Data Assimilation

Henry D. I. Abarbanel
Department of Physics,
and
Marine Physical Laboratory,
Scripps Institution of Oceanography,
University of California San Diego
9500 Gilman Drive
La Jolla, CA 92093

habarbanel@ucsd.edu

Contents

1 Introduction

A core issue in Physics in this century is the study of networks of nonlinear systems or high-dimensional nonlinear systems. Networks may consist of connected collections of low dimensional nonlinear dynamical systems each of whose properties we have learned to analyze [Aba96, KS04]. That analysis tells us how to evaluate some invariants associated with the attractor on which orbits of the dynamical system move after transients of the motion become unimportant. These include a collection of fractal dimensions characteristic of the density on the attractor, $\rho(\mathbf{r})$:

$$\rho(\mathbf{r}) = \frac{1}{N}\sum_{n=1}^{N} \delta^D(\mathbf{r} - \mathbf{x}(t_n)) \tag{1}$$

where $\mathbf{r}$ is a location in the D-dimensional space in which solutions $\mathbf{x}(t)$ of the system differential equation

$$\frac{dx_a(t)}{dt} = F_a(\mathbf{x}(t));\ a = 1, 2, ..., D \tag{2}$$

reside. The density follows the orbit $\mathbf{x}(t)$ starting at t_1 in discrete time, and $t_n = t_1 + (n-1)\Delta t$.

Also included among these invariants are the D global Lyapunov exponents [Aba96, KS04].

Characterizing properties of the attractor of a nonlinear dynamical system from time series observations of a measured orbit while the system is driven by some given forces does little to answer the question of paramount importance in determining the equations of motion of the system producing that orbit. For that we need to use observations of a system such as Eq. (2) and transfer the information in those observations to establish various aspects of the "vector field" $\mathbf{F}(\mathbf{x})$.

The notion of **data assimilation** arose in the long-standing problem of predicting the weather using as a model of the ocean and atmosphere the fluid dynamical equations of motion. These equations are partial differential equations for quantities such as the temperature $T(\mathbf{r}, t)$ at all spatial points in the atmosphere and ocean $\mathbf{r} = (x, y, z)$ as a function of time. The equations do not, in general, have closed form, analytic solutions with the complicated boundary conditions imposed by the actual Earth system. Numerical solutions require a discretization of both space and time.

Further, and more important here than the accuracy of the numerical solutions, is the fact that the equations of motion exhibit chaotic trajectories [Lor63] leading to the amplification of small perturbations, for example, from numerical round-off error. This leads to uncertainties in solutions to the dynamical equations from even very precisely specified initial conditions. To address this, the idea was that if we could measure the dynamical variables, such as $T(\mathbf{r}, t)$, we would know where the real Earth system actually lies at the measurement time, and we could use these measurements to "guide" the numerics toward, even to, the correct orbits of the system.

So one must collect the required measurements, and these are always noisy. Further, the model dynamics are infinite dimensional as the state variables are fields, so we must settle for collecting noisy observations of a subset of the state variables at locations in space $\mathbf{r}$ where instruments have been placed. As we want to **predict** the weather, we must, at the end t_{final} of an observation window $[t_0 \leq t \leq t_{final}]$ in time, provide accurate estimates of the value of all state variables, at all grid points in space used for integrating the model, including those state variables which were **not** measured during the observation window. Further, yet, we must estimate all time independent parameters in the model in order to integrate the dynamical equations for $t > t_F$ and predict the weather.

How big are these numerical weather models, in terms of the number of ordinary differential equations at each selected spatial grid point? In 2019, as this is written, the most detailed models have approximately $1.5X10^{10}$ and have available about 1% of these as noisy measured items. So one must estimate about 99% of the state variables as well as numerous parameters in the model representing properties of the earth, for example, soil properties, and transport of momentum (viscosity) and energy (heat conductivity), ...; most of which we do not know very well.

When all this is addressed, we must still answer the questions of how many state variables must be measured at each observation time in order to contain enough information about the state of the model system to stabilize the transfer of information, and we must address how to represent errors in the models we have selected.

That is a lot of material to get right just to predict tomorrow's weather, or more or less any predict the future of any complex nonlinear system. Not discouraged by this we have put together a systematic way to proceed in this data assimilation effort, identifying some of the actual challenges one must address.

2 Formulation of Statistical DA

First, a bit of notation. We eschew partial differential equations for the physical and biological problems we address, so we begin with a state space of D-dimensional variables we call $x_a(t);\ a = 1, 2, ..., D$. These are taken to satisfy D (nonlinear) ordinary differential equations in continuous time

$$\frac{dx_a(t)}{dt} = F_a(\mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\theta}), \tag{3}$$

or discrete time versions of this telling how we move from the state at time t to the state at time $t + \Delta t$:

$$x_a(t + \Delta t) = f_a(\mathbf{x}(t), \mathbf{u}(t), \boldsymbol{\theta}). \tag{4}$$

$\mathbf{f}(\mathbf{x})$ is your selection of what method is employed for integrating Eq. (3) [PTVF07], and that dictates how the vector field $\mathbf{F}(\mathbf{x})$ is represented by $\mathbf{f}(\mathbf{x})$. In each formulation the quantities $\mathbf{u}(t)$ are a set of forces dictated from outside the collection of state variables $\mathbf{x}(t)$. These forces do not obey differential equations within this framework; they make the system go. The N_p quantities $\boldsymbol{\theta}$ are time independent parameters used within the model dynamics.

