Balanced data assimilation for highly-oscillatory mechanical systems

Maria Reinhardt[†]   Gottfried Hastermann[‡]   Rupert Klein[‡]   Sebastian Reich[†,§]

August 14, 2017

Abstract

Data assimilation algorithms are used to estimate the states of a dynamical system using partial and noisy observations. The ensemble Kalman filter has become a popular data assimilation scheme due to its simplicity and robustness for a wide range of application areas. Nevertheless, the ensemble Kalman filter also has limitations due to its inherent Gaussian and linearity assumptions. These limitations can manifest themselves in dynamically inconsistent state estimates. We investigate this issue in this paper for highly oscillatory Hamiltonian systems with a dynamical behavior which satisfies certain balance relations. We first demonstrate that the standard ensemble Kalman filter can lead to estimates which do not satisfy those balance relations, ultimately leading to filter divergence. We also propose two remedies for this phenomenon in terms of blended time-stepping schemes and ensemble-based penalty methods. The effect of these modifications to the standard ensemble Kalman filter are discussed and demonstrated numerically for two model scenarios. First, we consider balanced motion for highly oscillatory Hamiltonian systems and, second, we investigate thermally embedded highly oscillatory Hamiltonian systems. The first scenario is relevant for applications from meteorology while the second scenario is relevant for applications of data assimilation to molecular dynamics.

Keywords. Data assimilation, ensemble Kalman filter, balanced dynamics, highly oscillatory systems, Hamiltonian dynamics, Langevin dynamics, geophysics, molecular dynamics

AMS(MOS) subject classifications. 65C05, 62M20, 93E11, 62F15, 86A22

1 Introduction

In this paper, we discuss sequential data assimilation techniques for highly-oscillatory Hamiltonian systems with Hamiltonian

$$H^\varepsilon(p,q) = \frac{1}{2}p^{\mathrm{T}}M^{-1}p + \frac{1}{2\varepsilon^2}g(q)^{\mathrm{T}}Kg(q) + V(q), \tag{1}$$

where $M \in \mathbb{R}^{N \times N}$ is the diagonal mass matrix, $V : \mathbb{R}^N \to \mathbb{R}$ is a potential, $g : \mathbb{R}^N \to \mathbb{R}^L$ gives rise to rapid oscillations with diagonal matrix of force constants $K \in \mathbb{R}^{L \times L}$, and $\varepsilon > 0$ is a stiffness parameter. To simplify notation, we assume $M = I$. The associated Hamiltonian equations of motion are then given by

$$\dot{q} = p \tag{2}$$
$$\dot{p} = -\varepsilon^{-2}G(q)^{\mathrm{T}}Kg(q) - \nabla_q V(q), \tag{3}$$

where $G(q) := Dg(q) \in \mathbb{R}^{L \times N}$ denotes the Jacobian matrix of g at q. These equations pose challenges both in terms of their numerical treatment as well as sequential data assimilation techniques in the limit $\varepsilon \to 0$. To clearly identify the dependence of the solution behavior of (2)–(3) on the parameter ε, we denote solutions by $q^\varepsilon(t)$ and $p^\varepsilon(t)$, respectively.

We observe that bounded energy, i.e.

$$H^\varepsilon(p^\varepsilon(t),q^\varepsilon(t)) = \mathcal{O}(\varepsilon^0), \tag{4}$$

implies $g(q^\varepsilon(t)) = \mathcal{O}(\varepsilon)$. In other words, solutions $q^\varepsilon(t)$ have to stay close to the constraint manifold

$$\mathcal{M} = \{q \in \mathbb{R}^N : g(q) = 0\}. \tag{5}$$

[†]Universität Potsdam, Institut für Mathematik, Karl-Liebknecht-Str. 24/25, D-14476 Potsdam, Germany
[‡]Freie Universität Berlin, Institut für Mathematik, Arnimallee 6, D-14195 Berlin, Germany
[§]University of Reading, Department of Mathematics and Statistics, Whiteknights, PO Box 220, Reading RG6 6AX, UK

The rapid oscillations normal to this manifold are characterized by the Hamiltonian

$$H_{\text{osc}}^{\varepsilon}(q,p) = \frac{1}{2}(G(q)p)^{\text{T}}(G(q)G(q)^{\text{T}})^{-1}G(q)p + \frac{1}{2\varepsilon^2}g(q)^{\text{T}}Kg(q) \tag{6}$$

and hence we may conclude that bounded energy implies $G(q^{\varepsilon}(t))p^{\varepsilon}(t) = \mathcal{O}(\varepsilon^0)$.

When describing physical processes by models, there are several sources for uncertainties, such as model errors or an uncertainty about the initial conditions. Data assimilation combines model outputs with observational data (that is also prone to errors) to estimate the probability distribution of the model state conditioned on the observations. There are two main approaches. First, variational data assimilation where a trajectory of the process over a time interval is estimated by solving a minimisation problem, for example 4D-VAR, see e.g. page 186 of [14]. The other approach is sequential Bayesian data assimilation, where one alternates between a forecast step, in which the probability distribution at a certain point in time is evolved according to the model until an observation y_{obs} comes in. The resulting distribution is called the prior (or forecast) distribution. This is followed by an assimilation step in which one takes the observation into account by applying Bayes' theorem and obtains the posterior (or analysis) distribution

$$\pi^{\text{a}}(z|y_{\text{obs}}) \propto \pi(y_{\text{obs}}|z)\pi^{\text{f}}(z). \tag{7}$$

We denote the posterior by π^{a} and the prior by π^{f}. In this paper we focus on the second approach, and how to apply it to a model of the form (2)-(3).

For a Gaussian uncertainty in the initial conditions $z(t_0)$, a linear model $z(t_{n+1}) = z(t_n) + \Delta t\,(Fz(t_n) + b) + \sqrt{2\Delta t}\Xi(t_n)$ and linear observations $y_{\text{obs}}(t_k) = Hz(t_k) + \epsilon$ with Gaussian measurement error $\epsilon \sim \mathcal{N}(0, R)$, the prior and the posterior distributions will always be Gaussian distributions. Hence they are completely characterised by its means and covariances which are given by:

$$\bar{z}^{\text{f}}(t_k) = [I + \Delta tF]^{N_{\text{out}}}z^{\text{a}}(t_{k-1}) + \Delta t\sum_{i=1}^{N_{\text{out}}}[I + \Delta tF]^{i-1}b \tag{8}$$

$$P^{\text{f}}(t_k) = [I + \Delta tF]^{N_{\text{out}}}P^{\text{a}}(t_{k-1})\left([I + \Delta tF]^{N_{\text{out}}}\right)^{\text{T}} + 2\Delta t\sum_{i=1}^{N_{\text{out}}}[I + \Delta tF]^{i-1}Q\left([I + \Delta tF]^{i-1}\right)^{\text{T}} \tag{9}$$

$$\bar{z}^{\text{a}}(t_k) = \bar{z}^{\text{f}}(t_k) - K(H\bar{z}^{\text{f}}(t_k) - y_{\text{obs}}(t_k)) \tag{10}$$

$$P^{\text{a}}(t_k) = P^{\text{f}}(t_k) - KHP^{\text{f}}(t_k) \tag{11}$$

$$\text{with}\quad K := P^{\text{f}}(t_k)H^{\text{T}}(HP^{\text{f}}(t_k)H + R)^{-1}. \tag{12}$$

