Gaussian Process Approximations of Stochastic Differential Equations

Cédric Archambeau* Dan Cawford† Manfred Opper‡
John Shawe-Taylor⋆

May, 2006

1 Introduction

Some of the most complex models routinely run are numerical weather prediction models. These models are based on a discretisation of a coupled set of partial differential equations (the dynamics) which govern the time evolution of the atmosphere, described in terms of temperature, pressure, velocity, etc, together with parameterisations of physical processes such as radiation and clouds [3]. These dynamical models typically have state vectors with dimension $O(10^6)$ or more. A key issue in numerical weather prediction is the inference of the state vector, given a set of observations, which is referred to as data assimilation. Most modern data assimilation methods can be seen in a sequential Bayesian setting as the estimation of the posterior distribution of the state, given a prior distribution of the state that is derived from the numerical weather prediction model, and a likelihood that relates the observations to the state [4].

In general the prior distribution is approximated as a space only Gaussian process, the observations are often given Gaussian likelihoods, and thus inference of the posterior proceeds using well known Gaussian process methods [6]. It is possible to view almost all data assimilation methods as approximations to the Kalman filter / smoother. Much early work in data assimilation focussed on the static case where the prior distribution was assumed to have a climatological covariance structure whose form was dictated

*University of Southampton
†Aston University
‡Technical University Berlin

in many cases from the model dynamics, and a mean given by a deterministic forecast from the previous data assimilation cycles posterior mean [3]. At each time step, the covariance of the state was thus ignored and only the mean was propagated forward in time. Recently much work has been done to address the issue of the propagation of the uncertainty at initial time through the non-linear model equations. The most popular method is called the ensemble Kalman filter [5], and is a Monte Carlo approach. The basic idea is very simple:

1. take a sample from your prior distribution (in practice a sample size O(100) or less is used, even for high dimensional systems);
2. propagate each sample, integrating the model equations deterministically to produce a forecast ensemble;
3. use the forecast ensemble to estimate the forecast mean and covariance (avoid rank deficiency problems by localising this, which also minimises the effect of sampling noise);
4. update each ensemble member carefully using a Kalman filter derived update, so they are a sample from the corresponding posterior distribution and iterate to 2.

The ensemble approach has many advantages, but it is suboptimal from many aspects, for example the sampling noise can be large and uncontrolled, it is not trivial to incorporate non-Gaussian likelihoods, it requires several model integrations and it is impractical to implement as a smoother. See Appendix A for more information about the Kalman filter link.

In the following we present some initial results from the VISDEM project, where we seek a variational Bayesian treatment of the dynamic data assimilation problem which builds upon our variational Bayesian Gaussian process treatment of the static data assimilation problem [1]. In particular we focus on the issue of defining a Gaussian process approximation to the temporal evolution of the solution of a general stochastic differential equation with additive noise.

2 Modelling stochastic differential equations

We assume that the variables divide into two groups

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}^{\|} \\ \mathbf{x}^{\perp} \end{pmatrix}$$

such that the noise level is $\sigma^{\|}$ on the $\mathbf{x}^{\|}$ variables assumed to be k of the n, while for the remainder it is $\sigma^{\perp}$.

Assuming the noise to be equal variance and uncorrelated we have

$$\begin{aligned}\Delta\mathbf{x}_k^{\|} &= \mathbf{x}_{k+1}^{\|} - \mathbf{x}_k^{\|} = \mathbf{f}^{\|}(\mathbf{x}_k)\Delta t + \sigma^{\|}\mathbf{z}^{\|}\sqrt{\Delta t},\\ \Delta\mathbf{x}_k^{\perp} &= \mathbf{x}_{k+1}^{\perp} - \mathbf{x}_k^{\perp} = \mathbf{f}^{\perp}(\mathbf{x}_k)\Delta t + \sigma^{\perp}\mathbf{z}^{\perp}\sqrt{\Delta t}\end{aligned}$$