We will observe L quantities at each time τ_k; $k = 1, 2, ..., F$ measurements are made. If the discrete time increment is Δt, then we choose $\tau_k = n_k \Delta t$ for convenience; n_k is an integer. This is not necessary, but it prevents severe notation clutter. Let's collect all samples of the model state up to $t_k = t_0 + k\Delta t$ into the **path** $\mathbf{X}(k) = \{x(t_0), x(t_1), ..., x(t_k)\} = \{\mathbf{x}(0), \mathbf{x}(1), ..., \mathbf{x}(k)\}$ of the system while it moves in the observation window, and all measured quantities into the collection $\mathbf{Y}(k) = \{\mathbf{y}(t_0), \mathbf{y}(t_1), ..., \mathbf{y}(t_k)\} = \{\mathbf{y}(0), \mathbf{y}(1), ..., \mathbf{y}(k)\}$. If no measurement is made at some time τ_k, the vector $\mathbf{y}(k)$ is not in $\mathbf{Y}$.

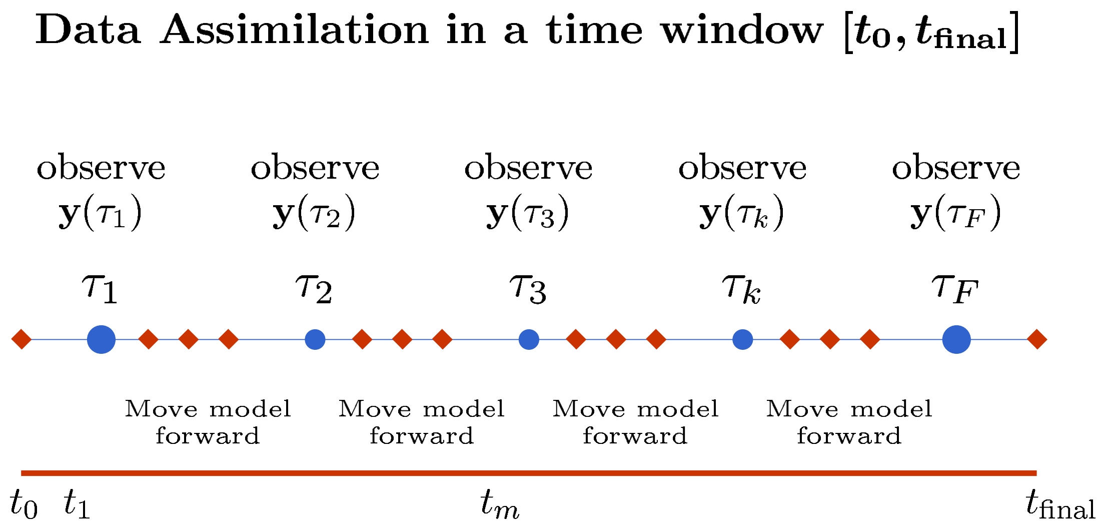
Data Assimilation in a time window $[t_0, t_{\text{final}}]$
observe $\mathbf{y}(\tau_1)$
observe $\mathbf{y}(\tau_2)$
observe $\mathbf{y}(\tau_3)$
observe $\mathbf{y}(\tau_k)$
observe $\mathbf{y}(\tau_F)$
τ_1
τ_2
τ_3
τ_k
τ_F
Move model forward
Move model forward
Move model forward
Move model forward
t_0
t_1
t_m
t_{final}

Figure 1: Timeline of observations $\mathbf{y}(\tau_k); \tau_k = t_0 + n_k \Delta t$ and progress of the model $\mathbf{x}(t + \Delta t) = \mathbf{f}(\mathbf{x}(t))$ through the observation window $[t_0 \leq t \leq t_{final}]$. The observations are taken, for convenience only, at some multiples n_k of the model sampling times Δt. If observations were made, for example, only at the two times $t = t_0 + 5\Delta t$ and $t = t_0 + 173\Delta t$, then only $\mathbf{y}(\tau_5)$ and $\mathbf{y}(\tau_{173})$ are nonzero; all other $\mathbf{y}(\tau)$ are absent.