This procedure is called the *Kalman filter* and was first described by Kalman 1960 [10]. When the model is nonlinear (like in our case) $\bar{z}^{\text{f}}(t_k)$ and $P^{\text{f}}(t_k)$ can be approximated by a Monte Carlo version, the *ensemble Kalman filter* (EnKf) [8]. In that case one samples $z_1(t_0), ..., z_M(t_0)$ from the initial distribution, feeds those initial conditions into the model, obtains an ensemble of model states at a point in time t_k and estimates $\bar{z}^{\text{f}}(t_k) \approx \frac{1}{M}\sum_{i=1}^{M}z_i(t_k) =: \bar{z}_M^{\text{f}}$ and $P^{\text{f}}(t_k) \approx \frac{1}{M-1}\sum_{i=1}^{M}(z_i(t_k) - \bar{z}_M^{\text{f}})(z_i(t_k) - \bar{z}_M^{\text{f}})^T$ by their empirical approximations. The posterior mean and covariance are then obtained via formula (10) and (11) respectively. To go back to the level of samples one has to define a transformation matrix D such that the analysis ensemble

$$z_j^{\text{a}}(t_k) = \sum_{i=1}^{M}z_i^{\text{f}}(t_k)d_{ij}(t_k) \qquad j = 1, \ldots, M \tag{13}$$

satisfies

$$\frac{1}{M}\sum_{i=1}^{M}z_i^{\text{a}}(t_k) = \bar{z}^{\text{a}}(t_k) \tag{14}$$

$$\frac{1}{M-1}\sum_{i=1}^{M}(z_i^{\text{a}} - \bar{z}^{\text{a}})(z_i^{\text{a}} - \bar{z}^{\text{a}})^{\text{T}} = P^{\text{a}}. \tag{15}$$

There are several possibilities to find such a matrix D. In our experiments we use the ensemble square root

filter. Let

$$A := \left[(z_1^f(t_k) - \bar{z}^f(t_k) \cdots (z_M^f(t_k) - \bar{z}^f(t_k)) \right], \tag{16}$$

$$S := I + \frac{1}{M-1}(HA)^{\mathrm{T}} R^{-1} HA, \tag{17}$$

$$w := \frac{1}{M} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} - \frac{1}{M-1} S^2 A^{\mathrm{T}} H^{\mathrm{T}} R^{-1} (H\bar{z}^f - y_{\mathrm{obs}}). \tag{18}$$

With w_i being the i-th entry of the column vector w, the entries of the matrix D are defined as

$$d_{ij} = w_i - \frac{1}{M} + s_{ij}, \tag{19}$$

see pages 211-212 of [14].

Our state variable is $z = (q^{\mathrm{T}}, p^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{2N}$. The Hamiltonian (1) is conserved under the model dynamics (2)–(3), i.e.

$$H^{\varepsilon}(z_i^{\varepsilon\,\mathrm{f}}(t_{q+1})) = H^{\varepsilon}(z_i^{\varepsilon\,\mathrm{a}}(t_q)), \tag{20}$$

but this is not the case for the data assimilation step (13). In particular, one often observes a severe increase in the oscillatory energy (6), i.e.

$$H_{\mathrm{osc}}^{\varepsilon}(z_i^{\varepsilon\,\mathrm{a}}(t_q)) \gg H_{\mathrm{osc}}^{\varepsilon}(z_i^{\varepsilon\,\mathrm{f}}(t_q)), \tag{21}$$

which, in practice, can lead to a destabilization of the simulation after a few data assimilation cycles.

The general problem of data assimilation procedures leading to unphysical predictions and introducing fast oscillations when considering systems with more than one time scale, and how to deal with that issue, has been investigated by several researchers. In [11] it is suggested to apply a digital filter after every assimilation step to filter out spurious fast oscillations and this is applied to a weather prediction model. Strategies that incorporate the observational data in the model evolution in a gradual and smooth way instead of using all the information about the observation at one single point in time have been suggested for example in [4] and [3]. In [7] a highly-oscillatory Hamiltonian form of the shallow water equations is investigated and a method for stabilizing the 4D-VAR technique is suggested, but the model system differs from the one we consider in this paper and sequential Bayesian DA techniques are not taken into account.

The main contribution of this paper is to investigate the phenomenon of sequential filters of the form (13) causing spurious fast oscillations in Hamiltonian models of the form (2)-(3) as well as to propose and test different stabilization techniques which overcome this issue. In this context we will investigate two different scenarios; Scenario A and Scenario B.

A Under this scenario, we will consider solutions $z^{\varepsilon}(t)$ of (2)–(3) with bounded first derivative, i.e. solutions which satisfy

$$\dot{q}^{\varepsilon}(t) = \mathcal{O}(\varepsilon^0) \quad \text{and} \quad \dot{p}^{\varepsilon}(t) = \mathcal{O}(\varepsilon^0). \tag{22}$$

In particular, $\dot{p}^{\varepsilon}(t) = \mathcal{O}(\varepsilon^0)$ together with (3) implies

$$\nabla_q V(q^{\varepsilon}) + \varepsilon^{-2} G(q^{\varepsilon}) K g(q^{\varepsilon}) = \mathcal{O}(\varepsilon^0) \tag{23}$$

and therefore $g(q^{\varepsilon}(t)) = \mathcal{O}(\varepsilon^2)$. Plugging this estimate into (6) leads via equipartitioning of energy to

$$H_{\mathrm{osc}}^{\varepsilon}(z^{\varepsilon}(t)) = \mathcal{O}(\varepsilon^2) \tag{24}$$

and consequently to $G(q^{\varepsilon}(t))p^{\varepsilon}(t) = \mathcal{O}(\varepsilon)$. In other words, for ε sufficiently small, solutions of (2)–(3) with $\dot{z}^{\varepsilon}(t) = \mathcal{O}(\varepsilon^0)$ stay ε-close to the tangent manifold defined by

$$T\mathcal{M} = \{(q,p) \in \mathbb{R}^{2N} : g(q) = 0, \, G(q)p = 0\}, \tag{25}$$

see for example [16] and [5]. Scenario A provides a simple class of toy problems which are representative of the various balanced flow regimes in atmosphere-ocean dynamics [15].

B The second scenario starts from the model (2)–(3) and couples it to a heat bath, giving rise to Langevin dynamics of the form

$$\dot{q} = p \tag{26}$$

$$\dot{p} = -\varepsilon^{-2} G(q)^{\mathrm{T}} K g(q) - \nabla_q V(q) - \gamma p + \sqrt{2\gamma \mathrm{k_B} T} \dot{W}, \tag{27}$$

where $\gamma > 0$ is the friction constant, T the temperature, $\mathrm{k_B}$ the Boltzmann constant, and $W(t)$ standard N-dimensional Brownian motion. Under this setting, the model (26)–(27) has invariant distribution

$$\pi^\varepsilon(q,p) \propto e^{-H^\varepsilon(q,p)/\mathrm{k_B} T} \tag{28}$$

and the energy is finite in the sense of

$$\mathbb{E}[H^\varepsilon(q,p)] = \int_{\mathbb{R}^{2N}} H^\varepsilon(z)\, \pi^\varepsilon(z)\, \mathrm{d}z = \mathcal{O}(\varepsilon^0). \tag{29}$$

This scenario arises from statistical mechanics and in particular from molecular dynamics simulations; see [12, 5, 13].

