Part IV

Sampling problems: From molecular dynamics to Bayesian statistics and data assimilation

Outline

1. Molecular dynamics and sampling

2. Bayesian statistics and data assimilation

3. The link between molecular dynamics and Bayesian statistics

4. Further remarks: modeling errors, from smoothing to filtering.

References: Eugenia Kalnay "Atmospheric modeling, data assimilation and predictability", John M Lewis et al "Dynamic data assimilation: A least squares approach", Jari Kaipio & Erkki Somersalo "Statistical and computational inverse problems", MP Allen & DJ Tildesley "Computer Simulation of Liquids", B Leimkuhler & S Reich "Simulating Hamiltonian Dynamics"

1. Constant energy molecular dynamics

We consider a (classical) molecular system with m degrees of freedom (DOFs) described by generalized coordinates $\mathbf{q} \in \mathbb{R}^m$, potential energy function $V(\mathbf{q})$ and symmetric (possibly non-constant) mass matrix $\mathcal{M}(\mathbf{q}) \in \mathbb{R}^{m \times m}$.

The corresponding equations of motion can be derived from the Lagrangian functional

$$L[\mathbf{q}] = \int_{t_0}^{t_1} \mathcal{L}(\dot{\mathbf{q}}(t), \mathbf{q}(t))\, dt, \quad \mathcal{L}(\dot{\mathbf{q}}, \mathbf{q}) = \frac{1}{2}\dot{\mathbf{q}} \cdot [\mathcal{M}(\mathbf{q})\dot{\mathbf{q}}] - V(\mathbf{q})$$

and the associated Euler-Lagrange equations

$$\frac{d}{dt}\frac{\partial \mathcal{L}}{\partial \dot{\mathbf{q}}} - \frac{\partial \mathcal{L}}{\partial \mathbf{q}} = 0.$$

We now switch to the Hamiltonian side. We introduce the momentum conjugate to $\dot{\mathbf{q}}$:

$$\mathbf{p} = \frac{\partial \mathcal{L}}{\partial \dot{\mathbf{q}}} = \mathcal{M}(\mathbf{q})\dot{\mathbf{q}} \in \mathbb{R}^m,$$

the Hamiltonian (energy):

$$\mathcal{H}(\mathbf{q}, \mathbf{p}) = \frac{\partial \mathcal{L}}{\partial \dot{\mathbf{q}}} - \mathcal{L} = \frac{1}{2}\mathbf{p} \cdot \left[\mathcal{M}(\mathbf{q})^{-1}\mathbf{p}\right] + V(\mathbf{q}),$$

and the canonical equations of motion:

$$\dot{\mathbf{q}} = +\nabla_{\mathbf{p}}\mathcal{H}(\mathbf{q}, \mathbf{p}), \quad \dot{\mathbf{p}} = -\nabla_{\mathbf{q}}\mathcal{H}(\mathbf{q}, \mathbf{p}).$$

Let $\Gamma = (\mathbf{q}^T, \mathbf{p}^T)^T \in \mathbb{R}^{2m}$ denote the phase space variable. Then the canonical equations of motion can be written in the form

$$\dot{\Gamma} = f(\Gamma) = J^{-1}\nabla H(\Gamma), \qquad J^{-1} = \begin{bmatrix} 0_m & I_m \\ -I_m & 0_m \end{bmatrix}.$$

We introduce the τ-flow map

$$\Gamma(t+\tau) = g_\tau(\Gamma(t)), \qquad g_\tau : \mathbb{R}^{2m} \to \mathbb{R}^{2m},$$

which satisfies $g_0(\Gamma) = \Gamma$ and

$$\frac{d}{dt} g_\tau(\Gamma) = f(g_\tau(\Gamma)) \qquad \forall \Gamma \in \mathbb{R}^{2m}, \forall \tau \in \mathbb{R}.$$

The flow map conserves energy, i.e., $H(g_t(\Gamma)) = H(\Gamma)$, the symplectic two-form $\omega = \frac{1}{2} d\Gamma \wedge J d\Gamma$, i.e.

$$\left[\partial_\Gamma g_t(\Gamma)\right]^T J \left[\partial_\Gamma g_t(\Gamma)\right] = J,$$

volume in phase space, i.e. $\nabla \cdot f(\Gamma) = 0$ or $\det\left[\partial_\Gamma g_t\right](\Gamma) = 1$, respectively, and is time-reversible, i.e., $\mathcal{F} \circ g_t \circ \mathcal{F} = g_{-t}$, where $\mathcal{F}$ denotes the momentum flip $(\mathbf{q}, \mathbf{p}) \to (\mathbf{q}, -\mathbf{p})$.

