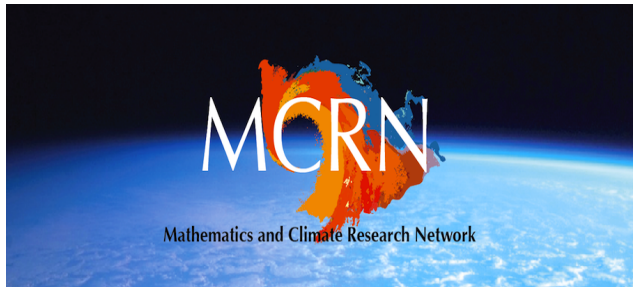
Efficient Mean Field Variational Algorithm for Data Assimilation

Michail D. Vrettas

m.vrettas@berkeley.edu

joint work with

Dan Cornford (IGI Ltd & Aston University) and Manfred Opper (TU Berlin)



University of California, Berkeley

http://www.berkeley.edu

09-13 Dec 2013 @ San Francisco, CA

Data assimilation

The goal of data assimilation is to estimate the (filtering or smoothing) posterior distribution of the system states given the available measurements (observations).

A Bayesian setup of the inference problem:

Prior measure defined by a Stochastic Differential Equation (Itō), [8]:

$$d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t)_\theta \; dt + \mathbf{\Sigma}^{1/2} \; d\mathbf{w}_t \; , \qquad d\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, dt\mathbf{I}) \; . \tag{1}$$

Observation model (**likelihood**) is corrupted by i.i.d. Gaussian white noise:

$$\mathbf{y}_k = h_k(\mathbf{x}_{t_k}) + \boldsymbol{\epsilon}_k \; , \qquad \boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k) \; . \tag{2}$$

Posterior measure is computed over paths (note the infinite dimensionality):

$$p_{post}(\{\mathbf{x}_t\}_{t\in T}|\mathbf{y}_{1:K}) = \frac{1}{Z} \times \left(\prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{x}_{t_k}) \right) \times p(\{\mathbf{x}_t\}_{t\in T}) \; . \tag{3}$$

A (new) variational approach to the estimation problem

The VGPA algorithm approximates the **true** posterior process by another that belongs to a family of **tractable** ones.

Gaussian process approximation (implies a linear SDE), [2]:

$$d\mathbf{x}_t = \mathbf{g}_L(\mathbf{x}_t)\, dt + \boldsymbol{\Sigma}^{1/2}\, d\mathbf{w}_t\ , \quad \text{with} \quad \mathbf{g}_L(\mathbf{x}_t) = -\mathbf{A}_t\mathbf{x}_t + \mathbf{b}_t\ , \tag{4}$$

with $\mathbf{A}_t \in \Re^{D\times D}$ and $\mathbf{b}_t \in \Re^D$ define the time varying linear drift in the approximating process.

The goodness of fit is measured with the Kullback-Liebler divergence (i.e. $KL[q(x)\|p(x|y)]$).

Variational free energy as an upper bound, [3, 4]:

$$\begin{aligned}\mathcal{F}(q(\mathbf{x}|\boldsymbol{\Sigma}), \boldsymbol{\theta}, \boldsymbol{\Sigma}) &= -\left\langle \ln \frac{p(\mathbf{x}, \mathbf{y}_{1:K}|\boldsymbol{\theta}, \boldsymbol{\Sigma})}{q(\mathbf{x}|\boldsymbol{\Sigma})} \right\rangle_{q(\mathbf{x}|\boldsymbol{\Sigma})} \\ &\geq -\ln p(\mathbf{y}_{1:K}|\boldsymbol{\theta}, \boldsymbol{\Sigma}) \end{aligned} \tag{5}$$

This is a 'by-product' of the variational framework and it comes for free!

VGPA - Core Equations

The Gaussian marginal at time 't' is defined as follows:

$$q(\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_t; \mathbf{m}_t, \mathbf{S}_t) \ , \quad t \in T \ , \tag{6}$$

where $\mathbf{m}_t \in \Re^D$ and $\mathbf{S}_t \in \Re^{D \times D}$, are respectively the marginal mean and covariance at time 't', given by:

Marginal moments (forward ODEs):

$$\frac{d\mathbf{m}_t}{dt} = -\mathbf{A}_t \mathbf{m}_t + \mathbf{b}_t \ , \tag{7}$$

$$\frac{d\mathbf{S}_t}{dt} = -\mathbf{A}_t \mathbf{S}_t - \mathbf{S}_t \mathbf{A_t}^\top + \mathbf{\Sigma} \ . \tag{8}$$

Solved numerically with explicit integration methods (e.g. Euler, Runge-Kutta).