Inevitably, the data we collect are noisy. Equally, the model we select to describe the production of those data has errors. This means we must, at the outset, address a conditional probability distribution $\pi(\mathbf{X}|\mathbf{Y})$ as our goal in the data assimilation transfer from $\mathbf{Y}$ to the model. In [Aba13] we describe how to use the Markov nature of the model dynamics $\mathbf{x}(n) \rightarrow \mathbf{x}(n + 1) = \mathbf{f}(\mathbf{x}(n), \boldsymbol{\theta})$ and the definition of conditional probabilities to derive the **recursion relation** connecting observations and dynamics at times t_{n+1}

and t_n:

$$
\begin{aligned}
&\pi(\mathbf{X(n+1)}|\mathbf{Y(n+1)}) = \frac{\pi(\mathbf{y}(n+1), \mathbf{x}(n+1), \mathbf{X}(n), \mathbf{Y}(n))}{\pi(\mathbf{y}(n+1), \mathbf{Y}(n))\, \pi(\mathbf{x}(n+1), \mathbf{X}(n), \mathbf{Y}(n))} \cdot \\
&\frac{\pi(\mathbf{x}(n+1), \mathbf{X}(n), \mathbf{Y}(n))}{\pi(\mathbf{X}(n), \mathbf{Y}(n))} \cdot \pi(\mathbf{X}(n), \mathbf{Y}(n)) \\
&= \frac{\pi(\mathbf{y}(n+1), \mathbf{x}(n+1), \mathbf{X}(n)|\mathbf{Y}(n))}{\pi(\mathbf{y}(n+1)|\mathbf{Y}(n))\, \pi(\mathbf{x}(n+1), \mathbf{X}(n+1)|\mathbf{Y}(n))} \cdot \\
&\pi(x(n+1)|\mathbf{X}(n), \mathbf{Y}(n)) \cdot \pi(\mathbf{X}(n)|\mathbf{Y}(n)) \\
= \quad & \exp[\mathrm{CMI}(\mathbf{y}(n+1), \mathbf{x}(n+1), \mathbf{X}(n)|\mathbf{Y}(n))] \cdot \pi(\mathbf{x}(n+1)|\mathbf{x}(n)) \cdot \pi(\mathbf{X}(n)|\mathbf{Y}(n)) \\
= \quad & \frac{\pi(\mathbf{y}(n+1)|\mathbf{x}(n+1), \mathbf{X}(n), \mathbf{Y}(n))}{\pi(\mathbf{y}(n+1)|\mathbf{Y}(n))} \cdot \pi(\mathbf{x}(n+1)|\mathbf{x}(n)) \cdot \pi(\mathbf{X}(n)|\mathbf{Y}(n)). \qquad (5)
\end{aligned}
$$

We have identified $\mathrm{CMI}(a, b|c) = \log[\frac{\pi(a,b|c)}{\pi(a|c)\,\pi(b|c)}]$. This is Shannon's conditional mutual information [Fan61] telling us how many bits (for $\log_2$) we know about a when observing b conditioned on c. For us $a = \{\mathbf{y}(n+1)\}, b = \{\mathbf{x}(n+1), \mathbf{X}(n)\}, c = \{\mathbf{Y}(n)\}$. We also used the Markov property of the noisy dynamics:
$\pi(\mathbf{x}(n+1)|\mathbf{X}(n), \mathbf{Y}(n)) = \pi(\mathbf{x}(n+1)|\mathbf{x}(n))$.

Using this recursion relation to move backwards from the end of the observation window from $t_{final} = t_0 + N\Delta t$ through the measurements at times τ_k to the start of the window at t_0, we may write, up to factors independent of $\mathbf{X}$

$$
\pi(\mathbf{X}|\mathbf{Y}) = \left\{ \prod_{k=0}^{F} \pi(\mathbf{y}(\tau_k)|\mathbf{X}(\tau_k), \mathbf{Y}(k-1)) \prod_{n=0}^{final-1} \pi(\mathbf{x}(n+1)|\mathbf{x}(n)) \right\} \pi(\mathbf{x}(0)). \qquad (6)
$$

If we now choose $\pi(\mathbf{X}|\mathbf{Y}) \propto \exp[-A(\mathbf{X})]$. $A(\mathbf{X})$ is the negative of the log likelihood; we call this the *action*.

Since the dimension of $\mathbf{X}$ is $D(N+1)$, and both the dimension of the model and the length of the time series may be large, we cannot expect to visualize the full conditional probability distribution $\pi(\mathbf{X}|\mathbf{Y})$. Instead we make (hopefully) educated guesses about quantities that characterize the Physics of the dynamical motion of the model. Call these $G(\mathbf{X})$, and concentrate attention on the expected values

$$
E[G(\mathbf{X})|\mathbf{Y}] = \langle G(\mathbf{X}) \rangle = \int d\mathbf{X}\, G(\mathbf{X}) \pi(\mathbf{X}|\mathbf{Y}) = \frac{\int d\mathbf{X}\, G(\mathbf{X}) \exp -[A(\mathbf{X})]}{\int d\mathbf{X} \exp -[A(\mathbf{X})]}. \qquad (7)
$$

Many discussions of SDA or ML discuss $\pi(\mathbf{X}|\mathbf{Y})$ itself, but it is hard to imagine knowing quite what to do with a function of such a large number of variables. Often, in discussions of statistics, one recognizes that much of the information one actually works with is contained in moments of this distribution

What $G(\mathbf{X})$ are of interest?