Our strategy for stabilizing ensemble-based data assimilation is to provide additional information to the data assimilation schemes in the form

$$g(\hat{q}_i^\varepsilon(t_q)) = \mathcal{O}(\varepsilon^2) \quad \text{and} \quad G(\hat{q}_i^\varepsilon(t_k))\hat{p}_i^\varepsilon(t_k)) = \mathcal{O}(\varepsilon) \tag{30}$$

for $i = 1, \dots, M$, in case of Scenario A or

$$g(\hat{q}_i^\varepsilon(t_q)) = \mathcal{O}(\varepsilon) \tag{31}$$

for $i = 1, \dots, M$, in case of Scenario B. These conditions can be enforced in different manners. In the remainder of this paper we will discuss and compare two such approaches for each of the two scenarios.

2 Mathematical background

2.1 Scenario A

In the following we will give a short review on existing results for the Hamiltonian model (2)–(3) subject to initial conditions of the form

$$q_0^\varepsilon := q_0^0 + \varepsilon^2 \tilde{q}_0 \qquad q_0^0 \in \mathcal{M}, \tilde{q}_0 \in \mathbb{R}^N \tag{32}$$

$$p_0^\varepsilon := p_0^0 + \varepsilon \tilde{p}_0 \qquad p_0^0 \in T_{q_0^0}\mathcal{M}, \tilde{p}_0 \in \mathbb{R}^N. \tag{33}$$

This form of the initial conditions guarantees that oscillatory energy (6) satisfies (24) at the initial time.

Formally, one can now apply asymptotic expansions to the solutions and derive the limit system for $\varepsilon \to 0$. More specifically, the main results of [16] state that solutions to the initial value problem

$$\dot{q}^\varepsilon = p^\varepsilon \qquad\qquad q^\varepsilon(0) = q_0^\varepsilon \tag{34}$$

$$\dot{p}^\varepsilon = -\varepsilon^{-2} G(q^\varepsilon)^{\mathrm{T}} K g(q^\varepsilon) - \nabla_q V(q^\varepsilon) \qquad p^\varepsilon(0) = p_0^\varepsilon \tag{35}$$

converge uniformly (on finite time intervals) to solutions (q, p) of the differential algebraic limit system

$$\dot{q} = p \qquad\qquad q(0) = q_0 \in \mathcal{M} \tag{36}$$

$$\dot{p} = -G(q)^{\mathrm{T}} K \lambda - \nabla_q V(q) \qquad p(0) = p_0 \in T_{q_0}\mathcal{M} \tag{37}$$

$$0 = g(q). \tag{38}$$

Hereby $\lambda \in \mathbb{R}^L$ is the corresponding Lagrange multiplier, which is determined by

$$0 = \ddot{g}(q) = -G(q) \left[\nabla V(q) + G(q)^{\mathrm{T}} K \lambda \right] + g_{qq}(q)[p,p]. \tag{39}$$

The zero-order balance relation $(q,p) \in T\mathcal{M}$ can be improved upon by the bounded derivative principle which applied to (2)–(3) states that

$$\ddot{g}(q^\varepsilon(t)) = \mathcal{O}(\varepsilon). \tag{40}$$

We know that

$$\ddot{g}(q^\varepsilon) = G(q^\varepsilon)\dot{p}^\varepsilon + g_{qq}(q^\varepsilon)[p^\varepsilon, p^\varepsilon] \tag{41}$$

$$= G(q^\varepsilon)\left(-\nabla V(q^\varepsilon) - \frac{1}{\varepsilon^2}G(q^\varepsilon)^T K g(q^\varepsilon)\right) + g_{qq}(q^\varepsilon)[p^\varepsilon, p^\varepsilon]. \tag{42}$$

In order for the constraint formulation (36)–(38) to be consistent with the balanced free dynamics, it is therefore necessary that

$$g(q^\varepsilon) = \varepsilon^2 \lambda(q^\varepsilon, p^\varepsilon), \tag{43}$$

where $\lambda(q,p)$ is determined by (39). Replacing (38) by (43) leads to the concept of soft or flexible constraints as introduced in [12, 18].

It should be noted that initial conditions of the form

$$q_0^\varepsilon := q_0^0 + \varepsilon \tilde{q}_0 \qquad q_0^0 \in \mathcal{M}, \tilde{q}_0 \in \mathbb{R}^N \tag{44}$$

$$p_0^\varepsilon := \tilde{p}_0 \qquad \qquad \tilde{p}_0 \in \mathbb{R}^N. \tag{45}$$

lead to an oscillatory energy (6) which is of order $\mathcal{O}(\varepsilon^0)$. In this case, an additional force term appears in the limiting equations (36)–(38). See [16, 17, 5] for more details.

We now describe in more detail the model systems, which we use for numerical experiments.

Example 2.1. The stiff elastic double pendulum corresponds to system (1)–(3) with $q, p \in \mathbb{R}^{N=4}$ and

$$g(q) = \begin{pmatrix} \|q_{1,2}\| - l_1 \\ \|q_{1,2} - q_{3,4}\| - l_2 \end{pmatrix}. \tag{46}$$

This system is a good prototypical example of a simple interacting multi scale model due to its chaotic behaviour. In Figure 1 we show how the positions of a solution with initial conditions of the form (32)–(33) develop in time compared to a solution that has more general initial conditions, and in Figure 2 we show the time evolution of the total energy (1) and the oscillatory energy (6) for both sets of initial conditions. As we do later for the numerical experiments, we used $\varepsilon = 10^{-3}$, $V(q) = 10(q_2 + q_4)$, $K = \text{diag}(1, 0.04)$ to create these figures. In Figure 3 one can see the experimental verification of the theory reviewed before for the balanced case. Here we rerun the balanced solution for a series of decreasing values of ε and monitor values of $g(q)$. As one would expect $g(q)$ is of order ε^2 and $\tilde{g} := g(q) - \varepsilon^2 \lambda(q,p)$ is slightly better preserved under the model dynamics than g, especially for small values of ε.

Since $g \in C(\mathbb{R}^4, \mathbb{R}^2)$, our system is of codimension $N - L = 2$ and the results of [16] are not longer valid. To ensure convergence of the solution as $\varepsilon \to 0$ in the non tangential case, we have to consider the smooth diagonalisation property [17]. To be more specific we have to prove the existence of a smooth basis $(e_1^\perp(q|_\mathcal{M}), \ldots, e_{N-L}^\perp(q|_\mathcal{M}))$ of $T_{q|_\mathcal{M}}\mathcal{M}^\perp$ such that the Hessian

$$\frac{\partial^2 g}{\partial e_i^\perp \partial e_j^\perp}(q|_\mathcal{M}) = \omega_i^2(q|_\mathcal{M})\delta_{i,j} \tag{47}$$

is in diagonal form and the diagonal entries ω_i depend smoothly on $q|_\mathcal{M}$. Furthermore they have to satisfy a non resonance condition given by

$$\omega_i = \omega_j \tag{48}$$
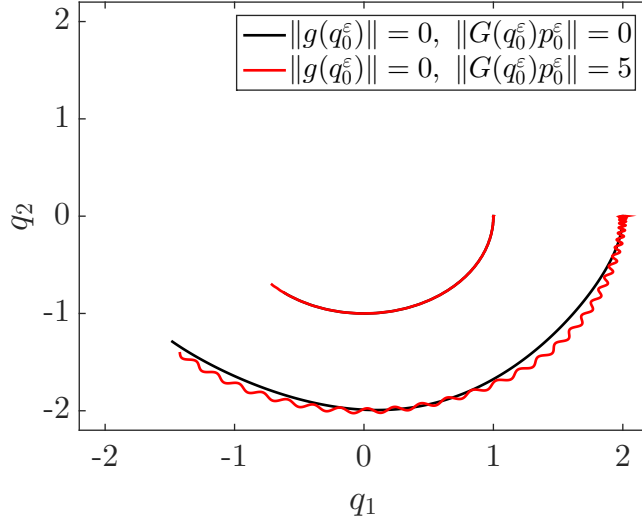
$$\omega_i = \pm\omega_j \pm \omega_k. \tag{49}$$