These ODEs define the time evolution of the approximate process and act as **constraints** in the optimization process.

VGPA - Core Equations

With a bit more effort Eq.(5) derives to:

Cost functional, [3, 4]:

$$\mathcal{F}(q(\mathbf{x}),\boldsymbol{\theta},\boldsymbol{\Sigma}) = \mathrm{KL}[q_0\|p_0] + \frac{1}{2}\int_{t_0}^{t_f}\left\langle (\mathbf{f}(\mathbf{x}_t)-\mathbf{g}_L(\mathbf{x}_t))^\top \boldsymbol{\Sigma}^{-1}(\mathbf{f}(\mathbf{x}_t)-\mathbf{g}_L(\mathbf{x}_t))\right\rangle_{q_t} dt$$
$$+\frac{1}{2}\sum_k \left\langle (\mathbf{y}_k-\mathbf{x}_{t_k})^\top \mathbf{R}^{-1}(\mathbf{y}_k-\mathbf{x}_{t_k})\right\rangle_{q_t} + const(\mathbf{R}). \quad (9)$$

And with the addition of the constraints we get:

Constraint optimization (what we actually minimize!!):

$$\mathcal{L} = \mathcal{F}(q(\mathbf{x}|\boldsymbol{\Sigma}),\boldsymbol{\theta},\boldsymbol{\Sigma}) - \int_{t_0}^{t_f} \boldsymbol{\lambda}_t^\top \underbrace{(\dot{\mathbf{m}}_t + \mathbf{A}_t\mathbf{m}_t - \mathbf{b}_t)}_{\text{ODE for the means}} + \mathrm{tr}\{\boldsymbol{\Psi}_t \underbrace{(\dot{\mathbf{S}}_t + \mathbf{A}_t\mathbf{S}_t + \mathbf{S}_t\mathbf{A}_t{}^\top - \boldsymbol{\Sigma})}_{\text{ODE for the covariances}}\}\, dt \quad (10)$$

Therefore we have transformed an inference problem to an optimization problem.

VGPA - how to estimate model parameters

When we estimate model parameters we target at the following distribution $p(\theta|y)$. The variational free energy provides an upper bound on the marginal likelihood, $p(y|\theta)$, and we propose two options for parameter estimation [3, 5].

1 - Discrete approximations to the posterior distributions:

- select a set of points $\mathtt{D}_{\boldsymbol{\theta}} = \{\boldsymbol{\theta}_i\}_{i=1}^{n_\theta}$, to approximate the distribution
- run VGPA to convergence; this yields to a corresponding set of free energy values $\mathtt{D}_{\mathcal{F}} = \{\mathcal{F}(q(\mathbf{x}|\boldsymbol{\Sigma}), \boldsymbol{\theta}_i, \boldsymbol{\Sigma})\}_{i=1}^{n_\theta}$
- evaluate $\exp\{-\mathcal{F}(q(\mathbf{x}|\boldsymbol{\Sigma}), \boldsymbol{\theta}_i, \boldsymbol{\Sigma})\}$ and use this value as a proxy of the true marginal likelihood $p(\mathbf{y}_{1:K}|\boldsymbol{\theta}, \boldsymbol{\Sigma})$ (see Eq. 10)
- the approximate posterior distribution (including a prior $\pi(\boldsymbol{\theta}_i)$) is given:

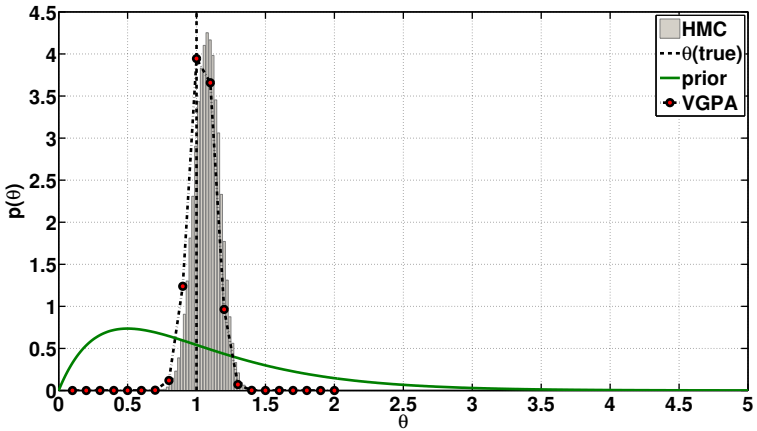
$$p(\boldsymbol{\theta}|\mathbf{y}_{1:K}) \propto \left\{ \exp\{-\mathcal{F}(q(\mathbf{x}|\boldsymbol{\Sigma}), \boldsymbol{\theta}_i, \boldsymbol{\Sigma})\} \times \pi(\boldsymbol{\theta}_i) \right\}_{i=1}^{n_\theta} \tag{11}$$

VGPA - DW example

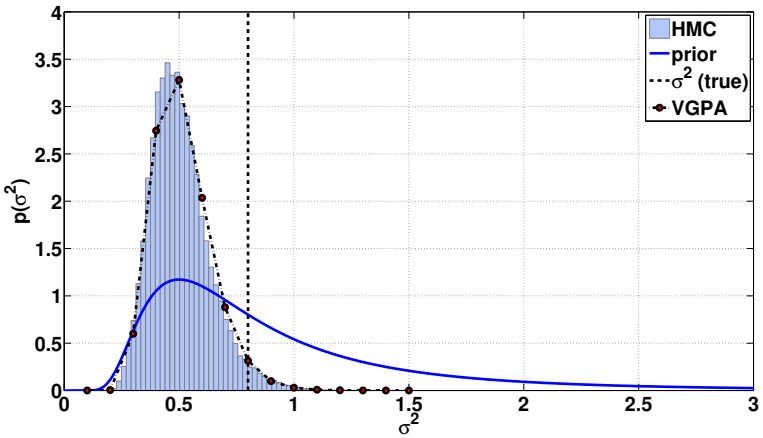
Double Well stochastic differential equation, $D = 1$:

$$dx_t = 4x_t(\theta - x_t^2)\ dt + \sigma dw_t\ ,\ \text{with}\ \theta\ , \sigma > 0\ , \tag{12}$$

driven by a double-well potential function $U(x_t) = -2x_t^2 + x_t^4$.



(a) Drift (θ) approximation, [5]



(b) Noise (σ^2) approximation, [5]

VGPA - how to estimate model parameters (continue).

2 - Maximum likelihood type-II (gradient based):

- compute the gradients of the cost function Eq.(10) w.r.t. the parameters of interest, example:

(a) $\nabla_{\theta}\mathcal{L} = \int_{t_0}^{t_f} \left\langle (\mathbf{f}(\mathbf{x}_t) - \mathbf{g}_L(\mathbf{x}_t))^{\top} \boldsymbol{\Sigma}^{-1} \nabla_{\theta} \mathbf{f}(\mathbf{x}_t) \right\rangle_{q_t} dt$

(b) $\nabla_{\boldsymbol{\Sigma}}\mathcal{L} = -\int_{t_0}^{t_f} \left(\frac{1}{2}\boldsymbol{\Sigma}^{-1} \left\langle (\mathbf{f}(\mathbf{x}_t) - \mathbf{g}_L(\mathbf{x}_t))(\mathbf{f}(\mathbf{x}_t) - \mathbf{g}_L(\mathbf{x}_t))^{\top} \right\rangle_{q_t} \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Psi}_t \right) dt$

- the procedure works in an inner/outer gradient optimization loop
- the parameters can be estimated either **jointly** or **marginally**

Other parameters of interest might be:

- initial posterior moments ($\mathbf{m}_0$ and $\mathbf{S}_0$)
- observation noise covariance $\mathbf{R}$

Mean field approximation (higher dimensions)

The variational framework [6], allows many approximations. To make VGPA applicable in higher dimensions we further adopt two more approaches.

1) Factorization of the posterior:

Assuming $\mathbf{A}_t$ is diagonal Eq.(7 & 8) can be solved for $\mathbf{A}_t$ and $\mathbf{b}_t$, allowing the linear drift to be re-written as: $\mathbf{g}_L(\mathbf{x}_t) = \dot{\mathbf{m}}_t - \frac{1}{2}(\boldsymbol{\Sigma} - \dot{\mathbf{S}}_t)\mathbf{S}_t^{-1}(\mathbf{x}_t - \mathbf{m}_t)$. This factorizes the posterior and allows for independent (and parallel) treatment for each dimension.