Certainly we would like to know the expected value of the path itself: $G(\mathbf{X}) = \mathbf{X}$. Within the observation window $[t_0, t_F]$, we make measurements and use methods discussed below to transfer information from these observations to

1. the **observed** state variables $\mathbf{x}_{\text{observed}}(t)$;
2. the **unobserved** state variables $\mathbf{x}_{\text{unobserved}}(t)$; and
3. the unknown parameters of the model dynamics.

This usually goes by the name “fit the model to the data”, though examining the unobserved variables are generally not considered as there is nothing to compare them with. However, if one is testing one’s methods for the transfer of information from the data to the model using data one has generated oneself—we call these exercises ‘twin experiments’–this can be quite instructive.

When time a has reached $t = t_{final}$, namely the end of observations, we have an estimate of the full state of the system $\mathbf{x}(t_{final})$ and the parameters $\boldsymbol{\theta}$. From this estimate and knowledge of the forces $\mathbf{u}(t \geq t_{final})$, we now may *predict* the behavior of the dynamical system beyond the observation window. In the prediction window $t \geq t_{final}$ no further information from observations is used.

Predicting is critical. As we only observe a subset of the model state variables, a good ‘fit’ within the observation window only permits us to conclude that the model may be consistent with the data. Success in predicting gives us validation (or not) of the properties of the model. Good predictions indicate both consistency of the data with the model we selected **and** a good estimate of the model state $\mathbf{x}(t_{final})$ and the model parameters at the termination of observations.

In machine learning ‘prediction’ is called ‘generalization’.

When we have $\langle \mathbf{X} \rangle$, we might want to know about the errors is the estimation of the model state. To this end we note that $\langle \mathbf{X} \rangle$ is a vector in $D(N+1) + N_p$ space. We give the components of this vector an index μ and write $\langle \mathbf{X}_\mu \rangle = \bar{\mathbf{X}}_\mu$. Our next choice for G($\mathbf{X}$) would then be

$$G_{\mu,\nu}(\mathbf{X}) = (\mathbf{X}_\mu - \bar{\mathbf{X}}_\mu)(\mathbf{X}_\nu - \bar{\mathbf{X}}_\nu), \tag{8}$$

and this is the sample covariance matrix of the model path.

3 Evaluating the Expected Value Integral

3.1 Standard Model for SDA

There is a standard set of assumptions about the ingredients entering the action. The three elements of the action are

- the conditional mutual information (CMI) representing the transfer from data $(\mathbf{y}(\tau_k)$ when observations are made: $\pi(\mathbf{y}(n+1)|\mathbf{x}(n+1), \mathbf{X}(n), \mathbf{Y}(n)$,
- the movement of the state variables from time t_n to time t_{n+1} represented in $\pi(\mathbf{x}(n+1)|\mathbf{x}(n))$, and
- the distribution of initial conditions $\pi(\mathbf{x}(0))$.

3.1.1 Measurement Error Term in the Action

The CMI term $\pi(\mathbf{y}(n+1)|\mathbf{x}(n+1), \mathbf{X}(n), \mathbf{Y}(n)$ tells us that the observation at time t_{n+1} $\mathbf{y}(n+1)$ can depend on the sequence of states up to t_{n+1} and the observations up to t_n. The usual interpretation of this is to consider the components of the L-dimensional observation to be a function of $\mathbf{x}(n+1)$ plus some noise. So one would write

$$y_l(n) = h_l(\mathbf{x}(n)) + \eta_l(n);\ l = 1, 2, ..., L, \tag{9}$$

and where $\mathbf{h}(\mathbf{x})$ describes the instrument(s) used in making observations. $\eta_l(n)$ is a random variable representing the noise in the measurements. This can be generalized in various ways, say by making the noise term dependent on the state $\mathbf{x}(n)$. Mostly it is simplified by taking $\mathbf{h}(\mathbf{x}) = \mathbf{x}$ and choosing $\eta_l(n)$ to be Gaussian. So $y_l(n) = x_l(n)+\eta_l(n)$ or

$$\pi(\mathbf{y}(n)|\mathbf{x}(n)) = \text{constant} \exp[-R_m/2 \sum_{l=1}^{L} (x_l(n) - y_l(n))^2], \tag{10}$$

where R_m can be a matrix and is known as the precision matrix of the Gaussian. The contribution to the action from measurement errors is then

$$\text{measurement error} = -\frac{R_m}{2} \sum_{k=1}^{F} \sum_{l=1}^{L} (x_l(\tau_k) - y_l(\tau_k))^2, \tag{11}$$

and constants cancel in the expected value integral.