where $\mathbf{z}^{\|}$ ($\mathbf{z}^{\perp}$) is a vector of dimension k ($n-k$) drawn from a multivariate Gaussian with identity covariance. Hence, the true probability of a sequence $\{\mathbf{x}_i\}_{i=1}^N$ is given by

$$\begin{aligned}P\left(\{\mathbf{x}_i\}_{i=1}^N\right) &= \prod_{i=1}^{N-1} \frac{1}{(2\pi\sigma^{\|2}\Delta t)^{k/2}(2\pi\sigma^{\perp 2}\Delta t)^{(n-k)/2}}\\ &\qquad \exp\left(-\frac{\|\Delta\mathbf{x}_i^{\|} - \mathbf{f}^{\|}(\mathbf{x}_i)\Delta t\|^2}{2\sigma^{\|2}\Delta t} - \frac{\|\Delta\mathbf{x}_i^{\perp} - \mathbf{f}^{\perp}(\mathbf{x}_i)\Delta t\|^2}{2\sigma^{\perp 2}\Delta t}\right).\end{aligned}$$

We will approximate this distribution by a distribution Q which we assume has the following form

$$\begin{aligned}Q\left(\{\mathbf{x}_i\}_{i=1}^N\right) &= \prod_{i=1}^{N-1} \frac{1}{(2\pi\sigma^{\|2}\Delta t)^{k/2}(2\pi\sigma^{\perp 2}\Delta t)^{(n-k)/2}}\\ &\qquad \exp\left(-\frac{\|\Delta\mathbf{x}_i^{\|} - (\mathbf{A}_i^{\|}\mathbf{x}_i + \mathbf{b}_i^{\|})\Delta t\|^2}{2\sigma^{\|2}\Delta t} - \frac{\|\Delta\mathbf{x}_i^{\perp} - (\mathbf{A}_i^{\perp}\mathbf{x}_i + \mathbf{b}_i^{\perp})\Delta t\|^2}{2\sigma^{\perp 2}\Delta t}\right),\end{aligned}$$

where the matrices $\mathbf{A}_i^{\|}$, $\mathbf{A}_i^{\perp}$ and vectors $\mathbf{b}_i^{\|}$, $\mathbf{b}_i^{\perp}$ are parameters that will be adjusted to minimise the KL divergence $\mathrm{KL}(Q\|P)$ between the two distributions.

We present here an approximation of the exact optimisation of the KL divergence that extends the Kalman filter approach. The detailed derivations are given in Appendix B. We summarise the results here.

If we define

$$\begin{aligned}\mathbf{A}_i &= -\langle(\mathbf{b}_i - \mathbf{f}(\mathbf{x}_i))\mathbf{x}_i'\rangle\langle\mathbf{x}_i\mathbf{x}_i'\rangle^{-1} = \mathbf{A}(\mathbf{m}_i, \mathbf{\Sigma}_i)\\ \text{and}\quad \mathbf{b}_i &= \langle\mathbf{f}(\mathbf{x}_i)\rangle - \mathbf{A}_i\langle\mathbf{x}_i\rangle = \mathbf{b}(\mathbf{m}_i, \mathbf{\Sigma}_i),\end{aligned}$$

where averages are over the Gaussian approximation at stage i with mean $\mathbf{m}_i$ and covariance $\mathbf{\Sigma}_i$. Taking limits as the interval tends to zero we obtain differential equations for these quantities which are now all functions of t:

$$\frac{d\mathbf{m}}{dt} = \mathbf{A}(\mathbf{m}, \mathbf{\Sigma})\mathbf{m} + \mathbf{b}(\mathbf{m}, \mathbf{\Sigma}) = \langle\mathbf{f}(\mathbf{x})\rangle,$$

and

$$\begin{aligned}\frac{d\boldsymbol{\Sigma}}{dt} &= \mathrm{diag}(\sigma) + \mathbf{A}(\mathbf{m}, \boldsymbol{\Sigma})\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\mathbf{A}(\mathbf{m}, \boldsymbol{\Sigma})' \\ &= \mathrm{diag}(\sigma) + \left\langle \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \right\rangle \boldsymbol{\Sigma} + \boldsymbol{\Sigma} \left\langle \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \right\rangle',\end{aligned}$$

These equations define the evolution of our approximation of the Q process.