2 Probability distributions, ergodicity, sampling

Let us assume that the initial state Γ_0 at $t = 0$ is not known precisely but that instead the initial value is treated as a random variable with given probability density function (PDF) $\rho_0(\Gamma)$.

Its time evolution is governed by the Liouville (transport) equation

$$\partial_t \rho + \nabla \cdot (\rho\, f) = 0$$

with initial condition $\rho(0, \Gamma) = \rho_0(\Gamma)$. Knowing the flow operator, the explicit solution is provided by

$$\rho(t, \Gamma) = \rho_0(g_{-t}(\Gamma)).$$

(This formula holds provided the vector field $f(\Gamma)$ is volume preserving, which holds since $f(\Gamma) = J^{-1}\nabla H(\Gamma)$).

An invariant PDF is a PDF which is stationary with respect to the Liouville equation, i.e.

$$\nabla \cdot (\rho\, f) = 0.$$

Assume that the energy is the only conserved quantity. A class of invariant PDFs is provided by

$$\rho(\Gamma) = \frac{1}{C}\xi(H(\Gamma)),$$

where ξ is an appropriate (generalized) function and C is normalization constant. Examples are provided by the micro-canonical (constant energy) PDF

$$\rho_E(\Gamma) = \frac{\delta(H(\Gamma) - E)}{\int \delta(H(\Gamma) - E)\, d\Gamma},$$

or the canonical (constant temperature) PDF

$$\rho_T(\Gamma) = \frac{\exp(-\beta H(\Gamma))}{\int \exp(-\beta H(\Gamma))\, d\Gamma}, \qquad \beta^{-1} = k_B T.$$

A constant energy hypersurface

$$\mathcal{M}_E = \{\Gamma : H(\Gamma) = E\}$$

is called ergodic if (i) all its invariant subsets $\mathcal{A} \subset \mathcal{M}_E$, i.e. $g_t(\mathcal{A}) = \mathcal{A}$, have either measure zero or one.

Ergodicity implies that

$$\bar{O}(\Gamma) := \lim_{T\to\infty} \int_0^T O(g_t(\Gamma))\, dt \;=\; \int O(\Gamma)\, \rho_E(\Gamma)\, d\Gamma := \langle O \rangle_E$$

for almost all Γ with $H(\Gamma) = E$ and all sufficiently smooth functions (observables) $O(\Gamma)$.

This is the famous time averages $\bar{O} = \bar{O}(\Gamma)$ equal to ensemble averages $\langle O \rangle_E$ relation for ergodic systems. It is nearly impossible to prove ergodicity except for highly idealized systems. But in practice the concept works remarkably well.

Molecular dynamics provides a tool to compute discrete trajectories $\{\Gamma_n\}_{n=0}^N$ from which we can compute time averages

$$\bar{O} \approx \lim_{N\to\infty} \sum_{n=0}^{N} O(\Gamma_n),$$

which, still assuming ergodicity, leads to an approximation to the ensemble average $\langle O \rangle_E$.

The most popular method for performing constant energy molecular dynamics (with constant mass matrices) is the Störmer-Verlet method

$$\begin{aligned}
\mathbf{p}^{n+1/2} &= \mathbf{p}^n - \frac{\Delta t}{2}\nabla V(\mathbf{q}^n), \\
\mathbf{q}^{n+1} &= \mathbf{q}^n + \Delta t M^{-1}\mathbf{p}^{n+1/2}, \\
\mathbf{p}^{n+1} &= \mathbf{p}^{n+1/2} - \frac{\Delta t}{2}\nabla V(\mathbf{q}^{n+1}).
\end{aligned}$$

Störmer-Verlet generates a numerical propagator

$$\Gamma^{n+1} = \Psi_{\triangle t}(\Gamma^n)$$

which is (i) symplectic, (ii) second order accurate, (iii) and time-reversible. Energy is not conserved, in general.

There is an ongoing theoretical debate whether or the not Störmer-Verlet method is capable of producing approximations to long time averages.

The main points in questions are (i) molecular dynamics conserves energy very well but not exactly and (ii) the global error between exact solutions and numerical approximations grows exponentially in time. We will come back to these issues in the following section.

In many applications one wishes to sample from the canonical PDF

$$\rho_T(\Gamma) = \frac{\exp(-\beta H(\Gamma))}{\int \exp(-\beta H(\Gamma))\, d\Gamma}, \qquad \beta^{-1} = k_B T,$$

i.e., one wishes to compute

$$\langle O \rangle_T = \int O(\Gamma)\, \rho_T(\Gamma)\, d\Gamma.$$

Again, instead of solving this problem by quadrature (curse of dimension), we generate time series and employ the ergodicity principle.