One place where it is useful to note the dependence of this term on measurements at earlier times, as allowed in the CMI formula, is when one wishes to use time delays to account for the waveform of the measurements to augment the information available to be transferred to the model [REK$^+$14, REM$^+$14, PCL16, PvLP18]

3.1.2 Model Error Term in the Action

The second term in the action moves the model state variables at time t_n to the state variables at time t_{n+1}. If the model were perfect, this term would be a delta function:

$$\pi(\mathbf{x}(n+1)|\mathbf{x}(n)) = \delta^D(\mathbf{x}(n+1) - \mathbf{f}(x(n));\ ;\text{perfect model}. \tag{12}$$

We must replace the delta function with something broader in $\mathbf{x}(n+1) - \mathbf{f}(\mathbf{x}(n)$, and we choose another Gaussian to represent the model error

$$\pi(\mathbf{x}(n+1)|\mathbf{x}(n)) \propto \exp\left[-\frac{R_f}{2} \sum_{a=1}^{D} (x_a(t_{n+1}) - f_a(t_n))^2\right]. \tag{13}$$

3.1.3 $\pi(\mathbf{x}(t_0))$

Statisticians call this the *prior*, and Physicists recognize this as the initial condition that the probability density for the model state variables must have as the time dependent probability density satisfies a first order partial differential equation in time, and we must give some statement of this to proceed.

If we know nothing or little about this, we can take it to be a uniform distribution in the model state variables and model parameters, representing our lack of knowledge. As we move through the observation window $[t_0, t_{final}]$ information flows from the measurements $\mathbf{y}(\tau_k)$ and in nonlinear systems the initial conditions are 'forgotten' and we can use the assumption of 'lack of knowledge' about the state of the system at t_0.

If we have some knowledge of $\mathbf{x}(t_0)$ either from a good guess or a previous set of observations assimilated into the model, we should use it. Often a Gaussian distribution is assumed so we could guess

$$\pi(\mathbf{x}(t_0)) \propto \exp\left[-\frac{B}{2}\sum_{a=1}^{D}(x_a(t_0) - \bar{x}_a)^2\right]. \tag{14}$$

In a sense this is not a resolution of how to address model error, but a way of ducking the issue. In practice the decision lies in the hands of the user.

Finally we assemble these pieces to arrive at an expression for the "standard model" for SDA:

$$\begin{aligned} A(\mathbf{X}) = {} & \frac{R_m}{2}\sum_{k=1}^{F}\sum_{l=1}^{L}(x_l(\tau_k) - y_l(\tau_k))^2 \\ & \frac{R_f}{2}\sum_{n=0}^{N}\sum_{a=1}^{D}(x_a(t_{n+1}) - f_a(t_n))^2 \\ & - \log[\pi(\mathbf{x}(t_0))]. \end{aligned} \tag{15}$$

It is important to note that even though we have taken the measurement and model errors to be Gaussian, the presence of the nonlinear vector field $\mathbf{f}(\mathbf{x}(n))$ in the discrete time dynamics makes the overall action non-Gaussian.

3.2 Variational Principles–Laplace's Method (1774)

If we could locate the path $\mathbf{X}^0$ which yields the smallest value of the action $A(\mathbf{X}^0)$, that would give the largest value for $\pi(\mathbf{X}|\mathbf{Y}))$, we suspect that would dominate the expected value integrals Eq. (7). To find the minima of $A(\mathbf{X})$ we must locate its extrema $\mathbf{X}^q; q = 0, 1, ...$

$$\left.\frac{\partial A(\mathbf{X})}{\partial \mathbf{X}_\nu}\right|_{\mathbf{X}^q} = 0, \tag{16}$$

where the matrix

$$\left.\frac{\partial^2 A(\mathbf{X})}{\partial \mathbf{X}_\nu \, \partial \mathbf{X}_\mu}\right|_{\mathbf{X}^q} \tag{17}$$

is positive definite.

This notion is due to Laplace [Lap74, Lap86]. As the action is nonlinear in the state variables and parameters of the model (the path variables) there may be many minima contributing to expected values.

The importance of the extremum can be seen if we expand the action about $\mathbf{X}^0$, a minimum,

$$\begin{aligned} A(\mathbf{X}) &= A(\mathbf{X}^0) + \frac{\partial A(\mathbf{X})}{\partial \mathbf{X}_\nu}\bigg|_{\mathbf{X}^0} (\mathbf{X} - \mathbf{X}^0)_\nu \\ &+ \frac{1}{2}\frac{\partial^2 A(\mathbf{X})}{\partial \mathbf{X}_\nu \, \partial \mathbf{X}_\mu}\bigg|_{\mathbf{X}^0} (\mathbf{X} - \mathbf{X}^0)_\nu (\mathbf{X} - \mathbf{X}^0)_\mu + \cdots, \\ &= A(\mathbf{X}^0) + \frac{1}{2} H_{\mu\nu}(\mathbf{X}^0)(\mathbf{X} - \mathbf{X}^0)_\nu (\mathbf{X} - \mathbf{X}^0)_\mu + \cdots, \end{aligned} \tag{18}$$

so if the first derivative of the action is zero and, for the moment we drop the higher derivatives, the expected value of $G(\mathbf{X})$ is approximately

$$\begin{aligned} &\langle G(\mathbf{X}) \rangle = \\ &\frac{\exp[-A(\mathbf{X}^0)] \int d\mathbf{X} \, \exp[-\tfrac{1}{2} H_{\mu\nu}(\mathbf{X}^0)(\mathbf{X} - \mathbf{X}^0_\nu)(\mathbf{X} - \mathbf{X}^0)_\mu] \, G(\mathbf{X}^0)}{\exp[-A(\mathbf{X}^0)] \int d\mathbf{X} \, \exp[-\tfrac{1}{2} H_{\mu\nu}(\mathbf{X} - \mathbf{X}^0)_\nu (\mathbf{X} - \mathbf{X}^0)_\mu]} \\ &= G(\mathbf{X}^0) \end{aligned} \tag{19}$$

Retaining additional terms in the Taylor expansion of $A(\mathbf{X})$ improves the accuracy of Laplace's method [ZJ02], it is known as perturbation theory in statistical physics.