3 Discussion

The final version of the paper will present results for a Gaussian process with the covariance function determined by the approximation of the process Q. A discussion of how the kernel can be obtained from the above computations is included in Appendix D.

References

[1] D Cornford, L Csató, D J Evans, and M Opper. Bayesian analysis of the scatterometer wind retrieval inverse problem: some new approaches. *Journal of the Royal Statistical Society B*, 66:609–626, 2004.

[2] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME, Journal of Basic Engineering*, 82:34–45, 1960.

[3] E Kalnay. *Atmospheric Modelling, Data Assimilation and Predictability.* Cambridge University Press, Cambridge, 2003.

[4] A. C. Lorenc. Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112:1177–1194, 1986.

[5] A C Lorenc. The potential of the ensemble Kalman filter for NWP? a comparison with 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, 129:3183–3203, 2003.

[6] Carl Edward Rasmussen and Christopher K.I. Williams. *Gaussian Processes for Machine Learning.* The MIT Press, Cambridge, Massachusetts, 2006.

A Link to the Kalman Filter

A standard approach for modelling dynamical systems with additive Gaussian measurement noise is the Kalman Filter (KF) [2]. When the system is nonlinear, one of its variants, for example the extended Kalman Filter (EKF), can be used. KF and EKF are concerned with propagating the two first moments of the filtering distribution $p(\mathbf{x}_{i+1}|\mathbf{y}_{0:i+1})$, which are given by

$$\bar{\mathbf{x}}_{i+1} = \mathbb{E}\{\mathbf{x}_i|\mathbf{y}_{0:i+1}\}, \tag{1}$$

$$\bar{\mathbf{S}}_{i+1} = \mathbb{E}\{(\mathbf{x}_i - \bar{\mathbf{x}}_{i+1})(\mathbf{x}_{i+1} - \bar{\mathbf{x}}_{i+1})^{\mathrm{T}}|\mathbf{y}_{0:i+1}\}, \tag{2}$$

where $\mathbf{y}_{0:i+1} \equiv \{\mathbf{y}_0, \ldots, \mathbf{y}_{i+1}\}$ are the observations up to time t_{i+1}. In order to propagate these quantities, they proceed in two steps. In the *prediction step*, the two moments are estimated given the observations up to time t_i:

$$p(\mathbf{x}_{i+1}|\mathbf{y}_{0:i}) = \int p(\mathbf{x}_{i+1}|\mathbf{x}_i)p(\mathbf{x}_i|\mathbf{y}_{0:i})d\mathbf{x}_i. \tag{3}$$

This allows computing the predicted state $\mathbb{E}\{\mathbf{x}_{i+1}|\mathbf{y}_{0:i}\}$ and the predicted state covariance $\mathbb{E}\{(\mathbf{x}_{i+1} - \bar{\mathbf{x}}_{i+1})(\mathbf{x}_{i+1} - \bar{\mathbf{x}}_{i+1})^{\mathrm{T}}|\mathbf{y}_{0:i}\}$. Next, the *correction step* consists in updating these estimates based on the new observation $\mathbf{y}_{i+1}$:

$$p(\mathbf{x}_{i+1}|\mathbf{y}_{0:i+1}) \propto p(\mathbf{y}_{i+1}|\mathbf{x}_{i+1})p(\mathbf{x}_{i+1}|\mathbf{y}_{0:i}), \tag{4}$$

which leads to the desired conditional moments (1) and (2). KF is particularly attractive for on-line learning as it is not required to keep trace of the previous conditional expectations. Unfortunately, the integral in (3) is in general intractable when the system is nonlinear or when the state transition probability $p(\mathbf{x}_{i+1}|\mathbf{x}_i)$ is non-Gaussian. Therefore, approximations are required.