Appropriate time series can be generated by stochastic equations (Brownian and/or Langevin dynamics), Monte Carlo methods (see Part V) or by solutions to modified Hamiltonian equations of motion (e.g. Nosé-Hoover-Poincaré formulation).

3. Symplectic integration and backward error analysis

Given a differential equation

$$\dot{\mathbf{z}} = f(\mathbf{z}),$$

We consider one-step methods

$$\mathbf{z}^{n+1} = \Psi_h(\mathbf{z}^n)$$

of order $p \geq 1$ in the step-size $h = \Delta t$. Provided that $\Psi_{\Delta t}$ and f are real-analytic, one can find a modified differential equation

$$\dot{\mathbf{z}}_h = f_h(\mathbf{z}_h)$$

with solution operator $g_{t;h}$ such that

$$\|\Psi_h(\mathbf{z}) - g_{h;h}(\mathbf{z})\| \leq c_1 e^{-c_2/h}$$

and

$$f(\mathbf{z}) - f_h(\mathbf{z}) = \mathcal{O}(h^p).$$

We define the modified equations iteratively, i.e., for $i = 1, \ldots, I$,

$$f_h^{[i]}(\mathbf{z}) := f_h^{[i-1]}(\mathbf{z}) + h^i \Delta f^{[i]}(\mathbf{z})$$

with $f_h^{[0]} = 0$ and

$$\Delta f^{[i+1]}(\mathbf{z}) = \lim_{h \to 0} \frac{\Psi_h(\mathbf{z}) - g_{h,h}^{[i]}(\mathbf{z})}{h^{i+1}}.$$

Optimal truncation at $I = I(h)$ yields the exponential smallness result of the previous page. See Neishtadt (1984), Benettin & Giorgilli (1994), Hairer & Lubich (1997), Reich (1999) for details.

An important aspect of backward error/modified equation analysis is its qualitative aspect.

We formally consider the "Lie algebra" $\mathfrak{g}$ of vector fields f and the associated "Lie group" G of diffeomorphism. The following linear subsets of $\mathfrak{g}$ (and their associated subsets of G) are of particular interest in the context of this talk:

- Hamiltonian vector fields (symplectic diffeomorphism),
- homogeneous vector fields of degree 1, i.e. $\sigma f(\mathbf{z}) = f(\sigma \mathbf{z})$ (homogeneous maps of degree 1, i.e., $\sigma \Psi(\mathbf{z}) = \Psi(\sigma \mathbf{z})$),
- time-reversible vector fields (time-reversible diffeomorphism).

Theorem (Reich, 1993, 1999): If $f \in \mathfrak{g}^*$, $\mathfrak{g}^*$ a linear subset of $\mathfrak{g}$, $\Psi_h \in \mathrm{G}^*$ for all $h \geq 0$, G^* the appropriate subset of G, then

$$f_h \in \mathfrak{g}^*.$$

Example. We consider a homogeneous vector field f of degree one and an associated numerical method Ψ_h, which is of degree one as well (e.g., the forward Euler or any other RK method).

We prove the induction step. Let us assume that the i-th modified vector field $f_h^{[i]}$ is of degree one. It follows that

$$\begin{aligned}
\Delta f^{[i+1]}(\sigma \mathbf{z}) &= \lim_{h \to 0} \frac{\Psi_h(\sigma \mathbf{z}) - g_{h,h}^{[i]}(\sigma \mathbf{z})}{h^{i+1}} \\
&= \lim_{h \to 0} \frac{\sigma \Psi_h(\mathbf{z}) - \sigma g_{h,h}^{[i]}(\mathbf{z})}{h^{i+1}} \\
&= \sigma \, \Delta f^{[i+1]}(\mathbf{z})
\end{aligned}$$

and

$$f_h^{[i+1]}(\mathbf{z}) := f_h^{[i]}(\mathbf{z}) + h^{i+1} \Delta f^{[i+1]}(\mathbf{z})$$

is homogeneous of degree one.

4. Practical methods to compute modified energies

We now return to the canonical Hamiltonian equations, denote the state variable by $\Gamma = (\mathbf{q}^T, \mathbf{p}^T)^T$ and apply a symplectic and time-reversible method Ψ_h.

There exist modified energies $\mathcal{H}_h$ which are preserved to "arbitrary" high order $q > p$, p the order of the method, along the numerical solutions Γ^n, $n = 0, \ldots, N$, i.e.

$$\mathcal{H}_h(\Gamma^n) - \mathcal{H}_h(\Gamma^0) = \mathcal{O}(h^q).$$

There are several ways to construct the modified Hamiltonians. A standard finite difference analysis has, for example, been used by Hairer & Lubich and Akhmaskaya & Reich.

A particularly elegant practical construction of an $\mathcal{H}_h$ for splitting methods goes back to Skeel & Hardy (2001), which we put here into the framework of general symplectic methods.