Figure 1: Evolution in time of a solution with initial conditions of the form (32)–(33) compared to a solution that has more general initial conditions

(a) $H^\varepsilon(q,p)$

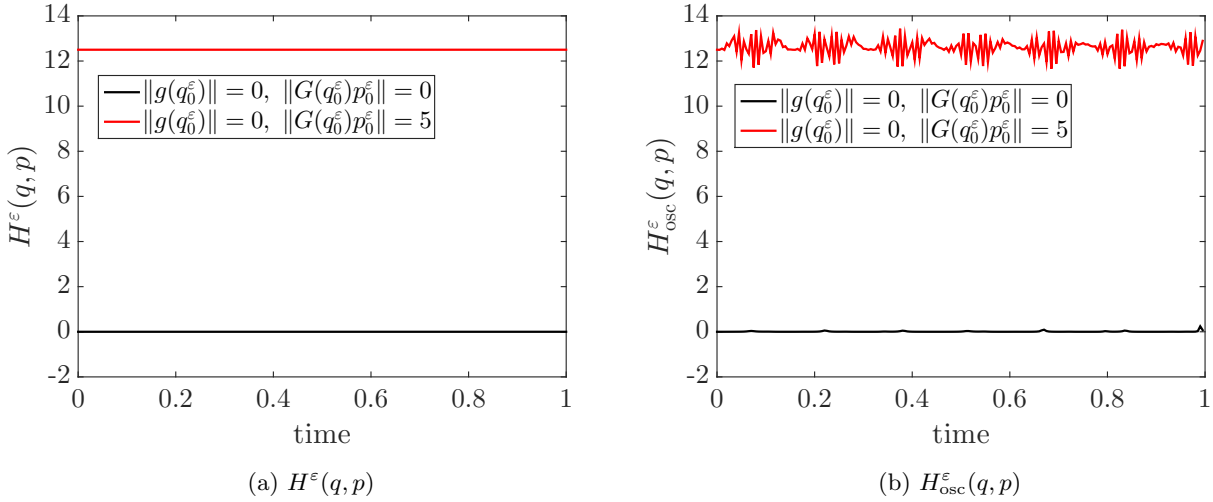(b) $H^\varepsilon_{\mathrm{osc}}(q,p)$

Figure 2: Evolution in time of the total energy and the energy of the fast variable for a solution with initial conditions of the form (32)–(33), shown in black, compared to a solution that has more general initial conditions, shown in red

2.2 Scenario B

The oscillatory energy (6) of thermally embedded systems is of order $\mathcal{O}(\varepsilon^0)$ in expectation and the influence of the rapid oscillations in $g(q^\varepsilon(t))$ on the overall dynamics cannot be ignored. The oscillatory Hamiltonian (6) is approximately harmonic in $x := g(q)$ and $p_x := (G(q)G(q)^{\mathrm{T}})^{-1}G(q)p$ with frequency

$$\omega^\varepsilon(q) = \varepsilon^{-1}(G(q)G(q)^{\mathrm{T}})^{1/2}, \tag{50}$$

where we have assumed $K = I$ for simplicity. We now introduce the action variable

$$J^\varepsilon(q,p) = \frac{H^\varepsilon_{\mathrm{osc}}(q,p)}{\omega^\varepsilon(q)} \tag{51}$$

and recall that the action of a harmonic oscillator with slowly varying frequency is nearly conserved in time, i.e. $J^\varepsilon \approx$ const. This is not the case for thermally embedded systems; instead, one can derive a stochastic
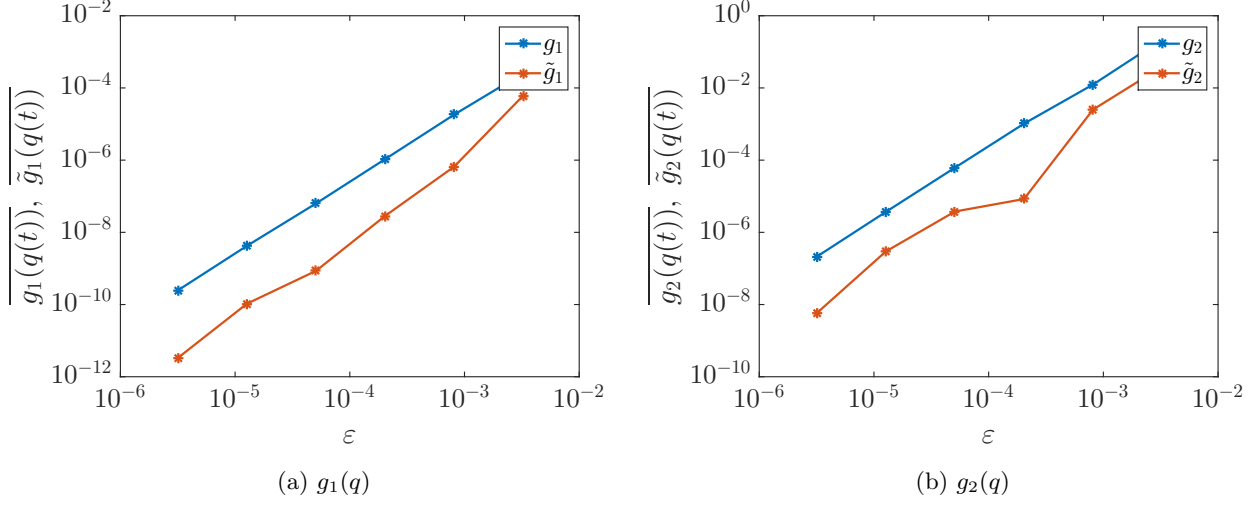
Figure 3: The time averaged values of $g_1(q(t)), g_2(q(t))$ and $\tilde{g}_1(q(t)), \tilde{g}_2(q(t))$ with $\tilde{g}(q) := g(q) - \varepsilon^2 \lambda(q, p)$ for a series of balanced solutions corresponding to decreasing values of the stiffness parameter ε.

differential equation in J^ε, see [13]. In either case one obtains a correcting force term

$$F_{\text{corr}}(q) := -J^\varepsilon(t)\nabla_q \omega^\varepsilon(q), \tag{52}$$

which represents the influence of the harmonic oscillations normal to $T\mathcal{M}$ on the motion of the system on the manifold $T\mathcal{M}$ itself. The existence of (52) poses severe challenges for data assimilation techniques as it requires estimation of $J^\varepsilon(t)$ to predict the slow motion along the constraint manifold $\mathcal{M}$.

