What is the correct cost functional for variational data assimilation?

Article

Accepted Version

Noname manuscript No.
(will be inserted by the editor)

What is the correct cost functional for variational data assimilation?

Jochen Bröcker

the date of receipt and acceptance should be inserted later

Abstract Variational approaches to data assimilation, and weakly constrained four dimensional variation (WC-4DVar) in particular, are important in the geosciences but also in other communities (often under different names). The cost functions and the resulting optimal trajectories may have a probabilistic interpretation, for instance by linking data assimilation with Maximum Aposteriori (MAP) estimation. This is possible in particular if the unknown trajectory is modelled as the solution of a stochastic differential equation (SDE), as is increasingly the case in weather forecasting and climate modelling. In this case, the MAP estimator (or "most probable path" of the SDE) is obtained by minimising the Onsager–Machlup functional. Although this fact is well known, there seems to be some confusion in the literature, with the energy (or "least squares") functional sometimes been claimed to yield the most probable path. The first aim of this paper is to address this confusion and show that the energy functional does not, in general, provide the most probable path. The second aim is to discuss the implications in practice. Although the mentioned results pertain to stochastic models in continuous time, they do have consequences in practice where SDE's are approximated by discrete time schemes. It turns out that using an approximation to the SDE and calculating its most probable path does not necessarily yield a good approximation to the most probable path of the SDE proper. This suggest that even in discrete time, a version of the Onsager–Machlup functional should be used, rather than the energy functional, at least if the solution is to be interpreted as a MAP estimator.

Keywords Variational Data Assimilation, Onsager–Machlup Functional, Stochastic Differential Equations

The author was supported by the UK Engineering and Physical Sciences Research Council under grant agreement EP/L012669/1. Fruitful discussions with Tobias Kuna, Dan Crisan, Andrew Stuart, Colin Cotter, and Horatio Boedihardjo are gratefully acknowledged. Referee Stéphàne Vannitsem and a second anonymous referee provided a number of important comments which helped to improve this manuscript.

School of Mathematical and Physical Sciences, University of Reading, United Kingdom,
E-mail: j.broecker@reading.ac.uk

1 Introduction

In the geosciences, the term data assimilation refers to a variety of mathematical and numerical techniques whereby time series of observations are employed to estimate states or trajectories of relevant dynamical models. In other words, plausible states or orbits are determined which, on the one hand, are consistent with a given dynamical model and, on the other hand, are consistent with a given set of observations. Many different approaches to data assimilation exist, based on very different philosophies and premises, see for instance Ide et al. (1997); Kalnay (2001); Evensen (2007), but this list is by no means complete.

Both within the atmospheric sciences, but also in other branches of physics and engineering, variational approaches have gained widespread attention (although the nomenclature may differ considerably). A particular instance of this idea is known as weakly constrained four dimensional variation (WC-4DVar) in atmospheric sciences; basically, a series of model states is found that minimises a cost functional which quantifies both the deviations from the observed data as well as the misfit with the given model. An early paper on discrete time WC-4DVar in atmospheric sciences is Derber (1989), see also Kalnay (2001). The cost function is almost invariably some form of quadratic error, and for this reason, the technique is known as the minimum energy estimator in the engineering community, see for instance Jazwinski (1970) or Mortensen (1968) (in the latter publication, the authors go further and derive an incremental version).

In the atmospheric sciences and in particular in climate modelling, stochastic models are becoming ever more important, despite having a long and distinguished history already (see for instance Imkeller and von Storch 2001; Franzke et al. 2015, and references therein). Mathematically speaking, climate models increasingly take the form of stochastic differential equations (SDE's). Consequently, data assimilation into such models needs well understood foundations. In particular, if variational data assimilation into SDE's is envisaged, the question arises as to what cost function should be used, and in particular whether the cost functions and the resulting optimal trajectories have any probabilistic interpretation. A possible avenue is to link variational data assimilation with Maximum Aposteriori (MAP) estimation. The MAP estimator of a random variable given some observations is essentially the maximiser of the posterior, that is, of the conditional density of the unknown random variable given the observations. In some sense, the MAP estimator can be interpreted as the "most probable value" of the unknown random variable given the observation. The concept of density generalises to situations where the unknown random variable is an entire function, given by the solution of a stochastic differential equation (SDE), and the MAP estimator becomes the "most probable path" of the SDE (see e.g. Zeitouni and Dembo (1987), Zeitouni and Dembo (1988); for MAP estimation in classical inverse problems but with random observations see Cotter et al. (2009); see also Apte et al. (2007); Stuart (2010) for applications to Bayesian estimation in stochastic dynamical systems). Contrary to what is sometimes claimed in the literature, the most probable path of an SDE is not a minimiser of the energy functional but rather of the Onsager–Machlup functional, which differs from the energy functional in that the latter contains extra terms. In other words, to find MAP estimators or most probable paths for SDE's, the Onsager–Machlup functional has to be minimised, rather than the energy functional.

The first aim of this paper is to illustrate this well known fact. The reader is referred to Zeitouni and Dembo (1987), Zeitouni and Dembo (1988) for a rigorous derivation of the Onsager–Machlup functional and discussion of the MAP estimator in the context of SDE's. The second aim is to show that although this is a result pertaining to stochastic models in continuous time, it does have consequences in discrete time. In practice, SDE's are approximated by discrete time schemes, for instance the Euler scheme which results in discrete time stochastic dynamical system with additive Gaussian errors. The (negative logarithm of the) density of solutions to this discrete time system is given by the energy functional. But we will argue that the appropriate functional in this situation should *still* be the Onsager–Machlup functional or a discrete time version thereof, at least if the solution is to be interpreted as a MAP estimator. The reason is that the MAP estimator (or most probable path) of an approximation to the SDE is not necessarily a good approximation to the most probable path of the SDE proper, as we will see. It is worth noting that this point involves the dynamics only and is entirely independent of whether observations are considered discrete or continuous in time.

In Section 2, we revisit the concepts of densities for random variables and the MAP estimator. In Section 3, we specialise to the situation were the unknown random variable is a trajectory of a stochastic differential equation, and demonstrate that the energy functional cannot be the correct functional to determine the MAP estimator. An expression for the Onsager–Machlup functional will also be provided. The findings will be supported by numerical simulations in Section 4. Further, these simulations illustrate that the Onsager–Machlup functional essentially provides the correct density for paths of SDE's even though the simulations are not truly continuous in time but rather use an approximation scheme that is discrete in time. Section 5 provides the Onsager–Machlup functional for more general SDE's that are not used in the present paper but which are relevant for the climate sciences, namely SDE's with multiplicative noise (see e.g. Franzke et al. 2015)[1]. Section 6 concludes with a discussion as to how our findings bear on discrete time simulations of SDE's. An informal derivation of the Onsager–Machlup functional is provided in Appendix A.

2 Definition of the Maximum Aposteriori (MAP) estimator

A fundamental concept in statistics in general and data assimilation in particular is the Maximum Aposteriori or MAP estimator. Let X, Y be random variables, where we interprete X as the unknown quantity (to be estimated) and Y as the observation. Let $p(x|y)$ denote the conditional probability density function of X given that Y assumes the value y. A *MAP estimator* of X given Y is a maximiser over x of the density $p(x|y)$. That is, the MAP estimator is a function $\hat{x}(y)$ so that for any y we have

$$p(\hat{x}(y)|y) = \sup_x p(x|y).$$

MAP estimators need not exist in general, nor are they unique.

Since the observations Y play the role of parameters in this problem, they will mostly suppressed in the notation for the sake of simplicity. That is, if X is a

[1] We are grateful to referee Stéphàne Vannitsem for stressing this point.

random variable with density p_X, we understand that p_X might in fact be the conditional density of X given some observations or parameters.