Assume that a Hamiltonian $\mathcal{H}$ is homogeneous of degree 2, i.e., $\sigma^2\mathcal{H}(\Gamma) = \mathcal{H}(\sigma\Gamma)$, then the Hamiltonian vector field $f = J^{-1}\nabla\mathcal{H}$ is homogeneous of degree 1 and

$$\mathcal{H}(\Gamma) = \frac{1}{2}\Gamma \cdot \left[J\dot{\Gamma}\right].$$

Why: Differentiate $\sigma^2\mathcal{H}(\Gamma) = \mathcal{H}(\sigma\Gamma)$ with respect to σ to obtain

$$2\sigma\mathcal{H}(\Gamma) = \Gamma \cdot \nabla\mathcal{H}(\sigma\Gamma)$$

and set $\sigma = 1$. Finally use $\nabla\mathcal{H}(\Gamma) = J\dot{\Gamma}$.

Assume that the method Ψ_h is symplectic and of degree 1. Then the modified Hamiltonians $\mathcal{H}_h$ are of degree 2 and, hence,

$$\mathcal{H}_h(\Gamma_h) = \frac{1}{2}\Gamma_h \cdot \left[J\dot{\Gamma}_h\right].$$

Finally we compute $\mathcal{H}_h(\widehat{\Gamma}_h)$ by interpolating $\Gamma_h(t)$ along the numerical trajectory, i.e. $\widehat{\Gamma}_h(t_n) = \Gamma^n$, to sufficiently high order.

Practical construction outline (Skeel & Hardy, 2001):

- Define the homogeneous extension of a Hamiltonian $\mathcal{H}$ by

$$H(\Gamma, \alpha, \beta) = \alpha^2 \mathcal{H}(\alpha^{-1}\Gamma),$$

  $\alpha, \beta \in \mathbb{R}$, and the extended structure matrix

$$\mathsf{J} = \begin{bmatrix} J & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}.$$

- Apply the numerical method (symplectic and homogeneous of degree 1) to the extended equations of motion with initial conditions $\alpha^0 = 1$ and $\beta^0 = 0$. Then $\alpha^n \equiv 1$ and

$$\mathcal{H}_h(\Gamma_h) = \frac{1}{2}\Gamma_h \cdot \left[J\dot{\Gamma}_h\right] - \dot{\beta}_h. \tag{1}$$

- Find approximations $\widehat{\Gamma}_h$ and $\widehat{\beta}_h$ from its known values, i.e., $\widehat{\Gamma}_h(t_n) = \Gamma^n$, $\widehat{\beta}_h(t_n) = \beta^n$ by appropriate interpolation.

5 Backward error analysis, shadowing, and long time averages

The only systems for which ergodicity (and in fact mixing) can be proved rigorously are uniformly hyperbolic systems. Such systems are characterized by the fact that, at each point $\Gamma \in \mathcal{M}_E$ on an constant energy surface, the associated tangent plane $T_\Gamma \mathcal{M}_E$ can be split into stable and unstable subspaces. Such systems are also structural stable under small perturbations. Furthermore, exact trajectories can be shadowed by pseudo-trajectories.

These results can be combined with modified equation analysis for symplectic methods to demonstrate that numerical trajectories shadow an exact trajectory over exponentially long periods of time (Reich, 1999).

6 Data assimilation

Let us assume that we need to estimate the initial conditions of a differential equation

