pi}(p|d_1^\text{obs}, \ldots, d_i^\text{obs}) = \prod_{j=1}^{i} \pi_{\text{meas}} \left( d_j^\text{obs} - d_m(p, t_j) \right) \pi(p)$$

(14)

This leads to cost-effective evaluations of the non-normalized density required by sampling methods (for example MCMC [10]). Moreover, PGD is highly beneficial in the sampling
procedure using transport maps. Indeed, the Transport Maps method is based on the minimization of the functional (9) which depends of \( \bar{\pi} \). With the PGD formulation, partial derivatives of the model with reference to the parameters \( p \) required by the computation of the derivatives of the functional (9) can be easily computed as:

\[
\frac{\partial^n u_m}{\partial p_j^n}(x, t, p) = \sum_{k=1}^{m} \Lambda_k(x) \lambda_k(t) \frac{\partial^n \alpha_{jk}}{\partial p_j^n}(p_j) \prod_{i=1}^{d} \alpha_{ik}(p_i)
\]

(15)

The parameter modes \( \alpha_{jk} \) being most often finite elements functions, the derivations are performed on one-dimensional shape functions. As a result, the problem (9) can be effectively solved by means of minimization algorithms using gradient or Hessian information which speeds up transport maps computations.

3. Numerical example

The considered problem is a welding control example introduced in [13]. Two metal plates are welded by a heat source whose center is moving along the geometry. To solve the problem, the coordinates system is moving at the same speed as the heat source. Thus, a convective term is added to the heat equation:

\[
\frac{\partial T}{\partial t} + v(Pe) \nabla T - \kappa \Delta T = s(\sigma)
\]

(16)

The problem unknown is the non-dimensional temperature \( T \) which is equal to 0 when the temperature is equal to the room temperature and 1 when the temperature is equal to the melting temperature of the material. On the boundary \( \Gamma_D \) (Figure 2) the temperature is supposed to be equal to the ambient temperature (\( T = 0 \)). The other boundaries are assumed insulated.

![Figure 2. Welding control example](image)

3.1. Inference problem

The stationary parameters of interest are \( \sigma \) et \( Pe \), which are respectively related to the spatial spreading and the speed of the volumic heat source (see (16)). The multi-parametric PGD solution of the problem detailed in [11] reads:

\[
T_k(x_k, y_k, t, \sigma, Pe) = \sum_{n=1}^{m} \Lambda_n(x_k, y_k) \lambda_n(t) \alpha_n^1(\sigma) \alpha_n^2(Pe), k=1,2,3
\]

(17)

\( T_1 \) and \( T_2 \) represent the temperatures at the measurement points and \( T_3 \) is the output of interest assumed to be unreachable by direct measurement (see Figure 2). From the measurement the purpose is to assess if the welding depth is sufficient (i.e: \( T_3 \geq 1 \)). For each time step \( t_i \), \( i \in \{1, ..., N_t\} \), the measurement \( T_1^{obs} \) and \( T_2^{obs} \) are assimilated to refine the knowledge of the parameters. The prior density on the parameters \( (\sigma, Pe) \) is supposed to be the product
of two independent Gaussian densities with means \((\mu_\sigma = 0.4, \mu_{Pe} = -60)\) and variances \((\sigma^2_\sigma = 0.003, \sigma^2_{Pe} = 7)\). Thus, at time step \(t_i, i \in \{1, \ldots, N_t\}\) the posterior density reads:

\[
\pi(\sigma, Pe|T^{obs,1:i}_1, T^{obs,1:i}_2) = \prod_{j=1}^{i} \pi_{t_j}(T^{obs,j}_1, T^{obs,j}_2|\sigma, Pe).\pi(\sigma, Pe)
\] (18)

3.2. Results

The PGD-Transport Map strategy is applied to this example. The solution of the heat equation (16) is used in its PGD form and derivatives of this solution with respect to the parameters to infer are computed in order to derive efficiently the transport maps. In Table 1 we represent the computation time required to compute at each assimilation step the applications \(M_1, \ldots, M_{N_t}\). The computation time shown also takes into account the transport of 20000 samples according to the reference density to obtain as many samples according to the posterior density for post-processing purposes. We compare the computation time with the different derivative orders provided to the minimization algorithm. With order 0 the minimization problem (9) is solved using a BFGS algorithm where the gradient is computed numerically. With order 1 the minimization is performed also using a BFGS algorithm but with the gradient given explicitly with respect to the PGD modes derivatives. With order 2 a conjugate gradient algorithm is used with an explicit formulation of both gradient and Hessian. The stopping criteria is a tolerance of \(10^{-3}\) on the variance diagnostic (12). The complexity of the maps (order of the Hermite polynomials) is increased until this tolerance is fulfilled. The first assimilation step is the most expensive because of the greater complexity of the transformation between the reference and the first posterior density (4-th order polynomials). The other transformations are much less expensive (time less than 1 second) because they can be easily represented by a linear transformation.