The presented definition of the MAP estimator will be referred to as the *de facto* definition (following Dutra et al. (2014)); there is an alternative definition which not only provides an intuitive interpretation but is more generally applicable. Roughly speaking, the MAP estimator of a random variable X is the center of a small ball positioned so as to have greatest possible probability of containing X, in the limit of the diameter of that ball going to zero. More formally, suppose that X is a random variable with values in some vector space V with norm $\|.\|$. Then the MAP estimator is a point $\hat{x}$ so that for any other point x

$$\limsup_{\epsilon\to 0} \frac{\mathbb{P}(\|X-x\|\le\epsilon)}{\mathbb{P}(\|X-\hat{x}\|\le\epsilon)} \le 1. \tag{1}$$

If observations are present, then these probabilities are conditional probabilities given those observations.

If a random variable X with values in $\mathbb{R}^d$ has a density p which is everywhere positive, then a MAP estimator according to the alternative definition (1) is a MAP estimator according to the *de facto* definition and vice versa. Indeed, if X has a positive density p, then for all $x\in\mathbb{R}^d$ the relation

$$p(x) = \lim_{\epsilon\to 0} \frac{\mathbb{P}(\|X-x\|\le\epsilon)}{\mathrm{vol}\{z\in\mathbb{R}^n; \|z\|\le\epsilon\}} \tag{2}$$

holds (except perhaps if x is in some exceptional set which has however volume zero; we will ignore this technical point). Here, vol denotes the standard volume on $\mathbb{R}^d$. Hence, if y is so that $p(y)>0$, then for any x we have

$$\begin{aligned}\frac{p(x)}{p(y)} &= \lim_{\epsilon\to 0} \frac{\mathbb{P}(\|X-x\|\le\epsilon)}{\mathrm{vol}\{z\in\mathbb{R}^d; \|z\|\le\epsilon\}} \lim_{\epsilon\to 0} \frac{\mathrm{vol}\{z\in\mathbb{R}^d; \|z\|\le\epsilon\}}{\mathbb{P}(\|X-y\|\le\epsilon)} \\ &= \lim_{\epsilon\to 0} \frac{\mathbb{P}(\|X-x\|\le\epsilon)}{\mathbb{P}(\|X-y\|\le\epsilon)}.\end{aligned} \tag{3}$$

The relation (3) shows that any point $\hat{x}\in\mathbb{R}^d$ which satisfies the de facto definition of a MAP estimator will also satisfy the alternative definition and vice versa.

A strong point of the *de facto* definition is that it provides a means to find a MAP estimator through an optimisation problem. An important insight from the alternative definition though is that it is not quite necessary to have a probability density function as in Equation (2) in order to define the MAP estimator. In particular the normalisation in Equation (2) need not be the standard volume; normalising in a different way would give a different density, but as long as the normalisation is the same for all reference points x and the resulting density is still everywhere positive, we would obtain the same MAP estimators, since the relation (3) would still be valid. For instance, if W is another random variable, we could normalise as follows

$$p^{(W)}(x) = \lim_{\epsilon\to 0} \frac{\mathbb{P}(\|X-x\|\le\epsilon)}{\mathbb{P}(\|W\|\le\epsilon)} \tag{4}$$

if the limit exists for every x; if $p^{(W)}$ is everywhere positive, $p^{(W)}$ can be used to calculate the MAP just as well.

It turns out that generalised densities as in Equation (4) might still be well defined even if X has values in some infinite dimensional space with norm $\|.\|$ for which there exists no generalisation of the standard volume.[2] This is precisely the situation when trying to find MAP estimators for trajectories of continuous time stochastic dynamical models; such a trajectory is a function (of time) and hence an infinite dimensional object. Hence the Definition (3) of a density does not apply in this situation but Definition (4) does, provided we find a suitable random variable W to normalise with.

3 MAP estimators for stochastic difference and differential equations

The link between MAP estimators and data assimilation in discrete time can be described as follows. The dynamics underlying the observations is modelled as a stochastic difference equation of the form

$$X_n = F(X_{n-1}) + R_n, \qquad n = 1, 2, \ldots, N \tag{5}$$

where F is some mapping on a vector space E (called the state space), and the $R_n, n = 1, 2, \ldots$ are taken as independent and identically distributed random variables with values in E. For simplicity's sake, we assume throughout that E is one dimensional (see however Sec. 5). Further, the $R_n, n = 1, 2, \ldots$ are assumed to be normal with mean zero and variance γ. We further set $X_0 = \xi$, where $\xi \in E$ is known.

The observations are assumed to be functions of the $X_1, \ldots, X_n$ further corrupted by noise. But as said earlier, they will enter the densities as parameters in some way which is not relevant for our purposes. It is then a simple matter to show that

$$\begin{aligned} \lim_{\epsilon \to 0} \frac{\mathbb{P}(\max_n |X_n - x_n| \leq \epsilon)}{\mathbb{P}(\max_n |R_n| \leq \epsilon)} &= \frac{p_{X_1,\ldots,X_N}(x_1, \ldots, x_N)}{p_{R_1,\ldots,R_N}(0, \ldots, 0)} \\ &= \exp\left(-\frac{1}{2\gamma} \sum_{n=1}^{N} (x_n - F(x_{n-1}))^2 \right), \end{aligned} \tag{6}$$

where we understand that $x_0 = \xi$. Since $(X_1, \ldots, X_N)$ is a random variable in E^N, we can interprete the right hand side of Equation (6) as a density of $(X_1, \ldots, X_N)$ according to Definition (4) with $V = E^N$ and norm $\|(x_1, \ldots, x_N)\| = \max_n |x_n|$.

Atmospheric and ocean dynamics are, however, continuous in time, as are many other processes in science and engineering where data assimilation is relevant. Considering data assimilation in discrete time is merely a concession to practical constraints. Indeed, there are several different processess that introduce time stepping in operational practice, for instance the integration of the model or the batch processing of the observations, but the relevant time steps can be very different. Accounting for "model error" with additive noise after discretising models in time will result in the solutions for different time stepping having different

[2] The problem is the translation invariance of the standard volume. In an infinite dimensional normed space, a ball of unit radius may contain infinitely many disjoint balls of sufficiently small but nonzero radius. By translation invariance, these balls must have the same volume. But this means that either the volume of the unit ball is infinity or the volume of a sufficiently small ball is zero.

statistical properties. Although this is to some extent inevitable, we still ought to have a formalism for comparing these different solutions, as they ultimately represent the same thing.

A convenient way to enable comparison of different discretisations (with noise added) is to formulate a stochastic model in continuous time, that is, a stochastic differential equation (SDE), and consider any discretisation as an approximation of that model. The question then arising is what is the MAP estimator, or more generally the density, for trajectories of an SDE? To put this question more precisely, let $I = [0, T]$ be an interval of the real line, and consider the SDE

$$\dot{X}_t = f(X_t) + \rho r_t, \qquad t \in I \tag{7}$$

where f is a vector field on E, $\rho > 0$, and $r_t, t \in I$ is white noise with zero mean and unit intensity (i.e. the correlation function is $\delta(t - s)$ with δ the Dirac delta function). Again, we set $X_0 = \xi$, where $\xi \in E$ is known.

Whatever the precise interpretation of the SDE (7), the solution is a random continuous function $\{X_t, t \in I\}$, and the density of it at some given reference trajectory $\{z_t, t \in I\}$ is defined as

$$p(\{z_t\}) = \lim_{\epsilon \to 0} \frac{\mathbb{P}(\sup_{t \in I} |X_t - z_t| \leq \epsilon)}{\mathbb{P}(\sup_{t \in I} \rho |W_t| \leq \epsilon)} \tag{8}$$

where $\{W_t, t \in I\}$ is the *Wiener* process, which can be seen as the time integral of white noise, that is

$$W_t = \int_0^t r_s \mathrm{d}s.$$

We will learn more about the Wiener process later. Normalisation with the Wiener process in the Definition (8) of the density will turn out to be convenient.