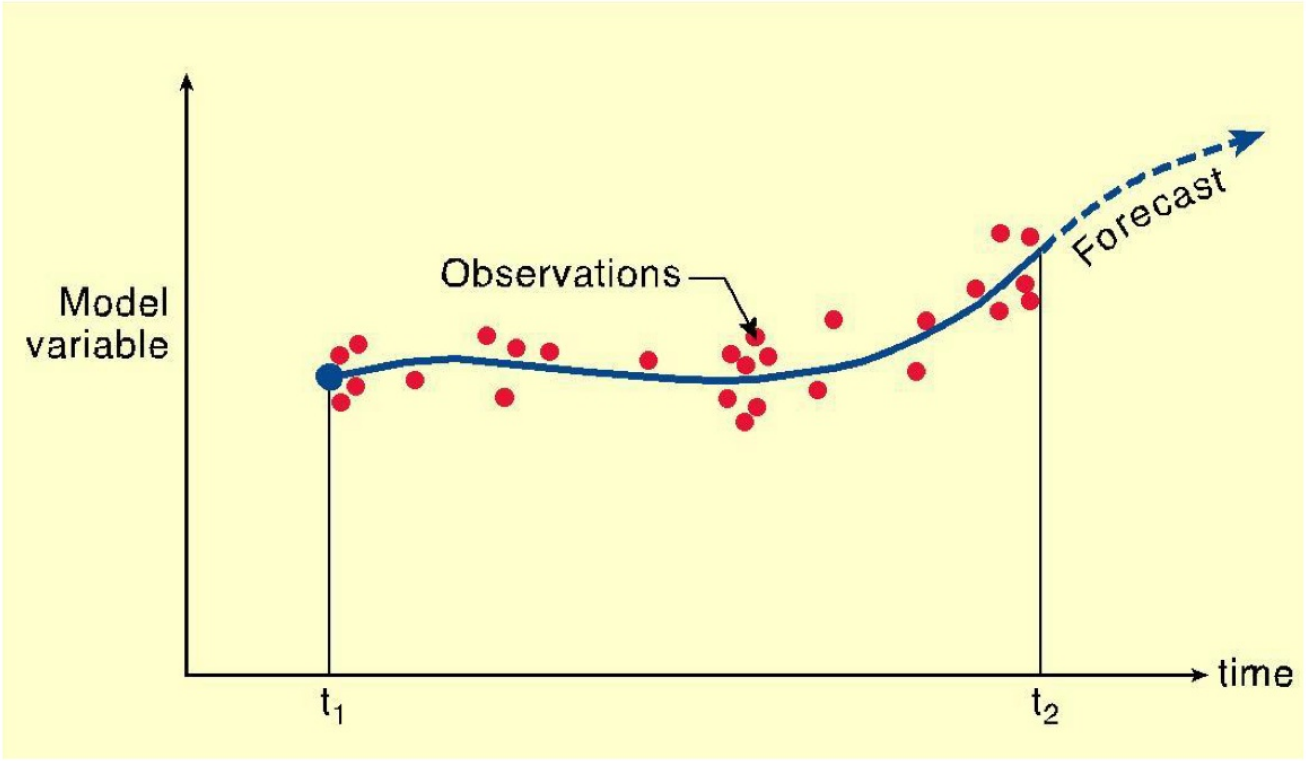
$$\dot{\mathbf{z}} = f(\mathbf{z})$$

from a given initial PDF $\rho_0(\mathbf{z})$ and available (partial) observations

$$\mathbf{y}(t_i) = h(\mathbf{z}(t_i)) + \eta_i, \qquad \mathbf{y} \in \mathbb{R}^k,$$

at times $t_i \geq 0$, where $\eta_i \sim [N(0, \sigma^2)]^k$ are i.i.d. normally distributed with mean zero and variance σ^2.

As a result we wish to find an optimal estimate for the initial $\mathbf{z}(0)$ based on the prior ρ_0 and the observations $\mathbf{y}_i$. This problem is called data assimilation.


Model variable
Observations
Forecast
time
t_1
t_2

7 Bayesian estimation method

Let $\mathbf{x} \in \mathbb{R}^n$ be the unknown to be estimated and $\mathbf{y} \in \mathbb{R}^l$ be the observations that contain information about the unknown $\mathbf{x}$ to be estimated.

The distinguished feature of the Bayesian framework is that it treats the unknown $\mathbf{x}$ as a random variable. It is assumed that a prior distribution $\rho(\mathbf{x})$ is known. The prior summarizes our initial knowledge/belief about the unknown. We are furthermore allowed to observe $\mathbf{y}$ whose conditional PDF $\rho(\mathbf{y}|\mathbf{x})$ is known.

From this information we wish to estimate the optimal value $\widehat{\mathbf{x}}$ for the true $\mathbf{x}$ with respect to some cost function $c : \mathbb{R}^n \to \mathbb{R}$.

Let $\rho(\mathbf{x}, \mathbf{y})$ denote the joint PDF of $\mathbf{x}$ and $\mathbf{y}$. Since

$$\rho(\mathbf{x}, \mathbf{y}) = \rho(\mathbf{y}|\mathbf{x})\,\rho(\mathbf{x}) = \rho(\mathbf{x}|\mathbf{y})\,\rho(\mathbf{y}),$$

we obtain the well-known Bayes' formula

$$\rho(\mathbf{x}|\mathbf{y}) = \frac{\rho(\mathbf{y}|\mathbf{x})\,\rho(\mathbf{x})}{\rho(\mathbf{y})}$$

where

$$\rho(\mathbf{y}) = \int \rho(\mathbf{x}, \mathbf{y})\,d\mathbf{x}$$

is the marginal distribution of $\mathbf{y}$. Note that $\rho(\mathbf{y})$ is just a normalization constant in Bayes' formula.

The conditional distribution $\rho(\mathbf{x}|\mathbf{y})$ is known as the posterior distribution of $\mathbf{x}$ given the observations $\mathbf{y}$.