Example 2.2. Let us consider a simple example, in which the fast motion can have a substantial influence on the slow one. The *elliptic* elastic pendulum is given by $q, p \in \mathbb{R}^2$ satisfying

$$H(q,p) = \frac{p^{\mathrm{T}}p}{2} + \frac{1}{2\varepsilon^2}g(q)^2 + c_g q_2 \tag{53}$$

$$g(q) = \sqrt{q^{\mathrm{T}}Aq} - 1. \tag{54}$$

$$\tag{55}$$

Hereby $A = \text{diag}(1, \alpha)$ gives the geometry of an ellipse in canonical form with $\alpha > 0$. For the moment we assume no stochastic forcing i.e. $\gamma = 0$. As a first step we introduce a coordinate change into fast and slow coordinates $(x, y) = T(q) := (g(q), b(q))$ where b is chosen as one of

$$b_1(q) = q_1 q_2^{-\alpha} \tag{56}$$

$$b_2(q) = -\alpha q_2 q_1^{-\frac{1}{\alpha}} \tag{57}$$

locally around the q_1 and q_2 axis. One can easily check local orthogonality of the coordinates i.e. $G(q)^{\mathrm{T}}B(q) = 0$ (G and B denote the gradients of g and b), and for the new conjugate momenta

$$\mathrm{D}T(p_x, p_y)^{\mathrm{T}} := p \tag{58}$$

we therefore get

$$p^{\mathrm{T}}p = \omega_1^2 p_x^2 + \omega_2^2 p_y^2 \tag{59}$$

with

$$\omega_1^2 := GG^{\mathrm{T}} = \frac{q^{\mathrm{T}}AAq}{q^{\mathrm{T}}Aq} \tag{60}$$

$$\omega_2^2 := BB^{\mathrm{T}}. \tag{61}$$

We can rewrite our Hamiltonian as

$$H^\varepsilon(x, y, p_x, p_y) = \omega_1(x, y)^2 \frac{p_x^2}{2} + \omega_2(x, y)^2 \frac{p_y^2}{2} + \frac{1}{2\varepsilon^2} x^2 + f(q(x, y)) \tag{62}$$

and by Taylor expansion we obtain

$$H^\varepsilon(x, y, p_x, p_y) = \omega_1(y)^2 \frac{p_x^2}{2} + \frac{1}{2\varepsilon^2} x^2 + h(y, p_y) + x\tilde{f}(x, y, p_x, p_y) \tag{63}$$

where $h(y, p_y)$ represents a slow Hamiltonian energy. Following [13] we use a canonical change of coordinates in rescaled Poisson brackets to conclude the form of the action variable to be

$$J^\varepsilon := \frac{1}{2\omega_1(q)} \left(\omega_1(q)^2 p_x^2 + \frac{g(q)^2}{\varepsilon^2} \right), \quad p_x = (G(q)G(q)^{\mathrm{T}})^{-1} G(q) p \tag{64}$$

This also enables us to find the correction force given by (52) with $\omega := \omega_1$ for the limit system with non tangential initial conditions. We also conclude that there is no additional correction term for the classical elastic pendulum (i.e. $\alpha = 1$), since $\omega_1 = \text{const}$ in this case.

To consider the stochastic forcing i.e. $\gamma \neq 0$ introduced in (26)-(27) one can repeat all coordinate changes as above, but on the level of the equations of motion. Therefore one must split up the noise into x and y components. For a detailed analysis of this case we refer to [13].

In Figure 4 we show the evolution of q_1 and q_2 in the thermally embedded elliptic pendulum with $\varepsilon = 0.001$, $\gamma = 1$, $k_B T = 16$, $\alpha = 36$ and $V(q) = (-1, 0)^{\mathrm{T}}$. The initial conditions are on the slow manifold but with an impuls, that is orthogonal to it. One run is done with the Störmer-Verlet method and the other one with the RATTLE method [1], a numerical integration method which approximates solutions of the constrained (limit) problem (36)–(38). Hereby the numerical solution is as close to $T\mathcal{M}$, c.f. (25), as a given tolerance independent of the time steph length. One can see that the two solutions behave very differently, which is caused by the RATTLE method setting the action variable (51), and therefore the correction force (52), to almost zero whereas the Störmer-Verlet method does not. This can be seen in Figure 5.
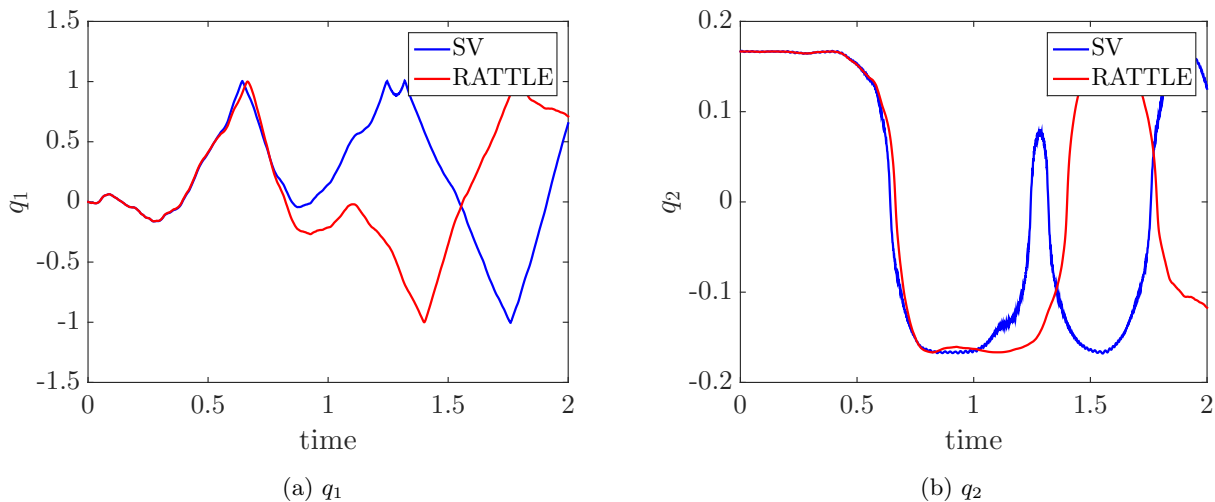

(a) q_1

(b) q_2

Figure 4: Evolution in time of the coordinates of the postion once obtained by the RATTLE and once by the Störmer-Verlet method

3 Balanced data assimilation

In this section, we describe computational methodologies to stabilize ensemble-based data assimilation schemes in the presence of highly oscillatory components in a mechanical system, for each of the two scenarios A and B.
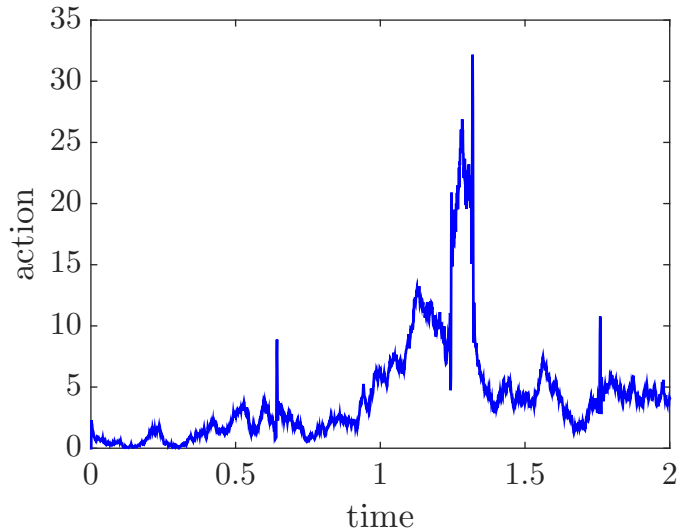
Figure 5: Evolution in time of the action variable of a solution obtained by the Störmer-Verlet method. In contrast to that, the action variable takes the value zero under the RATTLE approximation.