Here there is another point to be made. If there is another minimum at associated with the path $\mathbf{X}^1$, then another term enters with action value $A(\mathbf{X}^1)$. Gathering this term with the first, the leading term in $\langle G(\mathbf{X}) \rangle$ reads

$$\langle G(\mathbf{X}) \rangle = G(\mathbf{X}^0) + G(\mathbf{X}^1) \exp[-(A(\mathbf{X}^1) - A(\mathbf{X}^0)] \left(\frac{\det(H(A(\mathbf{X}^0))}{\det(H(A(\mathbf{X}^1))} \right)^{\mathcal{P}/2}, \tag{20}$$

where $\mathcal{P}$ is the dimension of the integral over $\mathbf{X}$. This shows that if $A(\mathbf{X}^1) > A(\mathbf{X}^0)$, the second term is exponentially smaller than the first. The minimum we seek in enforcing Eq. (16) is the one with the smallest value of the action, namely the global minimum.

Methods for finding that **global** minimum are discussed elsewhere [KRYA17].

3.3 Monte Carlo Methods

Monte Carlo methods [MRR$^+$53, Has70] were invented precisely to approximate integrals of the form Eq. (7). They employ a search method in path space $\mathbf{X}$ that samples $\pi(\mathbf{X}|\mathbf{Y})$ in one manner or another.

All start at some selected path $\mathbf{X}_0$ and make a 'proposal' choosing the next path $\mathbf{X}_1$ in a chain of accepted paths $\{\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \ldots\}$. Starting at $\mathbf{X}_0$, the proposal to move to $\mathbf{X}_1$ is evaluated by an acceptance criterion tuned to $\pi(\mathbf{X}|\mathbf{Y})$. If the proposal is accepted, the MC procedure moves to $\mathbf{X}_1$ and a second proposal is made to move to $\mathbf{X}_2$. If the proposal is not accepted, the MC procedure remains at $\mathbf{X}_0$ and makes another proposal for $\mathbf{X}_1$. This is now evaluated by the acceptance criterion, and accepted or not. By performing this "make proposal–evaluate by acceptance criterion" many times, a collection of accepted proposals is used to approximate the integral Eq. (7):

$$\langle G(\mathbf{X})\rangle \approx \frac{1}{\text{number of accepted paths}} \sum_{j=1}^{\text{number of accepted paths}} G(\mathbf{X}^j). \tag{21}$$

The error is approximately (number of accepted paths)$^{-0.5}$. A useful, entertaining discussion is found in [PTVF07].

As one can imagine over the many decades since [MRR$^+$53] this very straightforward procedure has been explored in some depth. This document is introductory and there would not be room for all the relevant references to MC procedures. So we will briefly describe two MC variants and let the reader explore from there.

3.3.1 Random Proposals; Metropolis Hastings

The Metropolis-Hastings (MH) procedure produces a Markov process defined by the transition probability $P(\mathbf{X}'|\mathbf{X})$ in path space which reaches a stationary distribution $\pi(\mathbf{X})$ after a large (in principle, infinite) number of repeated steps. A sufficient, but not necessary, condition is called *detailed balance* which states that the transition $\mathbf{X} \rightarrow \mathbf{X}'$ is reversible. This requires $\pi(\mathbf{X})P(\mathbf{X}'|\mathbf{X}) = \pi(\mathbf{X}')P(\mathbf{X}|\mathbf{X}')$.

Now start somewhere in path space $\mathbf{X}_0$ and through a rule, coming next, produce a sequence $\{\mathbf{X}_0, \mathbf{X}_1, ..., \mathbf{X}_N\}$. called a chain. The rule is this

- start at $\mathbf{X}_0$;
- select a candidate location in path space $\mathbf{X}_c$ and accept this location or reject it;
- **if you accept it**, the next location along the chain, $\mathbf{X}_1 = \mathbf{X}_c$;
- **if you reject it**, the next location along the chain, $\mathbf{X}_1 = \mathbf{X}_0$;
- do these steps starting at $\mathbf{X}_1$ and continuing to $\mathbf{X}_J$. Then stop.