It is worth stressing that the density in Definition (8) is a special case of the Definition (4) if we use the norm $\|z\| := \sup_{t \in I} |z_t|$ for trajectories over I. We also note that the density is zero for trajectories which do not start at the initial condition $z_0 = \xi$. For later use, we introduce the *ϵ–weight*

$$\alpha(\epsilon, \{z_t\}) = \frac{\mathbb{P}(\sup_{t \in I} |X_t - z_t| \leq \epsilon)}{\mathbb{P}(\sup_{t \in I} \rho |W_t| \leq \epsilon)}$$

of a trajectory $\{z_t, t \in I\}$. The ϵ–weight is the probability that the solution $\{X_t, t \in I\}$ of the SDE (7) falls entirely into a small strip or "sausage" of width ϵ around $\{z_t, t \in I\}$, relative to the probability that the Wiener process $\{W_t\}$ falls entirely into a "sausage" of width ϵ/ρ around zero. Figure 1 illustrates the situation. The density p according to Definition (8) is given by $p(\{z_t\}) = \lim_{\epsilon \to 0} \alpha(\epsilon, \{z_t\})$.

The density p can be written in the form

$$p(\{z_t\}) = \exp(-\mathcal{A}(\{z_t\})), \tag{9}$$

and several publications seem to imply that $\mathcal{A}(\{z_t\})$ should be equal to the *energy functional*

$$\mathcal{A}_E(\{z_t\}) = \frac{1}{2\rho^2} \int_I (\dot{z}_t - f(z_t))^2 \mathrm{d}t, \tag{10}$$

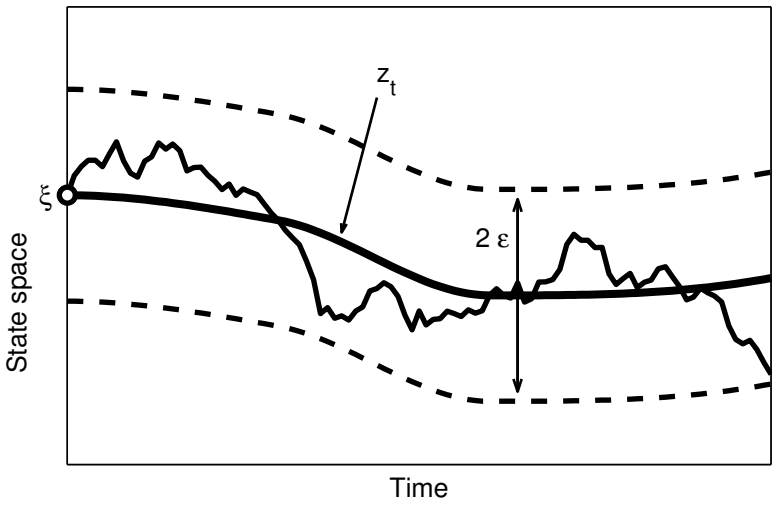

Fig. 1 The plot shows the event that the solution $\{X_t, t \in I\}$ of the SDE (7) (thin solid line) falls entirely into a small strip of width ϵ around the reference trajectory $\{z_t, t \in I\}$ (thick solid line). The strip is indicated with dashed lines. (Note that this is a schematic sketch rather than an actual simulation.)

or at least that the MAP estimator should be a minimiser of $\mathcal{A}_E$ (sometimes without clear reference to the concept of densities). In case observations are present, the energy estimator would carry another term pertaining to the observations.

As mentioned in the introduction already, the correct expression for the functional $\mathcal{A}$ in Equation (9) is not the energy functional but the *Onsager–Machlup* functional

$$\mathcal{A}_{OM}(\{z_t\}) = \mathcal{A}_E(\{z_t\}) + \frac{1}{2}\int_I f'(z_t)\mathrm{d}t. \tag{11}$$

An informal derivation of this expression will be given in Appendix A. Note however that for very small noise amplitudes, the energy functional $\mathcal{A}_E$ becomes the dominant term in the Onsager–Machlup functional, as this term scales inversely proportional with the noise, while the additional term does not depend on the noise at all. This suggests that data assimilation employing the energy functional does have a rigorous interpretation in the small noise limit. This is indeed the case, as discussed for instance in Vanden-Eijnden and Weare (2013), where the energy functional emerges from a large deviation principle. Furthermore, there are clearly other cases where the additional term in Equation (11) does not matter for the purposes of data assimilation, for instance if the dynamics is linear, as then the second term in Equation (11) is constant. In higher dimensions, the additional term is the integral over $\mathrm{div} f(z_t)$ (see Section 5) so that for systems with constant divergence, minimising the energy functional gives the same results as minimising the Onsager–Machlup functional.

In the remainder of this section, we will provide evidence that the expression (9) with the energy functional is *not* the correct density, and discuss possible reasons for this misconception. We write the SDE (7), somewhat more rigorously, as an

integral equation

$$X_t = \xi + \int_0^t f(X_s)\mathrm{d}s + \rho W_t, \qquad t \in I$$

where $W_t, t \in I$ is the standard Wiener process, which as we have seen can heuristically be interpreted as the integral of the white noise process r_t. In fact, from these heuristics, one can derive that the Wiener process ought to have the following properties:

1. $W_0 = 0$,
2. for $0 \le t_1 < t_2$ the *increment* $W_{t_2} - W_{t_1}$ is a normally distributed random variable with mean zero and covariance $t_2 - t_1$,
3. increments for nonoverlapping intervals are independent,

It is well known (see for instance Breiman (1973), Mörters and Peres (2010)) that a process $\{W_t, t \in I\}$ with the properties listed above exists and can be realised as a random continuous function of time. In view of this, the Equation (7) is a classical integral equation perturbed by a randomly selected function that is continuous in time.

Discretisation schemes for Equation (7) can be derived by observing that

$$X_{t_n} = X_{t_{n-1}} + \int_{t_{n-1}}^{t_n} f(X_s)\mathrm{d}s + \rho \cdot (W_{t_n} - W_{t_{n-1}}) \tag{12}$$

and approximating the integral in an appropriate way. For instance, using the approximation $\int_{t_{n-1}}^{t_n} f(X_s)\mathrm{d}s \cong f(X_{t_{n-1}})(t_n - t_{n-1})$ and assuming for simplicity a constant time step $(t_n - t_{n-1}) = \Delta$ results in the Euler scheme (also known as the Euler–Maruyama scheme, Milstein 1995)

$$X^{(\Delta)}_{t_n} = X^{(\Delta)}_{t_{n-1}} + f(X^{(\Delta)}_{t_{n-1}})\Delta + \rho \cdot (W_{t_n} - W_{t_{n-1}}). \tag{13}$$

(The superscript Δ indicates that this solution is obtained with the Euler scheme and time discretisation Δ). If we set $R_n = \rho \cdot (W_{t_n} - W_{t_{n-1}})$, then Equation (13) is precisely in the form of Equation (5) with $F(x) = x + f(x)\Delta$ and $\gamma = \Delta\rho^2$. Hence the density (6) for the solution $(X^{(\Delta)}_{t_1}, \ldots, X^{(\Delta)}_{t_N})$ of Equation (13) reads as

$$\begin{aligned} &\frac{p_{X^{(\Delta)}_{t_1},\ldots,X^{(\Delta)}_{t_N}}(x_1,\ldots,x_N)}{p_{R_1,\ldots,R_N}(0,\ldots,0)} \\ &= \exp\left[-\frac{\Delta}{2\rho^2}\sum_{n=1}^{N}\left(\frac{x_n - x_{n-1}}{\Delta} - f(x_{n-1})\right)^2\right]. \end{aligned} \tag{14}$$

It now seems tempting to take the "limit" $\Delta \to 0$ here. In fact, assuming that the $x_1, \ldots, x_n$ in Equation (14) are the values of some reference trajectory $\{z_t, t \in I\}$ at the points $t_1, \ldots, t_n$, we would by formally taking this limit indeed obtain Equation (9) for the density with the energy functional as in Equation (10).