3.1 Scenario A

3.1.1 Blended time-stepping

Motivated by the results of [2], we introduce a numerical time stepping scheme that extends a classical projection approach by subsequent blending steps. These steps continuously blend between the slow limit model (36) - (38) and the full model (34) - (35). This approach was originally developed in the context of incompressible fluid dynamics where the singular perturbation arises by the vanishing Mach number limit Ma $\to$ 0. In addition to the classical projection schemes introduced by [6], much effort was spent on developing asymptotic preserving low Mach number numerical schemes. The essential point is their ability to blend between the (weakly) compressible and the incompressible dynamics without additional stability constraints. As observed in [2], solving the incompressible model immediately after the assimilation and subsequently blending back smoothly to the compressible one over a few time steps can further reduce artificial imbalances caused by data assimilation relative to an approach that simply projects the system state onto the incompressible manifold in one step and then procedes with the compressible model.

To adapt this strategy to our situation we need an asymptotic preserving integrator. We therefore introduce the following numerical scheme. Our ingredients will be a projection method, which keeps the momenta tangential, denoted by ψ^{slow} and an *energy preserving* integrator for the full model, denoted by ψ^{fast}. Using both integration schemes together, our numerical method

$$q_{n+1} = \psi_h^\alpha(q_n) \tag{65}$$

is given by the flow

$$\psi_h^\alpha := \alpha\,\psi_h^{fast} + (1-\alpha)\,\psi_h^{slow}. \tag{66}$$

To choose suitable ψ^{fast} and ψ^{slow} we require

- ψ^{fast} to preserve the energy H (or at least a sufficiently close $\tilde{H}$),

- ψ^{slow} to preserve $G(q)p = 0$ up to floating point precision and

- the stability criterion of ψ^{slow} to be no more restrictive than the one for ψ^{fast}.

Concretely, we choose the classical symplectic Störmer Verlet method as integrator ψ^{fast} and our new method introduced in the appendix A as ψ^{slow}. The blended method is obviously consistent with the slow model for $\alpha = 0$ and with the full model for $\alpha = 1$, but we will introduce a consistency error with respect to the full model

9

by evolving the system with $\alpha \in [0, 1)$. We will argue that this consistency error is not of significant size in the context of data assimilation since even observational errors introduce larger errors in general. More importantly for $\alpha \in (0, 1)$ the solution relaxes to the nearly balanced slow one, which is exactly the behavior we would like to achieve. We therefore propose to use the above blending method as follows. We denote the *blending window* by $k - 1 \geq 1$ and start our forecast at time t_n. Let η be the number of forecast time integration steps, then the following two steps are repeated in every forecast cycle (c.f. Figure 6).

1. **Blending:** Let $\alpha \in \mathbb{R}^k$ such that $0 = \alpha_1 < \alpha_2 \leq \cdots \leq \alpha_{k-1} < \alpha_k = 1$. Integrate until t_{n+k} using

$$q_{n+k} = \left(\Pi_{i=1}^k \psi_h^{\alpha_i} \right)(q_n) \tag{67}$$

2. **Forecast:** Obtain forecast at $t_{n+\eta}$ by
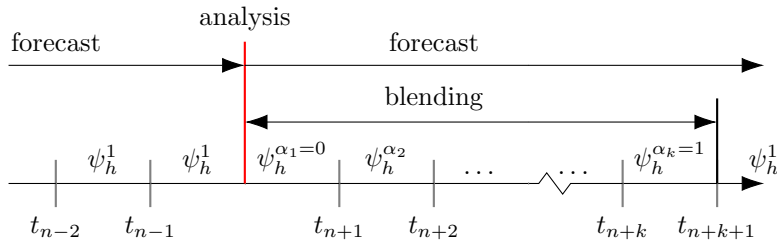
$$q_{n+\eta} = (\psi_h^1)^{\eta-k}(q_{n+k}) \tag{68}$$

Figure 6: Blending strategy applied after analysis with blending window $k - 1$. The numerical flow ψ_h^α with stepwidth h is given by (66).

To illustrate the dynamical behavior of the blending strategy we consider the stiff elastic double pendulum from example 2.1 with slightly unbalanced initial coordinates of the outer pendulum, but tangential ininital momenta. Figure 7 clearly shows that the energy of an initially unbalanced motion dissipates during the blending procedure until it asymptotically matches the initially balanced one. Simultaneously the imbalances reduce to the same order of magnitude as found for the initially balanced motion. Thus, the slow rotational motion is reasonably well approximated.

3.1.2   Ensemble-based penalty method

We now describe a penalty method that allows one to enforce balance relations on the analysis ensemble $\{z_i^{\mathrm{a}}(t_*)\}_{i=1}^M$. We drop the time-dependence for notational convenience from now on and introduce the notation $\hat{z}_i$ for the unbalanced analysis ensemble, i.e. before applying the penalty method.

Therefore, after an assimilation step of the form (13) is carried out, we have an ensemble $\hat{z}_i = (\hat{q}_i, \hat{p}_i)$. In the second step of our DA scheme, we use a variational approach with cost functional

$$J_i(q) = \frac{1}{2}(q - \hat{q}_i)^{\mathrm{T}}(B)^{-1}(q - \hat{q}_i) + \frac{\lambda}{2}g(q)^{\mathrm{T}}Kg(q), \tag{69}$$

where λ is a tuneable parameter, K is again the diagonal matrix of force constants and B could be the identity, a background covariance matrix as in 3DVar, or, as in our experiments, the q-part (the upper left quadrant) of the ensemble covariance matrix P^{a}. The covariance of the posterior distribution represents the uncertainty that we have about the analysis and is therefore a rational choice for weighting how close the solution of 69 should stay to $\hat{q}_i$.

The q-component of the balanced analysis ensemble $\{z_i^{\mathrm{a}}\}_{i=1}^M$ is then defined by

$$q_i^{\mathrm{a}} = \arg\min J_i(q), \tag{70}$$

$i = 1, \ldots, M$.

The derivative of J_i is

$$\nabla_q J_i(q) = (B)^{-1}(q - \hat{q}_i^{\mathrm{a}}) + \lambda \nabla_q g(q) K g(q) \tag{71}$$
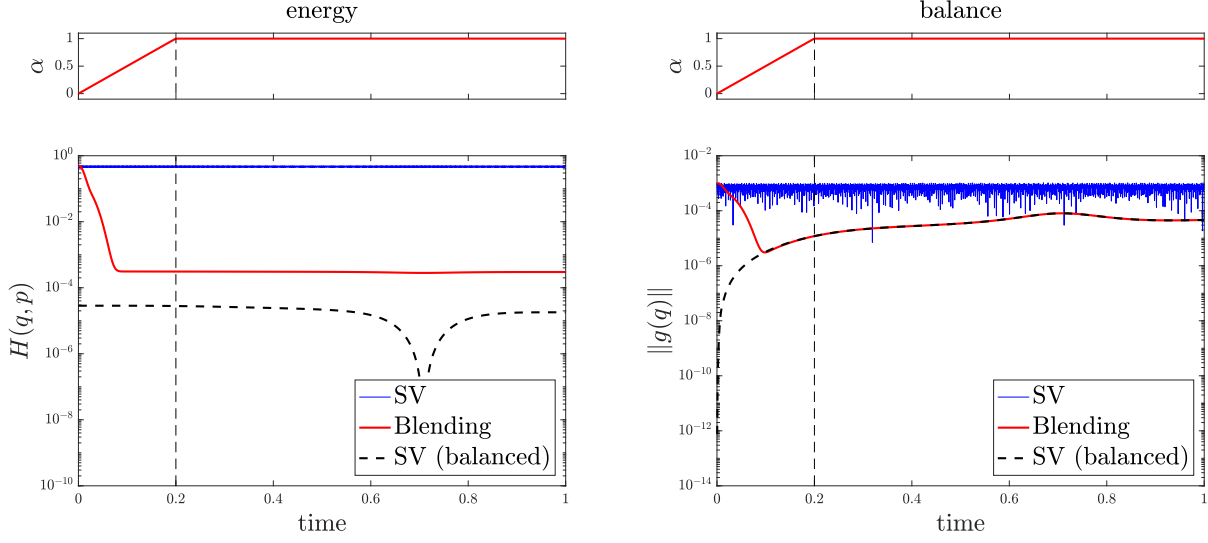
10