If we select the candidate location $\mathbf{X}_c$, when we are located at $\mathbf{X}_j$, according to the transition probability $P(\mathbf{X}_c|\mathbf{X}_j)$, the *acceptance rule* is to accept the candidate move from location $\mathbf{X}_j$ to location $\mathbf{X}_c$ at the rate

$$a(\mathbf{X}_j, \mathbf{X}_c) = \text{minimum}\left(1, \frac{\pi(\mathbf{X}_c)P(\mathbf{X}_j|\mathbf{X}_c)}{\pi(\mathbf{X}_c|\mathbf{X}_j)}\right). \tag{22}$$

This assures us that as the number of accepted new locations $J \rightarrow \infty$

$$\frac{1}{J}\sum_{j=1}^{J} G(\mathbf{X}_j) = E_{\pi(\mathbf{X})}[G(\mathbf{X})] = \int d\mathbf{X}\, G(\mathbf{X})\pi(\mathbf{X}), \tag{23}$$

regardless where one starts ($\mathbf{X}_0$).

Basically all Monte Carlo methods using MC chains differ by the target or desired density in path space and the rule for selecting candidate locations $P(\mathbf{X}_c|\mathbf{X}_j)$. The original [MRR+53, Has70] method chooses $\mathbf{X}_c$ using a draw from $P(\mathbf{X}_c|\mathbf{X}_j)$. We call this the random proposals (RP) MC procedure.

Wherever you start in path space, $\mathbf{X}_0$, you sample the whole distribution you desire $\pi(\mathbf{X})$ and perform the integral

$$\langle G(\mathbf{X})\rangle = \int d\mathbf{X}\, G(\mathbf{X})\pi(\mathbf{X}), \tag{24}$$

and this nice result is approached at the rate [PTVF07]

$$\int d\mathbf{X}\, G(\mathbf{X})\pi(\mathbf{X}) \approx \langle G(\mathbf{X})\rangle \pm \sqrt{\frac{\langle G^2(\mathbf{X})\rangle - \langle G(\mathbf{X})\rangle^2}{J}}, \tag{25}$$

as $J \to \infty$.

As one can never take $J = \infty$ steps from $\mathbf{X}_0$, the literature is built on methods to reduce the numerator and methods to speed up the random walks underlying making proposals for candidate locations. Two things often happen: (a) the acceptance rate is low, so it takes a long time to collect enough samples for the sum to approximate the integral, and/or (b) the size of an acceptable step from $\mathbf{X}_j \to \mathbf{X}_c$ is so small, many, many proposed candidate locations are required.

3.3.2 Hamiltonian Monte Carlo

Hybrid Monte Carlo or Hamiltonian Monte Carlo (HMC is shorthand for either) methods takes another path [DKPR87] from [MRR+53, Has70]. This procedure introduces a "canonically conjugate" path $\mathbf{P}$ changing the action by an additive function of $\mathbf{P}$, so

$$A(\mathbf{X}) \to A(\mathbf{X}) + h(\mathbf{P});\ \text{or} \pi(\mathbf{X}) \to \exp -[h(\mathbf{P}) + A(\mathbf{X})], \tag{26}$$

which leaves the expected value integral unchanged

$$\begin{aligned}\langle G(\mathbf{X})\rangle &= \frac{\int d\mathbf{X}\, G(\mathbf{X}) \exp[-A(\mathbf{X})]}{\int d\mathbf{X}\, \exp[-A(\mathbf{X})]} \\ \langle G(\mathbf{X})\rangle &= \frac{\int d\mathbf{X}d\mathbf{P}\, G(\mathbf{X}) \exp[-A(\mathbf{X}) + h(\mathbf{P})]}{\int d\mathbf{X}d\mathbf{P}\, \exp[-A(\mathbf{X}) + h(\mathbf{P})]}.\end{aligned} \tag{27}$$

Candidate proposals are made in $\{\mathbf{X}, \mathbf{P}\}$ space by a classical mechanics "canonical" transformation [GPS02] $\{\mathbf{X}_j, \mathbf{P}_j\} \to \{\mathbf{X}_c, \mathbf{P}_c\}$ which preserves the sum $H(\mathbf{X}, \mathbf{P}) = A(\mathbf{X}) + h(\mathbf{P})$ as well as the phase space volume $d\mathbf{X}_j\, d\mathbf{P}_j = \partial\mathbf{X}_c\, d\mathbf{P}_c$. A number of other quantities are also conserved as canonical transformations respect the underlying symmetry of symplectic transformations in $\{\mathbf{X}, \mathbf{P}\}$ phase space [GPS02].

Canonical transformations can be accomplished in many ways. If they carry a label "s" which we may think of as indicating a number of steps to be taken to go from $\{\mathbf{X}_j, \mathbf{P}_j\}$ to $\{\mathbf{X}_c, \mathbf{P}_c\}$, or one may call it a kind of 'time'. If the label is continuous, both $H(\mathbf{X}, \mathbf{P})$ and phase space volume are precisely conserved. However, as we are required to work in discrete "s", we cannot conserve **both** of these precisely [ZM88].

If both were conserved precisely, the HMC acceptance rate would be unity. Furthermore, a proposal in phase space $\{\mathbf{X}_j, \mathbf{P}_j\} \to \{\mathbf{X}_c, \mathbf{P}_c\}$, may include a 'large' jump in path space. This results in jumps in $\mathbf{P}$ space which leave expected values of $G(\mathbf{X})$ unchanged regardless of the choice of $h(\mathbf{P})$.