If we retrace the steps in our calculation though, we realise that we have not quite taken them in the order we should according to Definition (8) of the density. To discuss this, we introduce the *ϵ–weight* of a trajectory $\{z_t, t \in I\}$, but now with respect to the Euler approximation:

$$\alpha_\Delta(\epsilon, \{z_t\}) = \frac{\mathbb{P}(\sup_n |X^{(\Delta)}_{t_n} - z_{t_n}| \le \epsilon)}{\mathbb{P}(\sup_n \rho|W_{t_n}| \le \epsilon)}.$$

What we have done to arrive at the Equations (9,10) for the density is to take the limit $\epsilon \to 0$, then use Equation (6) in the special case of the Euler system (13), and finally take the limit $\Delta \to 0$. That is, we have proved

$$\lim_{\Delta\to 0}\lim_{\epsilon\to 0} \alpha_\Delta(\epsilon, \{z_t\}) = \exp(-\mathcal{A}_E(\{z_t\})). \tag{15}$$

However, Definition (8) basically requires to take these limits the other way round:

$$p(\{z_t, t \in I\}) = \lim_{\epsilon\to 0}\lim_{\Delta\to 0} \alpha_\Delta(\epsilon, \{z_t\}). \tag{16}$$

A simple example (following Dutra et al. (2014)) will show that interchanging these two limits will, in general, give different results. It is evident that the density should be independent of what scheme we use to approximate solutions of SDE's, and the Euler scheme is not the only scheme. To arrive at another scheme for numerically solving SDE's, we consider other approximations of the integral in Equation (12), for instance

$$\int_{t_{n-1}}^{t_n} f(X_s)\mathrm{d}s \cong (\lambda f(X_{t_{n-1}}) + (1-\lambda) f(X_{t_n}))\Delta$$

for some $\lambda \in [0, 1]$, leading to the implicit scheme

$$X_{t_n}^{(\Delta)} = X_{t_{n-1}}^{(\Delta)} + (\lambda f(X_{t_{n-1}}^{(\Delta)}) + (1-\lambda) f(X_{t_n}^{(\Delta)}))\Delta + \rho \cdot (W_{t_n} - W_{t_{n-1}}). \tag{17}$$

This is an equally valid approximation scheme for SDE's, see for instance Kloeden and Platen (1992), Chapter 12. Note however that $X_{t_n}^{(\Delta)}$ is now a nonlinear function of the noise $(W_{t_n} - W_{t_{n-1}})$. Using the same logic as before (see Appendix B) one arrives at the conclusion that the functional $\mathcal{A}$ in Equation (9) of the density should be

$$\mathcal{A}_\lambda = \exp\left[-\frac{1}{2\rho^2}\int_I (\dot{z}_t - f(z_t))^2 \,\mathrm{d}t - (1-\lambda)\int_I f'(z_t)\,\mathrm{d}t\right]. \tag{18}$$

So not only does another term $-(1-\lambda)\int_I f'(z_t)\,\mathrm{d}t$ appear in the exponent, but we can generate an entire spectrum of candidate functionals by varying λ. This result evidently draws the entire methodology into question.

We note that $\lambda = 1/2$ gives the Onsager–Machlup functional, that is, $\mathcal{A}_{1/2} = \mathcal{A}_{OM}$. This however does not prove that $\mathcal{A}_{OM}$ is indeed the correct functional. So far, we do not have any reason to believe that $\lambda = 1/2$ is in any way special.

4 Numerical experiment

It was already mentioned in the last section (and will be discussed further in the Appendix) that $\mathcal{A}_{OM}$ is the appropriate density functional for paths of a stochastic differential equation. In particular, this implies that the minimiser of $\mathcal{A}_{OM}$ can be interpreted as the MAP estimator or "most probable" path of the stochastic differential equation. We have also considered discrete time approximations to the stochastic differential equations, for instance the Euler scheme, and it emerged that the densities derived from these approximations do not, in general, agree with the

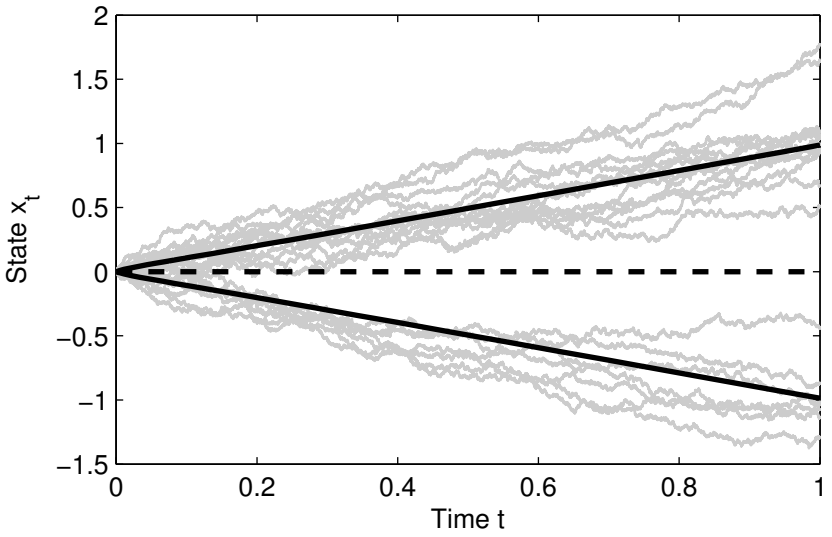

Fig. 2 Around 20 simulations of the SDE (7) with $f(x) = \frac{2}{\pi}\arctan(ax)$ and $a = 6$ and $\rho = 0.3$ are shown in grey, obtained with an Euler scheme with $\Delta = 1.14 \cdot 10^{-4}$. The two solid lines represent the most probable trajectories according to the Onsager–Machlup functional, and the dashed line represents the most probable trajectory according to the energy functional. It is evident that simulations are more likely to accumulate around the former.

Onsager–Machlup functional even approximately. This raises questions as to what the right functional should be *in practice*, since apart from the rare situation where explicit solutions are available, stochastic differential equations inevitably have to be approximated by numerical schemes which are discrete in time. But suppose we approximate a stochastic differential equation of the form (7) with the Euler scheme (13). We know that in this situation, Equation (14) is the *correct* density of solutions, so what is the link between solutions of the Euler scheme and the functional $\mathcal{A}_{OM}$, and why should we care about it?

We will examine the situation with a numerical example. We consider a stochastic differential equation of the form (7) with approximation by the Euler scheme (Equ. 13). Here $f(x) = \frac{2}{\pi}\arctan(ax)$, with $a = 6$ and $\rho = 0.3$. All solutions start from the fixed initial condition $\xi = 0$. Figure 2 shows 20 independent approximate solutions of Equation (7); "approximate" because these are solutions of the Euler scheme (13). The density of these solutions is given by Equation (14), and according to this expression the most probable solution is equal to zero for all times. The picture though we see in Figure 2 seems to contradict this. It is evident that very few solutions seem to be concentrating around the abscissa. This is easy to understand qualitatively. For small times, the variability of the solution grows exponentially as the origin is an unstable fixed point for this dynamics. Sooner or later, the solution will enter regions where the arctan is flat and the drift is essentially either $+1$ or -1. The solution might from time to time transit between these two regimes, but these transits become progressively rarer until it behaves essentially like a random walk with constant drift.