Figure 7: Energy, residual of balance relation and solution for the stiff elastic double pendulum c.f. example 2.1. An initially unbalanced solution, which is computed by the Störmer-Verlet method and the blending method is plotted in blue and red respectively. The corresponding parameter α for the blending is shown and the start of the pure forecasting region is marked by a dashed line. As the reference we display the data for an initially balanced solution computed again by the Störmer-Verlet method (dashed lines).

and we approximate $\nabla_q g(q) \approx \nabla_q g(\hat{q}_i) =: G_i^{\mathrm{T}}$ and apply a Newton method to minimize the functional J_i. We then obtain

$$q^{n+1} = q^n - \left((B)^{-1} + \lambda G_i^{\mathrm{T}} K (\nabla_q g(q^n))^{\mathrm{T}} \right)^{-1} \left((B)^{-1}(q^n - \hat{q}_i) + \lambda G_i^{\mathrm{T}} K g(q^n) \right). \tag{72}$$

In the first iteration and for $q^0 = \hat{q}_i$ this gives

$$q^1 = \hat{q}_i - \left((B)^{-1} + \lambda G_i^{\mathrm{T}} K G_i \right)^{-1} \lambda G_i^{\mathrm{T}} K g(\hat{q}_i). \tag{73}$$

Note that if we linearise (69) with

$$g(q) \approx g(\hat{q}_i) + \underbrace{(\nabla_q g(\hat{q}_i))}_{=:G_i}^{\mathrm{T}} (q - \hat{q}_i), \tag{74}$$

we obtain the derivative

$$\nabla_q J_i(q) = (B)^{-1}(q - \hat{q}_i^{\mathrm{a}}) + \lambda G_i^{\mathrm{T}} K \left(G_i q - G_i \hat{q}_i + g(\hat{q}_i) \right) \tag{75}$$

and therefore the solution

$$q_i^{\mathrm{a}} = \left((B)^{-1} + \lambda G_i^{\mathrm{T}} K G_i \right)^{-1} \left(\left((B)^{-1} + \lambda G_i^{\mathrm{T}} K G_i \right) \hat{q}_i - \lambda G_i^{\mathrm{T}} K g(\hat{q}_i) \right) \tag{76}$$

$$= \hat{q}_i - \left((B)^{-1} + \lambda G_i^{\mathrm{T}} K G_i \right)^{-1} \lambda G_i^{\mathrm{T}} K g(\hat{q}_i) \tag{77}$$

which conincides with (73).

With the Sherman-Morrison-Woodbury formula, (73) is equivalent to

$$q_i^{\mathrm{a}} = \hat{q}_i - B G_i^{\mathrm{T}} \left((\lambda K)^{-1} + G_i B G_i^{\mathrm{T}} \right)^{-1} g(\hat{q}_i), \tag{78}$$

(see [14], page 178). If $B = P^{\mathrm{a}}$ is the covariance matrix of the ensemble, this equation is similar to the Kalman formula for updating the mean when using g as the nonlinear observation operator, $(\lambda K)^{-1}$ as the measurement error variance and $y_{\mathrm{obs}} = 0$ as an artificial observation. We will make use of this in scenario B by applying the EnKF with perturbed observations.

11

3.2 Scenario B

3.2.1 EnKF with perturbed pseudo observations

The general description of the penalty method is the same as in scenario A, but because of the appearance of the correction force (52) in this scenario, λ must not be chosen too large. As already mentioned, due to the similarity of the minimiser of (73) to the Kalman formula, another approach to producing balanced analysis ensemble members could be to apply an assimilation step in form of the ensemble Kalman filter with perturbed observations to the already assimilated ensemble.

One knows that $g(q) \sim \mathcal{N}(0, \mathrm{k_B} T \varepsilon^2)$ and therefore we choose our "measurement error" covariance $\frac{1}{\lambda} = \mathrm{k_B} T \varepsilon^2$. When $g(q)$ is the observation operator, the EnKF with perturbed observations creates the assimilated ensemble via

$$q_i^{\mathrm{a}} = \hat{q}_i - \mathcal{K}_i(g(\hat{q}_i) + \xi_i - y_{\mathrm{obs}}) \tag{79}$$

where ξ_i are identically distributed Gaussians with mean 0 and covariance equal to the measurement error covariance $\mathrm{k_B} T \varepsilon^2$, and $y_{\mathrm{obs}} = 0$ is our pseudo observation. The Kalman gain matrix is

$$\mathcal{K}_i = P^{\mathrm{a}} G_i^{\mathrm{T}} \left(\mathrm{k_B} T \varepsilon^2 K^{-1} + G_i P^{\mathrm{a}} G_i^{\mathrm{T}} \right)^{-1}. \tag{80}$$

4 Numerical experiments

4.1 Scenario A

For all experiments we first create a reference trajectory denoted as z^{ref}, which serves as our truth. We use this to create our observations by applying the observation operator H to it after each Δt_{obs} time units and perturbing the resulting quantities by adding noise with mean zero and covariance matrix $R = \rho I$.

ε	K	g_0	δt	M	observeable	Δt_{obs}	ρ	initial variance	infl	total time
0.001	$\mathrm{diag}(1, 0.04)$	10	0.001	20	q	0.02	0.05	0.1	1.05	200

Table 1: Parameters for the numerical experiments with the double pendulum.

For the double pendulum experiment we use $\varepsilon = 0.001$ and a stiffness matrix $K = \mathrm{diag}(1, 0.04)$. We set the gravity constant to $g_0 = 10$, the ensemble size to $M = 20$ and we observe the coordinates of both pendulums every 0.02 time units with an observation error variance of $\rho = 0.05$. The initial uncertainty is Gaussian with a variance of 0.1 but all ensemble members are balanced initially. It is well known that due to a finite ensemblesize, the true covariance of the posterior distribution is underestimated in ensemble based DA methods. One technique to address that issue is ensemble inflation, where the spread of the ensemble is artificially increased around the ensemble mean by

$$z_i^{\mathrm{new}} := \bar{z} + \mathrm{infl}(z_i - \bar{z}). \tag{81}$$

We apply this with parameter $\mathrm{infl} = 1.05$ after every assimilation step. We let the experiment run for 200 time units in total. The parameters of our chosen setting are summarised in table 1.

Figure 8 displays the results of the penalty method for different values of λ in terms of the RMSE in the positions and the impulses in the tangential direction. We observe that the choice of λ is not that relevant as long as it is chosen not too small. Small values of λ cause the time averaged RMSE of the prediction to grow, since the oscillations on the fast time scale will be excited due to the assimilation. We once performed the method as in (72) and once we replaced $g(q^n)$ by $\tilde{g}(q^n)$. The example demonstrates that choosing $\tilde{g}$ instead of g in the Newton iteration yields a small improvement of the error.

Figure 9 depicts the dependency of the time and ensemble averaged fast energy on λ. As one would expect, it is very small for rather big values of λ and grows with λ decreasing.