There are many ways to construct a canonical transformation [CB09, GPS02]. One is to integrate Hamilton's equations for the elements of phase space $\{\mathbf{X}(s), \mathbf{P}(s)\}$ labeled by 's':

$$\frac{d\mathbf{X}(s)}{ds} = \frac{\partial H(\mathbf{X}(s), \mathbf{P}(s))}{\partial \mathbf{P}(s)}; \quad \frac{d\mathbf{P}(s)}{ds} = -\frac{\partial H(\mathbf{X}(s), \mathbf{P}(s))}{\partial \mathbf{X}(s)}, \tag{28}$$

starting at $s = 0$ where $\{\mathbf{X}(0), \mathbf{P}(0)\} = \{\mathbf{X}_j, \mathbf{P}_j\}$ to $\{\mathbf{X}(s), \mathbf{P}(s)\}$. Symplectic integrators which preserve phase space volume precisely and accurately, but not precisely, preserve $H(\mathbf{X}, \mathbf{P})$ are available [LR04, HLW06].

4 Closing Comments

There are three critical elements in statistical data assimilation (SDA)

- well curated data–namely know the instruments that collect features of observed systems of interest to you; noise in the data collection should definitely be on your mind;
- a model that you must provide that you argue acts dynamically to produce the observations you have collected;
- a tested method that transfers information contained in your data to your model.

After making the third step, you must use your well informed model, now having estimates of the full state of the model at the termination of observations $\mathbf{x}(t_{final})$ as well as estimates of all unknown parameters in the model, to make predictions of the response of your model for $t > t_{final}$ to new stimuli.

The next essay will discuss how this works in practice in establishing biophysical models for neurons using data collected in neurobiological experiments.

References

[Aba96] H. D. I. Abarbanel. *The Analysis of Observed Chaotic Data*. Springer-Verlag, New York, 1996.

[Aba13] Henry D. I. Abarbanel. *Predicting the Future: Completing Models of Observed Complex Systems*. Springer, 2013.

[CB09] John R. Cary and Alain J. Brizard. Hamiltonian theory of guiding-center motion. *Reviews of Modern Physics*, 81:693–738, 2009.

[DKPR87] Simon Duane, A. D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics Letter B*, 195(2):216–222, 1987.

[Fan61] Robert M. Fano. *Transmission of Information; A Statistical Theory of Communication*. MIT Press, 1961.

[GPS02] Herbert Goldstein, Charles P. Poole, and John L. Safko. *Classical Mechanics, 3rd Edition*. Pearson, 2002.

[Has70] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[HLW06] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer Heidelberg, Berlin, 2006.

[KRYA17] Nirag Kadakia, Daniel Rey, Jingxin Ye, and Henry D. I. Abarbanel. Symplectic structure of statistical variational data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 143(703):756–771, 2017.

[KS04] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis, 2nd ed.* Cambridge University Press, Cambridge, UK, 2004.

[Lap74] P. S. Laplace. Memoir on the probability of causes of events. *Mathématique et de Physique,Tome Sixiéme*, pages 621–656, 1774.

[Lap86] P.S. Laplace. Memoir of the probability of causes of events. *Statistical Science*, 1:365–378, 1986. Translation to English by S. M. Stigler.

[Lor63] Edward N Lorenz. Deterministic nonperiodic flow. *Journal of the atmospheric sciences*, 20(2):130–141, 1963.

[LR04] B. Leimkuhler and S. Reich. *Simulating Hamiltonian Dynamics*. Cambridge University Press, 2004.

[MRR$^+$53] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, 21:1087–1092, June 1953.

[PCL16] D. Pazo, A. Carrassi, and J. M. Lopez. Data assimilation by delay-coordinate nudging. *Quarterly Journal of the Royal Meteorological Society*, 142, 2016.

[PTVF07] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical recipes: the art of scientific computing.* Cambridge University Press, 2007.

[PvLP18] Flavia R. Pinheiro, Peter Jan van Leeuwen, and Ulrich Parlitz. An ensemble framework for time delay synchronization. *Quarterly Journal of the Royal Meteorological Society*, 142, 2018.

[REK$^+$14] Daniel Rey, Michael Eldridge, Mark Kostuk, Henry DI Abarbanel, Jan Schumann-Bischoff, and Ulrich Parlitz. Accurate state and parameter estimation in nonlinear systems with sparse observations. *Physics Letters A*, 378(11):869–873, 2014.

[REM$^+$14] Daniel Rey, M. Eldridge, U. Morone, H. D. I. Abarbanel, U. Parlitz, and J. Schumann-Bischoff. Using waveform information in nonlinear data assimilation. *Physical Review E*, 90:062916, 2014.

[ZJ02] Jean Zinn-Justin. *Quantum field theory and critical phenomena.* Oxford University Press, 2002.

[ZM88] Ge Zhong and Jerrold E. Marsden. Lie-poisson hamilton-jacobi theory and lie-poisson integrators. *Physics Letters A*, 133(3):134 – 139, 1988.