A common approach is to linearize the system, which corresponds to EKF. This leads to a Gaussian approximation of the transition probability. Moreover, if the likelihood $p(\mathbf{y}_{i+1}|\mathbf{x}_{i+1})$ is assumed to be Gaussian, then the filtering density is a Gaussian one at each t_{i+1}. An alternative approach is to resume the past information by the marginal $Q(\mathbf{x}_{i+1})$ and make predictions as follows

$$p(\mathbf{x}_{i+1}|\mathbf{y}_{0:i}) = \mathcal{N}(\mathbf{x}_{i+1}|\mathbf{m}_{i+1}, \mathbf{\Sigma}_{i+1}), \tag{5}$$

where

$$\mathbf{m}_{i+1} = \mathbf{m}_i + (\mathbf{A}\mathbf{m}_i + \mathbf{b}_i)\Delta t \tag{6}$$

$$\mathbf{\Sigma}_{i+1} = \mathbf{\Sigma}_i + (2\mathbf{A}\mathbf{\Sigma}_i + \mathrm{diag}\{\boldsymbol{\sigma}\})\Delta t. \tag{7}$$

This approximation is expected to be better than EKF, as the parameters $\mathbf{A}$ and $\mathbf{b}$ of the linear approximation are adjusted at each iteration. If we further assume that the likelihood $p(\mathbf{y}_{i+1}|\mathbf{x}_{i+1})$ is of the form $\mathcal{N}(\mathbf{y}_{i+1}|\mathbf{x}_{i+1}, \mathbf{R})$, then the correction step (4) is given by

$$p(\mathbf{x}_{i+1}|\mathbf{y}_{0:i+1}) = \mathcal{N}(\mathbf{y}_{i+1}|\bar{\mathbf{x}}_{i+1}, \bar{\mathbf{S}}_{i+1}), \tag{8}$$

with

$$\bar{\mathbf{x}}_{i+1} = \bar{\mathbf{S}}_{i+1}(\boldsymbol{\Sigma}_{i+1}^{-1}\mathbf{m}_{i+1} + \mathbf{R}^{-1}\mathbf{y}_{i+1}), \tag{9}$$

$$\bar{\mathbf{S}}_{i+1} = (\boldsymbol{\Sigma}_{i+1}^{-1} + \mathbf{R}^{-1})^{-1}. \tag{10}$$

Note that in this approach, only the filtering density (and its associated moments) are propagated through time. In contrast, GP framework allows us to define a distribution over the entire function space (i.e., over time). It is expected that this will have a smoothing effect and will lead to a better tracking of the state transitions.

B Approximation of KL divergence

If we use the variable $\mathbf{X}$ to denote the complete sequence $\{\mathbf{x}_i\}_{i=1}^N$, we can compute this as