| Derivatives order information: | 0 | 1 | 2 |
|-------------------------------|---|---|---|
| Number of iterations for step 1 | 107 | 33 | 10 |
| Computation time for step 1   | 33.85 s | 6.18 s | 4.60 s |
| Average number of iterations for steps \(\{2, \ldots, 45\}\) | 4.2 | 4.16 | 4.13 |
| Average computation time for steps \(\{2, \ldots, 45\}\) | 1.24 s | 0.92 s | 0.90 s |

Table 1. Computation costs of the transport maps depending on the derivatives order information given to the minimization algorithm.

Using gradient and Hessian information to solve the minimization problem needed to compute the transport maps leads to low computation times. As a comparison, the generation of 20000 samples with a classical MCMC method leads to a computation time of about 5 s. This time is close to the time required to compute the first transport map but much higher than the time to compute the other maps. Futhermore the 20000 samples generated by the Transport Map are independent which is not the case of the MCMC method. In the chain generated by MCMC method, the Integrated AutoCorrelation Time (IACT) is about 12. This means that the number of samples to generate in order to have 20000 independent samples is \(20000 \times \text{IACT}\) which takes about 60 s to compute. This consideration shows the benefit of using transport maps. Futhermore, the use of the PGD allows to have a large speed-up as regards the classical MCMC method.

Figure [3] represents all the marginals for each time step computed with the samples generated by the Transport Map method. The x-axis represents the time steps and the y-axis the parameters values. The color map gives the information on the probability density function.
values. Measurements are simulated with the model computed with the reference values \((\sigma = 0.4, Pe = -60)\).

![Graph](image)

**Figure 3.** Marginals computed with 20000 samples for each time step.

4. Conclusion
The method presented allows one to deal with sequential inference problems within the general framework of Bayesian inference. The Transport Map method is a deterministic method to sample from the posterior density with a clear convergence criterion. The PGD model order reduction method reduces model call costs and facilitates the calculation of the transport maps. In this context, the Transport Map method provides samples from the posterior density with a very low computation cost. The numerical example provided highlights the interest of the method, making it an attractive approach to consider real-time applications.

References
[1] Darema F. Dynamic Data Driven Applications Systems: A New Paradigm for Application Simulations and Measurements. *Computational Science - ICCS 2004*: 662–669.
[2] Tarantola A. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, 2005.
[3] Kaipio J, Somersalo E, Antman ESS, et al. Statistical and Computational Inverse Problems. 2004; 160: 346.
[4] Robert CP, Casella G. *Monte Carlo Statistical Methods*. Springer Texts in Statistics New York, NY: Springer New York, 2004.
[5] Stuart AM. Inverse problems: A Bayesian perspective. *Acta Numerica* 2010; 19: 451–559.
[6] Chinesta F, Keunings R, Leygue A. The proper generalized decomposition for advanced numerical simulations: A primer. In: No. 9783319028644. 2014 (pp. iv–v).
[7] El Moselhy TA, Marzouk YM. Bayesian inference with optimal maps. *Journal of Computational Physics* 2012; 231(23): 7815–7850. doi: 10.1016/J.JCP.2012.07.022.
[8] Marzouk Y, Moselhy T, Parno M, Spantini A. An introduction to sampling via measure transport. 2016.
[9] Spantini A, Bigoni D, Marzouk Y. Inference via low-dimensional couplings. 2017.
[10] Berger J, Orlande HRB, Mendes N. Proper Generalized Decomposition model reduction in the Bayesian framework for solving inverse heat transfer problems. *Inverse Problems in Science and Engineering* 2017; 25(2): 260–278. doi: 10.1080/17415977.2016.1160395.
[11] Rubio PB, Louf F, Chamoin L. Fast model updating coupling Bayesian inference and PGD model reduction. *Computational Mechanics* 2018: 1–25. doi: 10.1007/s00466-018-1575-8.
[12] Villani C. *Optimal Transport: Old and New*. Media 2007. doi: 10.1007/978-3-540-71050-9.
[13] Grepl MA. *Reduced-Basis Approximation and A Posteriori Error Estimation for Parabolic Partial Differential Equations*. PhD thesis. Massachusetts Institute of Technology, 2005.