2) Polynomial approximation:

Parameterizing $\mathbf{m}_t$ and $\mathbf{S}_t$ with low order polynomials (between observations), allows for an analytic expression for the cost function Eq.(9), therefore no discretization errors!

The Lagrange polynomial formula for $m(t)$ and $s(t)$ is chosen as it provides additional advantages:

$$m_i^j(t) = \sum_{k=0}^{3} \left\{ m_i^j(t_j + kh) \left(\prod_{\substack{0 \le l \le 3 \\ l \neq k}} \frac{t - (t_j + lh)}{t_j - (t_j + lh)} \right) \right\} \tag{13}$$

A similar expression is derived for the marginal variance $s_i^j(t)$.

Application in Hydrology

Richards' PDE describes the movement of liguids in unsaturated porous media [7].

Pressure head $\psi(z,t)$-based formulation

$$C(\psi)\frac{\partial \psi}{\partial t} = \frac{\partial}{\partial z}\left[K(\psi)\left(\frac{\partial \psi}{\partial z} - 1\right)\right] + S(z,t) \tag{14}$$

Assuming a fixed uniform spatial grid $\Delta z = z_{i+1} - z_i$ and using a simple explicit discretization scheme we get a set of coupled ODEs:

$$C_i\frac{\Delta \psi_i}{\Delta t} = \frac{1}{\Delta z}\left[K^n_{i+1/2}\left(\frac{\psi^n_{i+1} - \psi^n_i}{\Delta z} - 1\right) - K^n_{i-1/2}\left(\frac{\psi^n_i - \psi^n_{i-1}}{\Delta z} - 1\right)\right] + S^n_i \ , \tag{15}$$

where $\psi_i \equiv \psi(z_i, t)$, with $i = 1 \ldots L$, representing the spatial index and n is the time index.

Complexity & Limitations

1. Complexity
   - computing the expectations and the integrals can be challenging, but once this is done no further tunning is required (unlike EnKF)
   - no need for computing an adjoint model (unlike 4DVar)
   - mean field approximation allows fully parallelizable cost function (fast)
   - polynomial approximation elliminates time discretization (memory efficient)
2. Limitations
   - GP approximation suffers when posterior is strongly non-Gaussian (but the same is true for other DA methods such as EnKF or 4DVar)
   - absence of lower bound means there are no quarantees on the approximation error

References

M. D. Vrettas, M. Opper and D. Cornford, An efficient variational mean filed algorithm for data assimilation. Quartelry Journal of the Royal Meteorological Society. In preparation.

C. Archambeau, D. Cornford, M. Opper and J. Shawe Taylor. Gaussian process approximation of stochastic differential equations. Journal of Machine Learning Research, Workshop and Conference Proceedings, 1:1-16, 2007.

C. Archambeau, M. Opper, Y. Shen, D. Cornford, J. Shawe-Taylor. Variational Inference for Diffusion Processes. In C. Platt, D. Koller, Y. Singer and S. Roweis editors, Neural Information Processing Systems (NIPS) 20, pages 17-24, 2008. The MIT Press.

M. D. Vrettas. Approximate Bayesian techniques for inference in stochastic dynamical systems. Ph.D. Thesis. Aston University, U.K. 2010.

M. D. Vrettas, D. Cornford and M. Opper, Estimating parameters in stochastic systems: A variational Bayesian approach. Physica D: Non-linear phenomena, vol. 240, pp.: 1877-1900, 2011.

T. Jaakkola. Tutorial on variational approximation methods. In M. Opper and D. Saad, editors, Advanced Mean Field Methods: Theory and Practice, The MIT press, 2001.

L. A. Richards, Capillary conduction of liquids through porous mediums. Physics, vol. 1, pp.: 318-333, 1931.

P. E. Kloeden and E. Platen. Numerical Solutions of Stochastic Differential Equations, Springer-Verlag, Berlin, 1992.

Acknowledgements

- This research was funded from an EPSRC grant: Variational Inference in Stochastic Dynamic Environmental Models (**VISDEM**) (EP/C005848/1) and EC-FP7 under the **QeoViQua** project (ENV.2010.4.1.2-2; 265178).
- Participation to the AGU conference is supported by the NSF, Math and Climate Research Network (DMS-0940272).