We can now obtain an optimal estimate by minimizing Bayes' cost function

$$B(\hat{\mathbf{x}}|\mathbf{y}) = \mathbb{E}[c(\hat{\mathbf{x}})|\mathbf{y}] = \int c(\mathbf{x} - \hat{\mathbf{x}})\, \rho(\mathbf{x}|\mathbf{y})\, d\mathbf{x}.$$

If we use the uniform cost function

$$c(\mathbf{x}) = \begin{cases} 0, & \text{if } \|\mathbf{x}\| \leq \epsilon, \\ 1, & \text{otherwise} \end{cases},$$

for given $\epsilon > 0$. Then the associated Bayes' cost function $B(\hat{\mathbf{x}}|\mathbf{y})$ is minimized if and only if the posterior distribution $\rho(\mathbf{x}|\mathbf{y})$ is maximized, i.e.

$$0 = \nabla_{\mathbf{x}} \rho(\hat{\mathbf{x}}_U|\mathbf{y})$$

and $\hat{\mathbf{x}}_U$ is called the maximum a posteriori estimate (MAP).

If we use alternatively the cost function

$$c(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2,$$

then the optimal estimator is provided by the expectation value

$$\hat{\mathbf{x}}_{MS} = \mathbb{E}[\mathbf{x}|\mathbf{y}] = \int \mathbf{x}\, \rho(\mathbf{x}|\mathbf{y})\, d\mathbf{x}.$$

This is called Bayes' least squares estimator.

The estimator is unbiased ($\mathbb{E}[\mathbf{x} - \mathbf{x}_{MS}|\mathbf{y}] = 0$) and is also a minimum (error) variance estimate.

However, $\mathbf{x}_{MS}$ is more difficult to compute in general than $\mathbf{x}_U$. We will discuss an efficient method in Part V.

8 4D (variational) data assimilation

We us now return to differential equations

$$\dot{\mathbf{z}} = f(\mathbf{z}), \qquad \mathbf{z} \in \mathbb{R}^n.$$

Let us assume that the prior PDF for the initial state is given by a Gaussian distribution of the form

$$\rho_0(\mathbf{z}) = \frac{1}{C_0} \exp\left(\frac{1}{2}(\mathbf{z}-\mathbf{z}_b)^T B^{-1}(\mathbf{z}-\mathbf{z}_b)\right)$$

with given mean $\mathbf{z}_b$ and covariance matrix B. In general good approximations to $\mathbf{z}_b$ and B are difficult to obtain and cannot be obtained within the 4D Var framework. We will come back to this issue when discussing ensemble Kalman filtering.

Next we assume that we have a perfect forecasting method (i.e., we assume that a given differential equation represents reality exactly and that we know its flow map g_t).

Then we assume that the conditional probability of observing $\mathbf{y}_i$ at time t_i given an initial state $\mathbf{z}$ is also Gaussian, i.e.

$$\rho_i(\mathbf{y}_i|\mathbf{z}) = \frac{1}{C_i} \exp\left(\frac{1}{2}(\mathbf{y}_i - h(g_{t_i}(\mathbf{z})))^T R_i^{-1}(\mathbf{y}_i - h(g_{t_i}(\mathbf{z})))\right)$$

where R_i is the error covariance matrix and

$$\bar{y}_i = h(g_t(\mathbf{x}))$$

is the expected value of the observation for a given initial state.

Note that the flow operator g_t leads to a nonlinear dependence of ρ_i on $\mathbf{z}$.

Let us assume that we got N such observations at times $t_1, \ldots, t_N$, which we group together in a big vector $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_N)^T$, then

$$\rho(\mathbf{y}|\mathbf{z}) = \prod_{i=1}^{N} \rho_i(\mathbf{y}_i|\mathbf{z}).$$

Bayes' formula then implies

$$\rho(\mathbf{z}|\mathbf{y}) = \frac{1}{C}\exp(-J(\mathbf{z}))$$

with the functional ("cost function") J given by

$$\begin{aligned} J(\mathbf{z}) &= -\frac{1}{2}(\mathbf{z}-\mathbf{z}_b)^T B^{-1}(\mathbf{z}-\mathbf{z}_b) \\ &\quad -\sum_{i=1}^{N} \frac{1}{2}(\mathbf{y}_i - h(g_{t_i}(\mathbf{z})))^T R_i^{-1}(\mathbf{y}_i - h(g_{t_i}(\mathbf{z}))). \end{aligned}$$

Note that the observed values $\mathbf{y}_i$ are known (have been observed) at this stage!

We can now compute the MAP estimator for the most likely initial state by minimizing J, i.e.

$$\nabla J(\widehat{\mathbf{z}}_U) = 0.$$

This requires an efficient way of computing the gradient $\nabla J(\mathbf{z})$. We will discuss this on the next few slides.