The solid lines in Figure 2 represent the optimal paths of the Onsager–Machlup functional $\mathcal{A}_{OM}$. These have been calculated numerically by solving the Euler Lagrange equations associated with the Onsager–Machlup functional $\mathcal{A}_{OM}$ (the functional displays a symmetry whence there are two solutions symmetric about the abscissa). These solutions seem to capture much better the "big picture", indicating where solutions of our simulations tend to be. So it seems that the Onsager–Machlup functional provides a better description of the density, even though the solutions have been obtained with a discrete time system and thus strictly speaking Equation (14) provides the correct density.

To resolve this apparent paradox, we remember that the density at some reference path $\{z_t\}$ is the probability that the solution of our dynamics lies in a thin sausage of width ϵ around that reference path, relative to the probability that the driving Wiener process lies in a thin sausage of width ϵ around zero. These probabilities, or rather the ϵ–weight $\alpha_\Delta(\epsilon, \{z_t\})$ can be estimated using a Monte Carlo approach in order to study the dependence on ϵ and Δ. For simplicity, the reference path was taken to be zero. Note that this is the most probable path according to $\mathcal{A}_E$. In Figure 3, $\alpha_\Delta(\epsilon, \{z_t\})$ is shown as a function of ϵ (on the abscissa), with different curves (different marker symbols) corresponding to different values of Δ (curves corresponding to smaller values of Δ tend to be more to the left in the plot). Two time windows of different length were considered; the solid lines correspond to $T = 0.2$, while the dashed lines correspond to an experiment with $T = 0.4$.

The discussion in Section 3 revealed that taking the limits $\lim_{\epsilon\to 0}$ and $\lim_{\Delta\to 0}$ of $\alpha_\Delta(\epsilon, \{z_t\})$ in different order gives different results, see Equations (15,16). Along the particular path considered here, $\mathcal{A}_E = 0$ (independent of the value of Δ), meaning that

$$\lim_{\Delta\to 0}\lim_{\epsilon\to 0} \alpha_\Delta(\epsilon, \{z_t\}) = \exp(-\mathcal{A}_E) = 1,$$

while interchanging these limits gives the values $\exp(-\mathcal{A}_{OM}) = 0.68$ for $T = 0.2$ and $\exp(-\mathcal{A}_{OM}) = 0.47$ for $T = 0.4$ (obtained by simply evaluating $\mathcal{A}_{OM}$ along our reference path). The fact that interchanging the limits gives different results manifests itself in the plot in Figure 3 which shows an interesting crossover behaviour. With ϵ decreasing, α first approaches the value given by the Onsager–Machlup functional. If ϵ reaches a sufficiently small value though (depending on Δ), the curves start to diverge from this value and approach one, consistent with the energy functional. The smaller Δ, the longer α stays close to the Onsager–Machlup value for decreasing ϵ, or in other words, for smaller Δ a smaller ϵ has to be chosen in order for $\exp(-\mathcal{A}_E)$ to become a relevant approximation for α.

For a rough estimate on how small ϵ has to be in order for the crossover to take place, we observe that for a reference path z,

$$X^{(\Delta)}_{t_{n+1}} - z_{t_{n+1}} = X^{(\Delta)}_{t_n} - z_{t_n} + \left(f(X^{(\Delta)}_{t_n}) - \frac{z_{t_{n+1}} - z_{t_n}}{\Delta} \right) \Delta + \rho(W_{t_{n+1}} - W_{t_n}). \quad (19)$$

Hence for fixed Δ, the increments of $X^{(\Delta)}_{t_n} - z_{t_n}$ in Equation (19) have a characteristic size (at time t_n), namely

$$|X^{(\Delta)}_{t_{n+1}} - z_{t_{n+1}} - (X^{(\Delta)}_{t_n} - z_{t_n})| \cong |f(z_{t_n}) - \dot{z}_{t_n}|\Delta + \rho\sqrt{\Delta},$$

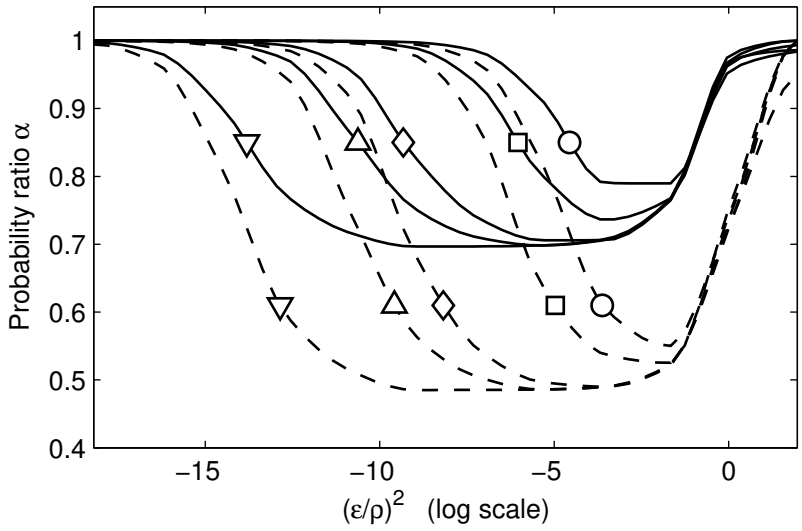

Fig. 3 The ϵ–weight $\alpha(\epsilon, \Delta)$ as a function of ϵ for several values of Δ. The abscissa shows $\log((\frac{\epsilon}{\rho})^2)$. (See text for the reason for this scaling.) The values for $\log(\Delta)$ are -13.7 ($\bigtriangledown$), -10.5 ($\bigtriangleup$), -9.1 ($\Diamond$), -5.9 ($\square$), -4.5 ($\bigcirc$). The solid lines represent results for a shorter time window $T = 0.2$, while the dashed lines represent results for $T = 0.4$.

It seems plausible that α starts to approach the energy functional as soon as ϵ becomes smaller than the typical increment of $X_{t_n}^{(\Delta)} - z_{t_n}$, which means

$$\epsilon \cong |f(z_{t_n}) - \dot{z}_{t_n}|\Delta + \rho\sqrt{\Delta},$$

which is just $\epsilon \cong \rho\sqrt{\Delta}$ in our case, or $\log((\frac{\epsilon}{\rho})^2) = \log(\Delta)$. For the experiments shown in Figure 3, we used the following values of $\log(\Delta)$: -13.7 ($\bigtriangledown$), -10.5 ($\bigtriangleup$), -9.1 ($\Diamond$), -5.9 ($\square$), -4.5 ($\bigcirc$). This appears to be roughly consistent with the values of of $\log((\frac{\epsilon}{\rho})^2)$ at which the crossover takes place.

5 The Onsager–Machlup functional in higher dimensions and for multiplicative noise

In this section we will provide additional (and well known) results regarding the Onsager–Machlup functional in higher dimensions and with multiplicative noise. We will see that in the case of multiplicative noise, further terms appear in the Onsager–Machlup functional; the effect of these terms in data assimilation applications remains to be investigated. We consider a general SDE

$$\dot{X}_t = f(X_t) + \rho(X_t) \cdot r_t, \qquad t \in [0, T] \tag{20}$$

where the state space E is the d–dimensional Euclidean space, f is a vector field on E and ρ is a state dependent d–by–d matrix. For SDE's with multiplicative noise as in Equation (20), different mathematical interpretations are possible, most prominently the Itô and the Stratonovič interpretation (see e.g. Øksendal 1998; Ikeda

and Watanabe 1989). We will interprete the SDE (20) in the sense of Stratonovič; and Itô equation can always be converted to a Stratonovič equation. The expression for the Onsager–Machlup functional given in Equation (22) below is valid if the noise is nondegenerate, that is $\rho(x)\rho^T(x) \geq \alpha \mathbb{1}$ for some $\alpha > 0$. In this situation, the matrix $g(x) = (\rho(x)\rho^T(x))^{-1}$ defines a Riemannian metric. For any vector field f, the divergence $\mathrm{div} f$ will be understood with respect to this metric, that is

$$\mathrm{div} f = \frac{1}{\sqrt{|g|}} \sum_{k=1}^{d} \partial_k (\sqrt{|g|} f^{(k)})$$