$$\begin{aligned}
\mathrm{KL}(Q\|P) &= \mathbb{E}_{\mathbf{X}\sim Q}\left[\log \frac{Q(\mathbf{X})}{P(\mathbf{X})}\right] \\
&= \sum_{i=1}^{N-1} \mathbb{E}_{\mathbf{X}\sim Q}\left[\log \frac{Q(\mathbf{x}_{i+1}|\mathbf{x}_i)}{P(\mathbf{x}_{i+1}|\mathbf{x}_i)}\right] \\
&= \sum_{i=1}^{N-1} \int d\mathbf{x}_i Q(\mathbf{x}_i) \int d\mathbf{x}_{i+1} Q(\mathbf{x}_{i+1}|\mathbf{x}_i) \\
&\quad \left[\frac{\|\Delta\mathbf{x}_i^{\|} - \mathbf{f}^{\|}(\mathbf{x}_i)\Delta t\|^2 - \|\Delta\mathbf{x}_i^{\|} - (\mathbf{A}_i^{\|}\mathbf{x}_i + \mathbf{b}_i^{\|})\Delta t\|^2}{2\sigma^{\|2}\Delta t}\right. \\
&\quad \left. + \frac{\|\Delta\mathbf{x}_i^{\perp} - \mathbf{f}^{\perp}(\mathbf{x}_i)\Delta t\|^2 - \|\Delta\mathbf{x}_i^{\perp} - (\mathbf{A}_i^{\perp}\mathbf{x}_i + \mathbf{b}_i^{\perp})\Delta t\|^2}{2\sigma^{\|2}\Delta t}\right] \\
&= \sum_{i=1}^{N-1} \int d\mathbf{x}_i Q(\mathbf{x}_i) \left[\frac{\Delta t}{2\sigma^{\|2}}\|(\mathbf{A}_i^{\|}\mathbf{x}_i + \mathbf{b}_i^{\|}) - \mathbf{f}^{\|}(\mathbf{x}_i)\|^2\right. \\
&\qquad \left. + \frac{\Delta t}{2\sigma^{\perp 2}}\|(\mathbf{A}_i^{\perp}\mathbf{x}_i + \mathbf{b}_i^{\perp}) - \mathbf{f}^{\perp}(\mathbf{x}_i)\|^2\right],
\end{aligned}$$

where the last equality follows from using the equality

$$\mathbb{E}_{x\sim N(\mu,\sigma)}[(x-a)^2] = \sigma^2 + (\mu - a)^2$$

for the k components of the first vector and $n-k$ components of the second.

In order to minimise the KL divergence, we must take the derivative with respect to the parameters, set to zero and solve. We obtain

$$\begin{aligned}
2\mathbf{A}_i^{\|}\langle \mathbf{x}_i\mathbf{x}_i'\rangle + 2\langle(\mathbf{b}_i^{\|} - \mathbf{f}^{\|}(\mathbf{x}_i))\mathbf{x}_i'\rangle &= \mathbf{0} \\
\text{and} \quad 2(\mathbf{A}_i^{\|}\langle \mathbf{x}_i\rangle + \mathbf{b}_i^{\|} - \langle \mathbf{f}^{\|}(\mathbf{x}_i)\rangle) &= \mathbf{0}, \quad \text{and} \\
2\mathbf{A}_i^{\perp}\langle \mathbf{x}_i\mathbf{x}_i'\rangle + 2\langle(\mathbf{b}_i^{\perp} - \mathbf{f}^{\perp}(\mathbf{x}_i))\mathbf{x}_i'\rangle &= \mathbf{0} \\
\text{and} \quad 2(\mathbf{A}_i^{\perp}\langle \mathbf{x}_i\rangle + \mathbf{b}_i^{\perp} - \langle \mathbf{f}^{\perp}(\mathbf{x}_i)\rangle) &= \mathbf{0}
\end{aligned}$$

where the angle brackets indicate expectations with respect to the marginal distribution $Q(\mathbf{x}_i)$, which is Gaussian with mean and covariance $\mathbf{m}_i$ and $\mathbf{\Sigma}_i$ respectively (note that these are over the full variable set). We obtain the equations for the parameters as

$$\begin{aligned}
\mathbf{A}_i &= -\langle(\mathbf{b}_i - \mathbf{f}(\mathbf{x}_i))\mathbf{x}_i'\rangle\langle \mathbf{x}_i\mathbf{x}_i'\rangle^{-1} = \mathbf{A}(\mathbf{m}_i, \mathbf{\Sigma}_i) \\
\text{and} \quad \mathbf{b}_i &= \langle \mathbf{f}(\mathbf{x}_i)\rangle - \mathbf{A}_i\langle \mathbf{x}_i\rangle = \mathbf{b}(\mathbf{m}_i, \mathbf{\Sigma}_i),
\end{aligned}$$