Alternatively, Bayes' least square estimator is obtained by computing the expectation value

$$\mathbf{z}_{MS} = \mathbb{E}[\mathbf{z}|\mathbf{y}] = \int \mathbf{z}\, \rho(\mathbf{z}|\mathbf{y})\, d\mathbf{z}.$$

This task can be tackles using Monte Carlo methods. Efficient Monte Carlo methods will also require knowledge of the gradient $\nabla J(\mathbf{z})$!

9 The adjoint formulation

Here we describe a recursive procedure to compute the gradient $\nabla J(\mathbf{z})$.

We first make the following observation with respect to the Jacobian $G(t_i, t_{i+1}) \in \mathbb{R}^{n\times n}$ of the flow operator g_t along the solution $\mathbf{z}(t) = g_t(\mathbf{z})$. The Jacobian is the solution to the variation equation

$$\dot{\mathbf{Z}} = Df(\mathbf{z}(t))\,\mathbf{Z}$$

with initial condition $\mathbf{Z}(t_i) = I \in \mathbb{R}^{n\times n}$, i.e., $G(t_i, t_{i+1}) = \mathbf{Z}(t_{i+1})$. We have the composition rule

$$G(t_0, t_i) = \prod_{j=i}^{1} G(t_{j-1}, t_j) = G(t_{i-1}, t_i)\, G(t_{i-2}, t_{i-1}) \cdots G(t_0, t_1).$$

Given the standard inner product $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^T \mathbf{b}$ in $\mathbb{R}^n$, the gradient is defined by

$$\frac{J(\mathbf{z} + \varepsilon\, \delta\mathbf{z}) - J(\mathbf{z})}{\varepsilon} = \langle \delta\mathbf{z}, \nabla J(\mathbf{z}) \rangle$$

for all $\delta\mathbf{z}$ as $\varepsilon \to 0$.

After a bit of algebra, this results in

$$\begin{aligned} \nabla J(\mathbf{z}) &= -B^{-1}(\mathbf{z} - \mathbf{z}_b) + \\ &\quad + \sum_{i=1}^{N} G(t_0, t_i)^T Dh(\mathbf{z}(t_i))^T \left[R^{-1}(\mathbf{y}_i - h(\mathbf{z}(t_i))) \right] \\ &= -B^{-1}(\mathbf{z} - \mathbf{z}_b) + \sum_{i=1}^{N} G(t_0, t_i)^T \mathbf{d}_i \end{aligned}$$

with

$$\mathbf{d}_i = Dh(\mathbf{z}(t_i))^T \left[R^{-1}(\mathbf{y}_i - h(\mathbf{z}(t_i))) \right].$$

Note that

$$G(t_0, t_i)^T = \prod_{j=1}^{i} G(t_{j-1}, t_j)^T = G(t_0, t_1)^T \, G(t_1, t_2)^T \cdots G(t_{i-1}, t_i)^T.$$

The operators $G(t_{i-1}, t_i)^T$ are the solutions to the adjoint variational equation

$$\dot{\mathbf{M}} = Df(\mathbf{z}(t))^T \, \mathbf{M}$$

with "initial" condition $\mathbf{M}(t_i) = I$ and

$$G(t_{i-1}, t_i)^T = -\mathbf{M}(t_{i-1}).$$

i.e., we solve the adjoint equation backwards in time!

Hence we obtain a recursive formulation for the gradient of the form ($i = 2$):

$$G(t_0, t_2)\mathbf{d}_2 + G(t_0, t_1)\mathbf{d}_1 = G(t_0, t_1)^T \left[\mathbf{d}_1 + G(t_1, t_2)^T \mathbf{d}_2\right].$$

Let us assume for a moment that we approximate the differential equation by the explicit Euler method. Then the numerical $G(t_{i-1}, t_i)$ is given by

$$G(t_{i-1}, t_i) = I + \Delta t \, Df(\mathbf{z}_{i-1}), \qquad \Delta t = t_i - t_{i-1},$$

from this we obtain

$$G(t_{i-1}, t_i)^T = I + \Delta t \, Df(\mathbf{z}_{i-1})^T,$$

which is obtained by solving the adjoint variational equation

$$\dot{\mathbf{M}} = -Df(\mathbf{z}_{i-1}) \, \mathbf{M}$$

by the implicit Euler method backwards in time from t_i to t_{i-1} with "initial" condition $\mathbf{M}_i = I$, i.e.

$$\frac{I - \mathbf{M}_{i-1}}{t_i - t_{i-1}} = -Df(\mathbf{z}_{i-1})I.$$

Given a Hamiltonian system

$$\dot{\mathbf{q}} = M^{-1}\mathbf{p}, \qquad \dot{\mathbf{p}} = -\nabla V(\mathbf{q}),$$

the variational equation becomes

$$\dot{\mathbf{Q}} = M^{-1}\mathbf{P}, \qquad \dot{\mathbf{P}} = -A(t)\,\mathbf{Q}$$

with

$$A(t) = V_{\mathbf{qq}}(\mathbf{q}(t)).$$

The adjoint variational equation is then given by

$$\dot{\mathbf{Q}} = A(t)\,\mathbf{P}, \qquad \dot{\mathbf{P}} = -M^{-1}\mathbf{Q}.$$

The same analysis can be performed directly on the discretization level (e.g. Störmer-Verlet method).