Further, let $R(x)$ be the scalar (Ricci) curvature and $m(x, y)$ the (geodesic) distance between points $x, y \in E$. These concepts are defined with respect to the metric g as well (see Gallot et al. 2004, for an introduction to Riemannian geometry). Then the Onsager–Machlup functional is defined as

$$\exp(-\mathcal{A}_{\mathrm{OM}}(\{z_t\})) = \lim_{\epsilon \to 0} \frac{\mathbb{P}(\sup_{t \in I} m(X_t, z_t) \leq \epsilon)}{\mathbb{P}(\sup_{t \in I} |W_t| \leq \epsilon)}, \tag{21}$$

and as is proved for instance in Ikeda and Watanabe (1989); Zeitouni and Dembo (1987), it has the expression

$$\begin{aligned} \mathcal{A}_{\mathrm{OM}}(\{z_t\}) = \frac{1}{2} \int_I (\dot{z}_t - f(z_t))^T g(z_t)(\dot{z}_t - f(z_t)) \, \mathrm{d}t \\ + \frac{1}{2} \int_I \mathrm{div} f(z_t) \mathrm{d}t - \frac{1}{12} \int_I R(z_t) \, \mathrm{d}t. \end{aligned} \tag{22}$$

As was already discussed in Section 4 (in the context of a one–dimensional example), the effect of the second term (containing $\mathrm{div} f$) is to discourage the most probable path from staying in regions where the dynamics is unstable, as this causes strong amplification of the noise and thus typical solutions of the SDE quickly escape from such regions. The effect of the second term involving the Ricci curvature is not so clear at this point and is subject to future investigation.

In the remainder of this section we discuss what terms need adding to the Onsager–Machlup functional if observations are present. Let the observations be a discrete time series $\{Y_n, n = 1, \ldots, N\}$. The Onsager–Machlup functional is now defined as

$$\exp(-\mathcal{F}_{\mathrm{OM}}(\{z_t\}, \{Y_n\})) = \lim_{\epsilon \to 0} \frac{\mathbb{P}(\sup_{t \in I} m(X_t, z_t) \leq \epsilon | \{Y_n, n = 1, \ldots, N\})}{\mathbb{P}(\sup_{t \in I} |W_t| \leq \epsilon)}.$$

(We will use the notation $\mathcal{F}_{\mathrm{OM}}$ to designate the Onsager–Machlup functional with observations; $\mathcal{A}_{\mathrm{OM}}$ still defined as in Eq. 21.) A commonly made assumption is that the observations are conditionally independent given the underlying trajectory $\{X_t, t \in [0, T]\}$, and that the distribution of Y_n depends on X_{t_n} only for $n = 1, \ldots, N$ and a series of sampling times $t_1, \ldots, t_N$. Let $q_n(y, x)$ be the density of Y_n given X_{t_n}. In this case, the full Onsager–Machlup functional reads as

$$\mathcal{F}_{\mathrm{OM}}(\{z_t\}; \{Y_n\}) = \mathcal{A}_{\mathrm{OM}}(\{z_t\}) - \sum_{n=1}^{N} \log(q_n(Y_n, z_{t_n}))$$

If for instance Y_n given X_{t_n} is Gaussian with mean $h(X_{t_n})$ and variance γ (where h and γ are often called the observation function and observation error covariance,

respectively), then the additional term in the Onsager–Machlup functional reads as

$$-\sum_{n=1}^{N} \log(q_n(Y_n, z_{t_n})) = \frac{1}{2}\sum_{n=1}^{N} (y_n - h(z_{t_n}))^T \gamma^{-1} (y_n - h(z_{t_n})).$$

6 Conclusions for discrete time simulations and data assimilation

When modelling a dynamical process with a stochastic differential equation, then any practical implementation will use a discrete time approximation of one form or another. If (as part of a data assimilation experiment for instance) one is interested in a most probable path of that dynamical process, then our considerations imply that the appropriate functional is the Onsager–Machlup functional (or a discrete time approximation thereof), even though the density of discrete time approximations might differ from the Onsager–Machlup functional. The Onsager–Machlup functional provides results which are robust with respect to the particular approximation scheme, and in particular with respect to the chosen time discretisation, which does not have any intrinsic meaning in terms of the problem specification. More specifically, the Onsager–Machlup functional gives approximately the ϵ-weight of a reference path, that is the probability that the solutions of the stochastic differential equation stay in an ϵ sausage around the reference path, and a discrete time approximation of the SDE will assign approximately the same ϵ-weight to that path, *unless* ϵ reaches the scale of typical increments in that approximation. In other words, the Onsager–Machlup functional provides an approximation to the ϵ-weight of a path with respect to the stochastic differential equation *and* approximations thereof, save approximations that employ increments which are typically larger than ϵ. Such approximations do not appropriately represent the fast fluctuations of the Wiener process that are still relevant for the dynamics, even when the amplitude of Wiener process is constrained to be small.

For these reasons, most probable paths should be determined using the Onsager–Machlup functional, since such paths carry the largest possible ϵ-weight, no matter if this weight is calculated from the stochastic differential equation or any reasonable approximation, as long as that approximation uses increments which are smaller than ϵ. Paths which are minimisers of the energy functional or any other functional do not possess this universality property. The implication for data assimilation is that minimising paths of the Onsager–Machlup functional are more typical for the dynamics and in fact carry a rigorous interpretation as MAP estimators, different from maximum energy paths which do not.

These arguments do not apply though if the process under consideration is intrinsically discrete in time. In this situation, it does not make sense to consider the limit $\Delta \to 0$ which brings about the extra term in the Onsager–Machlup functional. Systems like this might appear in the context of seasonal or diurnal cycles, or more generally systems with an internal clocking mechanism.

Appendix

A Derivation of the correct functional

We will attempt a more careful calculation of the ϵ–weight of a path which will not only allow us to take the limits in the right order and obtain the correct expression for the density, but also to identify the reason why interchanging these limits gives a different result. We will later restrict our attention to linear dynamics. It should be said that for linear dynamics, the additional term in the Onsager–Machlup functional (11) does not depend on the reference trajectory and hence minimising $\mathcal{A}_{OM}$ or $\mathcal{A}_E$ gives the same results in this case. However, the functionals are still different and only the Onsager–Machlup functional provides the correct density.

First we note the following simple but important fact. Let $Z^{(1)}, Z^{(2)}$ be random variables with values in $\mathbb{R}^N$ with densities p_1, p_2 respectively, and $p_2(z) > 0$ for all $z \in \mathbb{R}^N$. Further, let ϕ be a function on $\mathbb{R}^N$. Then the identity

$$\mathbb{E}(\phi(Z^{(1)})) = \mathbb{E}(\phi(Z^{(2)})\frac{p_1(Z^{(2)})}{p_2(Z^{(2)})})$$

holds, since

$$\mathbb{E}(\phi(Z^{(1)})) = \int_{\mathbb{R}^n} \phi(z)p_1(z)\mathrm{d}z = \int_{\mathbb{R}^n} \phi(z)\frac{p_1(z)}{p_2(z)}p_2(z)\mathrm{d}z = \mathbb{E}(\phi(Z^{(2)})\frac{p_1(Z^{(2)})}{p_2(Z^{(2)})}). \tag{23}$$