where the matrix $\mathbf{A}_i$ is formed by concatenating $\mathbf{A}_i^{\|}$ and $\mathbf{A}_i^{\perp}$ and similarly $\mathbf{b}_i$. The expressions for $\mathbf{A}(\mathbf{m}, \mathbf{\Sigma})$ and $\mathbf{b}(\mathbf{m}, \mathbf{\Sigma})$ can be given as

$$\begin{aligned}
\mathbf{A}(\mathbf{m}, \mathbf{\Sigma})(\mathbf{\Sigma} + \mathbf{m}\mathbf{m}') &= \langle \mathbf{f}(\mathbf{x})\mathbf{x}'\rangle - \mathbf{b}\mathbf{m}' \\
\mathbf{b}(\mathbf{m}, \mathbf{\Sigma}) &= \langle \mathbf{f}(\mathbf{x})\rangle - \mathbf{A}\mathbf{m}.
\end{aligned}$$

Substituting $\mathbf{b}$ from the second equation in the first gives

$$\begin{aligned}
\mathbf{A}(\mathbf{m}, \mathbf{\Sigma})(\mathbf{\Sigma} + \mathbf{m}\mathbf{m}') &= \langle \mathbf{f}(\mathbf{x})\mathbf{x}'\rangle - \langle \mathbf{f}(\mathbf{x})\rangle\mathbf{m}' + \mathbf{A}\mathbf{m}\mathbf{m}' \\
\Rightarrow \mathbf{A}(\mathbf{m}, \mathbf{\Sigma})\mathbf{\Sigma} &= \langle \mathbf{f}(\mathbf{x})(\mathbf{x}' - \mathbf{m}')\rangle \\
&= \left\langle \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \right\rangle \mathbf{\Sigma}, \qquad (11)
\end{aligned}$$

with the last equality following from an integration by parts. Writing expressions for the generation of $\mathbf{x}_{i+1}$ we have

$$\begin{aligned}
\mathbf{x}_{i+1}^{\|} &= \mathbf{x}_i^{\|} + (\mathbf{A}^{\|}\mathbf{x}_i + \mathbf{b}^{\|})\Delta t + \sigma^{\|}\mathbf{u}^{\|}\sqrt{\Delta t} \\
\mathbf{x}_{i+1}^{\perp} &= \mathbf{x}_i^{\perp} + (\mathbf{A}^{\perp}\mathbf{x}_i + \mathbf{b}^{\perp})\Delta t + \sigma^{\perp}\mathbf{u}^{\perp}\sqrt{\Delta t}
\end{aligned}$$

where $\mathbf{u}$ is an n-dimensional vector of independent zero mean Gaussian variables with unit variance. We can combine these into the single equation

$$\mathbf{x}_{i+1} = \mathbf{x}_i + (\mathbf{A}\mathbf{x}_i + \mathbf{b})\Delta t + \mathrm{diag}(\sigma)\mathbf{u}\sqrt{\Delta t}$$

where $\mathrm{diag}(\sigma)$ denotes the diagonal matrix whose first k entries are $\sigma^{\|}$ and remaining entries $\sigma^{\perp}$. We also know that $\mathbf{x}_i$ is generated by a Gaussian with mean $\mathbf{m}_i$ and covariance Σ_i, so that

$$\mathbf{x}_i = \mathbf{m}_i + \sqrt{\Sigma_i}\mathbf{v}$$

where $\mathbf{v}$ is a vector of zero mean unit variance Gaussian variables. Hence, we obtain the following expression for $\mathbf{x}_{i+1}$

$$\mathbf{x}_{i+1} = \mathbf{m}_i + \sqrt{\Sigma_i}\mathbf{v} + \mathbf{A}\mathbf{m}_i\Delta t + \mathbf{A}\sqrt{\Sigma_i}\mathbf{v}\Delta t + \mathbf{b}\Delta t + \mathrm{diag}(\sigma)\mathbf{u}\sqrt{\Delta t}$$