8 The link between molecular dynamics and Bayesian statistics

In molecular dynamics one is often interested in expectation values $\mathbb{E}[f(\mathbf{q})]$ with respect to the canonical PDF

$$\rho_T(\mathbf{q}) = \frac{1}{C}\exp(-\beta V\mathbf{q}), \qquad \beta^{-1} = k_B T,$$

in the atomistic positions.

A well known technique is provided by stochastic Brownian dynamics

$$d\mathbf{q} = -\nabla V(\mathbf{q})\,dt + \sqrt{2\beta^{-1}}d\mathbf{W}(t)$$

where $\mathbf{W}(t)$ is a m-dimensional Wiener process. It can be shown that, under appropriate conditions on the potential V, the canonical PDF $\rho_T(\mathbf{q})$ is stationary and that

$$\mathbb{E}[f(\mathbf{q})] = \lim_{T\to\infty}\frac{1}{T}\int_0^T f(\mathbf{q}(t))\,dt.$$

In some cases one might want to use the second-order Langevin equation instead:

$$d\mathbf{q} = M^{-1}\mathbf{p}\,dt, \qquad d\mathbf{p} = -\nabla V(\mathbf{q})\,dt - \gamma\mathbf{p}\,dt + \sqrt{2\beta^{-1}\gamma M}\,d\mathbf{W}(t),$$

where γ is a friction parameter. In this case the canonical density

$$\rho_T(\Gamma) = \frac{1}{C}\exp(-\beta H(\Gamma))$$

associated with the complete Hamiltonian is stationary under appropriate conditions on H.

In molecular dynamics, the second order formulation is often preferred since it is "closer" to the constant energy formulation. For $\gamma \to \infty$, Langevin dynamics reduces to first order Brownian dynamics.

Brownian and Langevin dynamics can also be used in the context of data assimilation where we only need to replace the atomistic positions $\mathbf{q}$ by the initial state $\mathbf{z}$, the potential energy function $V(\mathbf{q})$ by $J(\mathbf{z})$, and set $\beta = 1$.

In case of second-order Langevin dynamics we introduce artificial momenta $\pi \in \mathbb{R}^n$ and an Hamiltonian (energy)

$$H(\Gamma) = \frac{1}{2}\pi^T M^{-1}\pi + J(\mathbf{z})$$

for an appropriate "mass" matrix M and $\Gamma = (\mathbf{z}^T, \pi^T)^T$.

In either case Bayes' least square estimator is approximated by

$$\mathbf{z}_{MS} = \lim_{T\to\infty} \frac{1}{T}\int_0^T \mathbf{z}(t)\,dt$$

along solutions $\mathbf{z}(t)$ of the Langevin equations.

10 From smoothing to filtering

The 4D Var off-line analysis can be turned into an on-line technique (ensemble Kalman filter).

Given any method which samples from the Bayesian distribution

$$\rho(\mathbf{z}|\mathbf{y}) = \frac{1}{C}\exp(-J(\mathbf{z})),$$

we can also sample from the implied PDF

$$\rho_N(\mathbf{z}|\mathbf{y}) = \rho(g_{-t_N}(\mathbf{z})|\mathbf{y})$$

for the solutions $\mathbf{z} = \mathbf{z}(t_N)$ at time t_N (assuming for simplicity that g_t is volume conserving). Furthermore, we can search for the best fit of $\rho_N(\mathbf{z}|\mathbf{y})$ to a normal distribution, i.e., we need to estimate the expectation value $\mathbf{z}_b = \mathbb{E}[\mathbf{z}|\mathbf{y}]$ and the covariance matrix B. This information can be used as a prior PDF for a new data assimilation window.

11 Model errors

So far we have assumed that the map g_t is a true representation of the dynamics. However, in general, we have to take modeling errors and errors due to numerical approximation into account.

Furthermore, the observations $\mathbf{y}_i$ obtain information that is not representable by the model and/or the numerics.

These aspects should be built into the data assimilation process. Some progress has been made under the assumption that the errors can be treated as random variables as well.