On the other hand, note that

$$\mathbb{P}(\max_k |X_{t_k} - z_{t_k}| \leq \epsilon) = \mathbb{E}(H(\frac{\max_k |X_{t_k} - z_{t_k}|}{\epsilon} - 1)), \tag{24}$$

where H is the Heaviside function. We might use Equation (23) in (24) with

$$\begin{aligned}
\phi(z) &= H(\frac{\max_k |z_k|}{\epsilon} - 1),\\
Z^{(1)} &= (X^{(\Delta)}_{t_1} - z_{t_1}, \ldots, X^{(\Delta)}_{t_N} - z_{t_N}),\\
Z^{(2)} &= (W_{t_1}, \ldots, W_{t_N}),
\end{aligned}$$

where $(X^{(\Delta)}_{t_1}, \ldots, X^{(\Delta)}_{t_N})$ is a solution to the Euler approximation (13). Note that $(X^{(\Delta)}_{t_1} - z_{t_1}, \ldots, X^{(\Delta)}_{t_N} - z_{t_N})$ is then a solution of the system (19). We therefore obtain

$$\mathbb{P}(\max_k |X_{t_k} - z_{t_k}| \leq \epsilon) = \mathbb{E}\Big[H(\frac{\max_k |W_{t_k}|}{\epsilon} - 1)\exp(A + B + C)\Big] \tag{25}$$

with

$$\begin{aligned}
A &= -\frac{\Delta}{2\rho^2}\sum_{n=1}^{N}(\frac{z_{t_n} - z_{t_{n-1}}}{\Delta} - f(z_{t_{n-1}} + \rho W_{t_{n-1}}))^2\\
B &= -\frac{1}{\rho}\sum_{n=1}^{N}(\frac{z_{t_n} - z_{t_{n-1}}}{\Delta})(W_{t_n} - W_{t_{n-1}})\\
C &= \frac{1}{\rho}\sum_{n=1}^{N}(f(z_{t_{n-1}} + \rho W_{t_{n-1}}))(W_{t_n} - W_{t_{n-1}})
\end{aligned} \tag{26}$$

In terms of the limits $\Delta \to 0$ and $\epsilon \to 0$, the first two terms A and B will converge to

$$\lim_{\epsilon\to 0}\lim_{\Delta\to 0} A = -\frac{1}{2\rho^2}\int_0^T (\dot{z_t} - f(z_t))^2\mathrm{d}t \tag{27}$$

and zero, respectively, no matter in which order the limits are taken. The third term however shows different behaviour depending on whether $\Delta \to 0$ or $\epsilon \to 0$ first. If we take $\Delta \to 0$ first, it can be shown that a well defined random variable obtains[3] which can be written as an Ito integral

$$\lim_{\Delta\to 0} C = \frac{1}{\rho}\int_0^T f(z_t + \rho W_t)\mathrm{d}W_t.$$

We do not expect the reader to be familiar with the theory of Ito integrals – relevant here is that the limit of this expression for $\epsilon \to 0$ will *not* be zero but

$$\lim_{\epsilon\to 0}\lim_{\Delta\to 0} C = -\frac{1}{2}\int_0^T f'(z_t)\mathrm{d}t. \tag{28}$$

A demonstration of this fact[4] for the case where f is linear is given here for illustration. If $f(x) = ax$ for some $a \in \mathbb{R}$, then

$$\begin{aligned} C &= \frac{a}{\rho}\sum_{n=1}^{N}(z_{t_{n-1}} + \rho W_{t_{n-1}})(W_{t_n} - W_{t_{n-1}}) \\ &= \frac{a}{\rho}\sum_{n=1}^{N} z_{t_{n-1}}(W_{t_n} - W_{t_{n-1}}) + a\sum_{n=1}^{N} W_{t_{n-1}}(W_{t_n} - W_{t_{n-1}}) \\ &= \frac{a}{\rho}C_1 + aC_2. \end{aligned} \tag{29}$$

It is easy to see that $C_1 \to 0$ if $\Delta \to 0$ and $\epsilon \to 0$, no matter in which order these limits are taken. After some algebra, we can write C_2 as

$$\begin{aligned} C_2 &= \sum_{n=1}^{N} W_{t_{n-1}}(W_{t_n} - W_{t_{n-1}}) \\ &= \frac{1}{2}W_T^2 - \frac{1}{2}\sum_{n=1}^{N}(W_{t_n} - W_{t_{n-1}})^2 \end{aligned}$$

Considering the mean and the variance of the second term, we obtain $\frac{1}{2}T$ and $\frac{1}{2}T\Delta$, respectively, implying that (at least in a mean square sense) the second term converges to its mean $\frac{1}{2}T$ if $\Delta \to 0$. Hence

$$\lim_{\Delta\to 0} C_2 = \frac{1}{2}W_T^2 - \frac{1}{2}T \tag{30}$$

Therefore, taking the limits $\Delta \to 0$ and then $\epsilon \to 0$ in Equation (29) and using Equation (30) we obtain

$$\lim_{\epsilon\to 0}\lim_{\Delta\to 0} C = -\frac{a}{2}T$$

which is the same as Equation (28) for this special case.

Using Equation (28) and the expression in Equation (27) in (25) we obtain that for small ϵ

$$\begin{aligned} \mathbb{P}(\sup_t |X_t - z_t| \le \epsilon) \cong \mathbb{E}(H(&\frac{\sup_t |W_t|}{\epsilon} - 1) \\ &\cdot \exp\Big[-\frac{1}{2\rho^2}\int_0^T (\dot{z}_t - f(z_t))^2\mathrm{d}t - \frac{1}{2}\int_0^T f'(z_t)\mathrm{d}t.\Big] \end{aligned}$$

[3] The limit is in fact in the L_2 sense.

[4] Strictly speaking this "fact" is only correct in a much weaker sense but still sufficient to derive the Onsager–Machlup functional; The correct statement is that

$$\mathbb{E}\Big[\exp\Big(\int_0^T f(z_t + W_t)\mathrm{d}W_t + \frac{1}{2}\int_0^T f'(z_t)\mathrm{d}t\Big)\,\Big|\,\sup_t |W_t| \le \epsilon\Big] \to 1$$

for $\epsilon \to 0$, see Ikeda and Watanabe (1989).

so that we can conclude

$$\begin{aligned}
&\lim_{\epsilon\to 0} \frac{\mathbb{P}(\sup_t |X_t - z_t| \le \epsilon)}{\mathbb{P}(\sup_t |W_t| \le \epsilon)} \\
&= \exp\Big[-\frac{1}{2\rho^2}\int_0^T (\dot{z}_t - f(z_t))^2 \mathrm{d}t - \frac{1}{2}\int_0^T f'(z_t)\mathrm{d}t.\Big] \\
&= \exp(-\mathcal{A}_{OM}).
\end{aligned}$$

Note that if we used Equations (25,26) as a starting point but subsequently took the limits in the wrong order, that is, first $\epsilon \to 0$ and then $\Delta \to 0$, we would have $B, C \to 0$, so we would obtain the energy estimator $\mathcal{A}_E$.

As a final remark, by looking back at the calculations the reader will see that the only term that does not permit interchange of the limits is a second order or "quadratic" term $\sum_{n=1}^N (W_{t_n} - W_{t_{n-1}})^2$ which would vanish with $\Delta \to 0$ if W were a differentiable function but converges to T in case of the Wiener process. Roughly speaking, this is because $W_{t_n} - W_{t_{n-1}}$ is of order $\sqrt{\Delta}$, which more generally gives rise to the extra terms in the Ito calculus.