Hence, we can compute

$$\mathbf{m}_{i+1} = \mathbb{E}[\mathbf{x}_{i+1}] = \mathbf{m}_i + \mathbf{A}\mathbf{m}_i\Delta t + \mathbf{b}\Delta t.$$

Taking the difference between means, dividing by Δt and taking limits gives

$$\frac{d\mathbf{m}}{dt} = \mathbf{A}(\mathbf{m}, \mathbf{\Sigma})\mathbf{m} + \mathbf{b}(\mathbf{m}, \mathbf{\Sigma}) = \langle \mathbf{f}(\mathbf{x}) \rangle.$$

Next consider

$$\begin{aligned}
\Sigma_{i+1} &= \mathbb{E}\left[(\mathbf{x}_{i+1} - \mathbf{m}_{i+1})(\mathbf{x}_{i+1} - \mathbf{m}_{i+1})'\right] \\
&= \mathbb{E}\left[\sqrt{\Sigma_i}\mathbf{v}\mathbf{v}'\sqrt{\Sigma_i}\right] + \mathbb{E}\left[\sqrt{\Sigma_i}\mathbf{v}\mathbf{v}'\sqrt{\Sigma_i}\mathbf{A}'\Delta t\right] + \mathbb{E}\left[\mathbf{A}\sqrt{\Sigma_i}\mathbf{v}\mathbf{v}'\sqrt{\Sigma_i}\Delta t\right] \\
&\qquad + \mathrm{diag}(\sigma)\Delta t + O\left((\Delta t)^2\right) \\
&= \Sigma_i + \Sigma_i\mathbf{A}'\Delta t + \mathbf{A}\Sigma_i\Delta t + \mathrm{diag}(\sigma)\Delta t + O\left((\Delta t)^2\right)
\end{aligned}$$

Taking the difference between means, dividing by Δt and taking limits gives

$$\begin{aligned}
\frac{d\mathbf{\Sigma}}{dt} &= \mathrm{diag}(\sigma) + \mathbf{A}(\mathbf{m}, \mathbf{\Sigma})\mathbf{\Sigma} + \mathbf{\Sigma}\mathbf{A}(\mathbf{m}, \mathbf{\Sigma})' \\
&= \mathrm{diag}(\sigma) + \left\langle \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \right\rangle \mathbf{\Sigma} + \mathbf{\Sigma}\left\langle \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \right\rangle',
\end{aligned}$$

where we have made use of equation (11). The result of this computation hold for all values of $\sigma^{\|}$ and $\sigma^{\perp}$, so that we can consider the case where we let $\sigma^{\perp}$ tend to zero. This allows us to encode quite general noise models and covariances as the examples in Appendix C illustrate.

C Examples of different noise models

General covariance If we wish to introduce noise with a known fixed covariance Σ_0 as for example given by spatial relations between the locations of a climate grid model, we can double the number of variables to $n = 2k$ to obtain

$$\begin{aligned} \Delta \mathbf{x}_k^{\parallel} &= \mathbf{x}_{k+1}^{\parallel} - \mathbf{x}_k^{\parallel} = \mathbf{z}\sqrt{\Delta t}, \\ \Delta \mathbf{x}_k^{\perp} &= \mathbf{x}_{k+1}^{\perp} - \mathbf{x}_k^{\perp} = \mathbf{f}(\mathbf{x}_k^{\perp})\Delta t + \sqrt{\Sigma_0}\mathbf{x}_k^{\parallel}, \end{aligned}$$

where $\mathbf{z}$ is k dimensional zero mean unit variance Gaussian random variables and $\mathbf{f}(\mathbf{x}^{\perp})$ is the system being studied.