B Derivation of Equation (18)

In this section, we will derive the Equation (18), that is, we follow same steps as for the Euler scheme and take the limits as in Equation (15), but starting with the implicit scheme (17) instead of the Euler scheme. If we set $R_n = \rho(W_{t_n} - W_{t_{n-1}})$, then the implicit scheme (17) can be written in the form

$$X_{t_n} = X_{t_{n-1}} + F_1(X_{t_n}) + F_2(X_{t_{n-1}}) + R_n$$

which can be expressed as $(R_1, \dots, R_N) = \Psi(X_{t_1}, \dots, X_{t_N})$ with

$$\Psi_n(x_1, \dots, x_N) = x_n - x_{n-1} - F_1(x_n) - F_2(x_{n-1}) \qquad \text{for } n = 1, \dots, N.$$

According to basic probability calculus, we have for the densities

$$p_{X_{t_1},\dots,X_{t_N}}(x_1, \dots, x_N) = p_{R_1,\dots,R_N}(\Psi(x_1, \dots, x_N)) \cdot \left|\frac{\partial \Psi}{\partial x}(x_1, \dots, x_N)\right| \tag{31}$$

Since $\frac{\partial \Psi_k}{\partial x_l} = 0$ for $k < l$, the Jacobi matrix of Ψ is lower left triangular and hence

$$\begin{aligned}
\left|\frac{\partial \Psi}{\partial x}\right|(x_1, \dots, x_N) &= \prod_{n=1}^N \frac{\partial \Psi_k}{\partial x_k}(x_1, \dots, x_N) \\
&= \prod_{n=1}^N 1 - F_1'(x_k) \\
&= \prod_{n=1}^N 1 - (1-\lambda)\Delta f'(x_k) \\
&= \exp\left(\sum_{n=1}^N \log(1 - (1-\lambda)\Delta f'(x_k))\right).
\end{aligned}$$

We evaluate this expression with $x_k = z_{t_k}$ for $k = 1, \dots, N$ where $\{z_t\}$ is some trajectory on the interval $I = [0, T]$ and $N = T/\Delta$. Since $\log(1 + w) \cong w$ for small w, we can write the exponent approximately as

$$\sum_{n=1}^N \log(1 - (1-\lambda)\Delta f'(z_{t_n})) \cong -(1-\lambda)\Delta \sum_{n=1}^N f'(z_{t_n})$$

which is a Riemann sum converging to $-(1-\lambda)\int_I f'(z_t)\,\mathrm{d}t$. The first factor in Equation (31), after normalisation and when evaluated along a trajectory, reads as

$$\frac{p_{R_1,\ldots,R_N}(\Psi(z_{t_1},\ldots,z_{t_N}))}{p_{R_1,\ldots,R_N}(0,\ldots,0)} = \exp\left(-\frac{\Delta}{2\rho^2}\sum_{n=1}^{N}\left(\frac{z_{t_n}-z_{t_{n-1}}}{\Delta}-(1-\lambda)f(z_{t_n})-\lambda f(z_{t_{n-1}})\right)^2\right).$$

Again, the exponent is a Riemann sum which converges to $-\frac{1}{2\rho^2}\int_I(\dot{z}_t-f(z_t))^2\mathrm{d}t$ for $\Delta\to 0$. In summary, we get Equation (18).

References

A. Apte, M. Hairer, A.M. Stuart, and J. Voss. Sampling the posterior: An approach to non-gaussian data assimilation. *Physica D: Nonlinear Phenomena*, 230(1):50 – 64, 2007. ISSN 0167-2789. doi: https://doi.org/10.1016/j.physd.2006.06.009. URL `http://www.sciencedirect.com/science/article/pii/S016727890600217X`. Data Assimilation.

Leo Breiman. *Probability*. Addison-Wesley, Reading, Mass, 1973.

S. L. Cotter, M. Dashti, J. C. Robinson, and A. M. Stuart. Bayesian inverse problems for functions and applications to fluid mechanics. *Inverse Problems*, 25(11): 115008, 43, 2009. ISSN 0266-5611. doi: 10.1088/0266-5611/25/11/115008. URL `http://dx.doi.org/10.1088/0266-5611/25/11/115008`.

J.C. Derber. A variational continuous assimilation technique. *Monthly Weather Review*, 117 (11):2437–2446, 1989.

Dimas Abreu Dutra, Bruno Otvio Soares Teixeira, and Luis Antonio Aguirre. Maximum a posteriori state path estimation: Discretization limits and their interpretation. *Automatica*, 50(5):1360 – 1368, 2014. ISSN 0005-1098. doi: http://dx.doi.org/10.1016/j.automatica.2014.03.003. URL `http://www.sciencedirect.com/science/article/pii/S0005109814000958`.

Geir Evensen. *Data Assimilation. The Ensemble Kalman Filter*. Springer-Verlag, New York, 2007.

Christian L. E. Franzke, Terence J. O'Kane, Judith Berner, Paul D. Williams, and Valerio Lucarini. Stochastic climate theory and modeling. *Wiley Interdisciplinary Reviews: Climate Change*, 6(1):63–78, 2015. ISSN 1757-7799. doi: 10.1002/wcc.318. URL `http://dx.doi.org/10.1002/wcc.318`.

Sylvestre Gallot, Dominique Hulin, and Jacques Lafontaine. *Riemannian geometry*. Universitext. Springer-Verlag, Berlin, third edition, 2004. ISBN 3-540-20493-8. URL `https://doi.org/10.1007/978-3-642-18855-8`.

K. Ide, P. Courtier, M. Ghil, and A. C. Lorenc. Unified notation for data assimilation: Operational, sequential and variational. *Journal of the Meteorologcial Society of Japan*, 75(1B): 181–189, 1997.

Nobuyuki Ikeda and Shinzo Watanabe. *Stochastic differential equations and diffusion processes*, volume 24 of *North-Holland Mathematical Library*. North-Holland Publishing Co., Amsterdam, second edition, 1989.

Peter Imkeller and Jin-Song von Storch, editors. *Stochastic climate models*, volume 49 of *Progress in Probability*, 2001. Birkhäuser Verlag, Basel. ISBN 3-7643-6520-X. URL `https://doi.org/10.1007/978-3-0348-8287-3`.

Andrew H. Jazwinski. *Stochastic Processes and Filtering Theory*, volume 64 of *Mathematics in Science and Engineering*. Academic Press, 1970. ISBN 9780123815507.

Eugenia Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, first edition, 2001.

P. E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer Verlag, 1992.

G. N. Milstein. *Numerical integration of stochastic differential equations*, volume 313 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1995. ISBN 0-7923-3213-X. URL `https://doi.org/10.1007/978-94-015-8455-5`. Translated and revised from the 1988 Russian original.

R. E. Mortensen. Maximum-likelihood recursive nonlinear filtering. *Journal of Optimization Theory and Applications*, 2:386–394, 1968.

Peter Mörters and Yuval Peres. *Brownian motion*, volume 30 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2010. ISBN 978-0-521-76018-8. doi: 10.1017/CBO9780511750489. URL http://dx.doi.org/10.1017/CBO9780511750489. With an appendix by Oded Schramm and Wendelin Werner.

Bernt Øksendal. *Stochastic differential equations*. Universitext. Springer-Verlag, Berlin, fifth edition, 1998. ISBN 3-540-63720-6. URL https://doi.org/10.1007/978-3-662-03620-4. An introduction with applications.

A. M. Stuart. Inverse problems: A bayesian perspective. *Acta Numerica*, 19:451ŏ2013559, 2010. doi: 10.1017/S0962492910000061.

Eric Vanden-Eijnden and Jonathan Weare. Data assimilation in the low noise regime with application to the kuroshio. *Monthly Weather Review*, 141(6):1822–1841, 6 2013. ISSN 0027-0644. doi: 10.1175/MWR-D-12-00060.1.

O. Zeitouni and A. Dembo. A maximum a posteriori estimator for trajectories of diffusion processes. *Stochastics*, 20(3):221, 1987.

O. Zeitouni and A. Dembo. An existence theorem and some properties of maximum a posteriori estimators of trajectories of diffusions. *Stochastics*, 23 (2):197, 1988. ISSN 0090-9491. doi: 10.1080/17442508808833490. URL http://www.informaworld.com/10.1080/17442508808833490.