Ornstein-Uhlenbeck process coloured noise Consider the two variable stochastic differential equation:

$$\begin{aligned} \Delta y_k &= y_{k+1} - y_k = -m y_k \Delta t + \sigma z \sqrt{\Delta t}, \\ \Delta x_k &= x_{k+1} - x_k = (f(x_k) + y_k)\Delta t, \end{aligned}$$

where z is unit variance, zero mean Gaussian and $f(\cdot)$ is some possibly non-linear function. We can approximate this system using the above model with $n = 2$ and $k = 1$, $\sigma^{\parallel} = \sigma$ and $\sigma^{\perp} \to 0$. It gives a general one-dimensional system driven by coloured noise.

D On the kernel

The problem is: Consider

$$d\mathbf{x} = (\mathbf{A}\mathbf{x} + \mathbf{b})dt + \sigma d\mathbf{W}$$

with $\mathbf{W}$ a standard Wiener process (assuming uncorrelated noise with equal variance). The goal is to find the two time covariance kernel

$$\mathbf{K}(t_1, t_2) = \mathbb{E}\left[(\mathbf{x}(t_1) - \mathbf{m}(t_1))(\mathbf{x}(t_2) - \mathbf{m}(t_2))'\right]$$

Let $\mathbf{U}$ be a solution to homogenous (nonstochastic) equation

$$\frac{d\mathbf{U}}{dt} = \mathbf{A}\mathbf{U}$$

with $\mathbf{U}(0) = \mathbf{I}$. Then we can solve the inhomogenous equation as

$$\mathbf{x}(t) = \mathbf{U}(t)\mathbf{x}(0) + \sigma \int_0^t \mathbf{U}(t-s)d\mathbf{W}(s)$$

The naive play with the differentials seems to be justified for the linear case! We understand that the first part is nonrandom, the second part is of zero mean. Hence, if we want the covariance, we just have

$$\mathbf{K}(t_1, t_2) = \sigma^2 \int_0^{t_1} \int_0^{t_2} \mathbf{U}(t_1 - s_1)\mathbb{E}[d\mathbf{W}(s_1)d\mathbf{W}'(s_2)]\mathbf{U}(t_2 - s_2)$$

We also have $\mathbb{E}[d\mathbf{W}(s_1)d\mathbf{W}(s_2)'] = \mathbf{I}\,\delta(s_1 - s_2)ds_1 ds_2$. So we should end up with

$$\mathbf{K}(t_1, t_2) = \sigma^2 \int_0^{\min(t_1,t_2)} ds\ \mathbf{U}(t_1 - s)\mathbf{U}'(t_2 - s)$$

So far we have not used any properties of time independence. The simplification that comes with a constant $\mathbf{A}$ is that we can immediately write

$$\mathbf{U}(t) = e^{\mathbf{A}t}$$

For the time dependent case, finding explicit solutions is problematic by the fact that matrices at different times may not commute, i.e. $\mathbf{A}(t)\mathbf{A}(t') \neq \mathbf{A}(t')\mathbf{A}(t)$. The *time independent case* gives (assume $t_1 < t_2$)

$$\mathbf{K}(t_1, t_2) = \sigma^2 e^{\mathbf{A}(t_2 - t_1)} \int_0^{t_1} ds\ e^{\mathbf{A}(t_1 - s)} e^{\mathbf{A}'(t_1 - s)}$$

Now, the integral (with σ^2) is precisely the kernel at equal times (the variance) and we have the final result.

$$\mathbf{K}(t_1, t_2) = e^{\mathbf{A}(t_2 - t_1)}\mathbf{K}(t_1, t_1) \qquad \text{for } t_1 < t_2$$

and similar

$$\mathbf{K}(t_1, t_2) = \mathbf{K}(t_2, t_2)e^{\mathbf{A}'(t_1 - t_2)} \qquad \text{for } t_2 < t_1$$