In Figure 10 we show the dependence of the RMSE, measured in the positions and the tangential impulses, on the window size of the blending method. For blending windows shorter than shown in the graphs, the errors will be very large. Thus, for window sizes of 5 and above, the blending methods yields an even smaller error
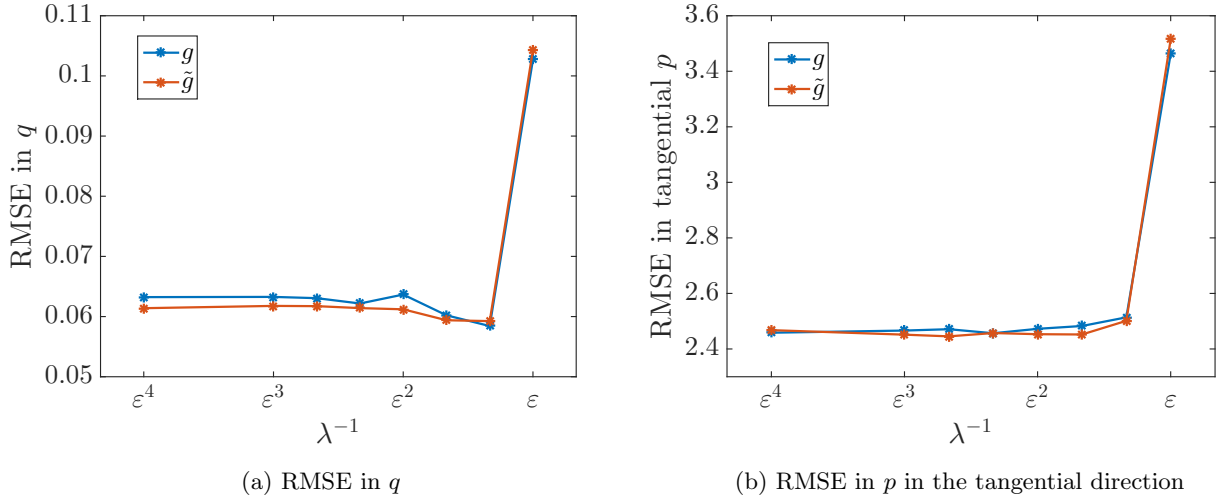
(a) RMSE in q

(b) RMSE in p in the tangential direction

Figure 8: Time averaged values of the RMSE in q and p in the tangential direction for the penalty method applied to the double pendulum, plotted against different values of λ
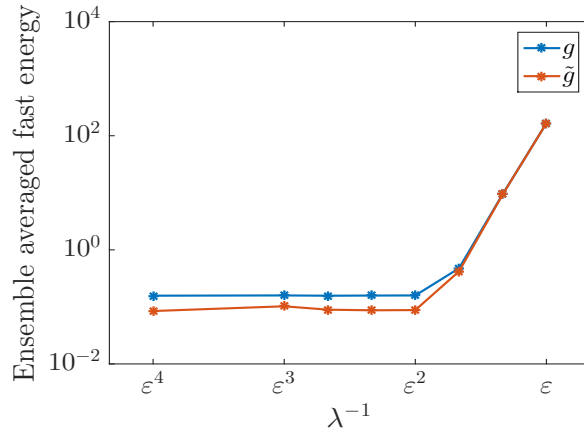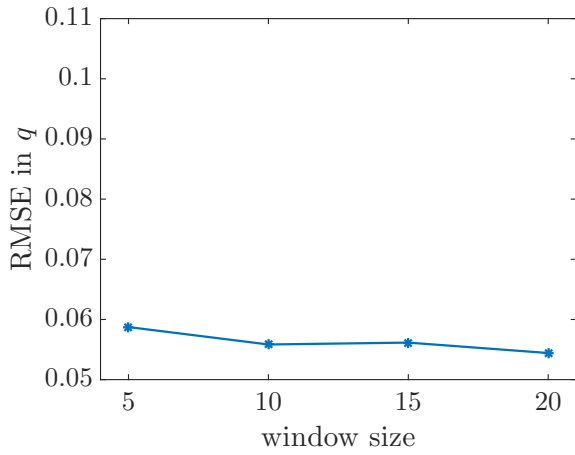

Figure 9: Time averaged values of the ensemble average of the fast energy for the penalty method applied to the double pendulum, plotted against different values of λ

than the penalty method. The biggest window size ($= 20$) coincides with the whole length of the forecast interval.

We show the dependency of the time and ensemble averaged fast energy $H_{\mathrm{osc}}^{\varepsilon}(q, p)$ on the size of the blending window in Figure 11 and observe that it decreases with increasing window size but remains larger than with the penalty method.

It should be noted that for the penalty method M nonlinear optimization problems have to be solved after each assimilation step whereas the blending scheme requires $M \times$ size of the blending window nonlinear equation solvings.
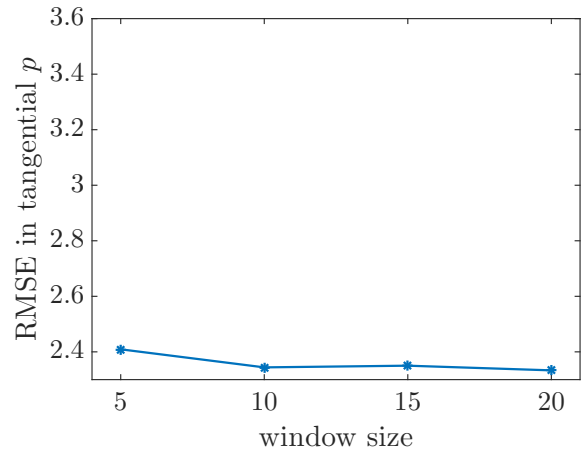
(a) RMSE in q (b) RMSE in p in the tangential direction

Figure 10: Time averaged values of the RMSE in q and in p in the tangential direction for the blending method applied to the double pendulum, plotted against different sizes of the blending window
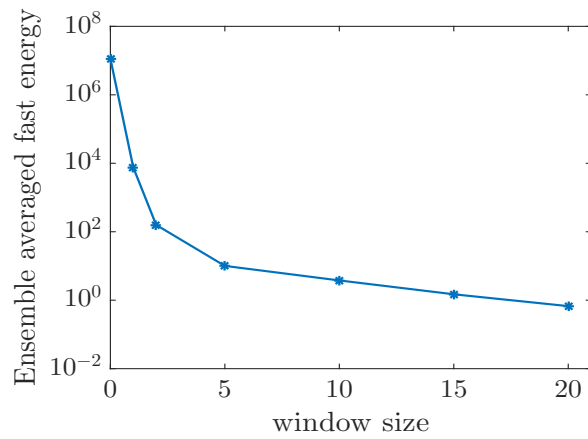

Figure 11: Time averaged values of the ensemble average of the fast energy for the blending method applied to the double pendulum, plotted against different sizes of the blending window

4.2 Scenario B

ε	K	A	g_0	γ	k_BT	δt	M	observeable	Δt_{obs}	ρ	initial variance	time
0.001	1	$\text{diag}(1, 36)$	0	1	16	5^{-6}	20	p	0.01	0.1	0.1	200

Table 2: Parameters for the numerical experiments with the thermally embedded system

For our experiments with the thermally embedded elliptic elastic pendulum (example 2.2), we choose $g(q) = \sqrt{q^{\text{T}}Aq} - 1$ with $A = \text{diag}(1, 36)$. We set $\varepsilon = 0.001$, $V(q) = 0$, $\gamma = 1$ and $k_BT = 16$. This time we observe the impulses of the pendulum every 0.01 time units with an observation error variance of $\rho = 0.1$. Initial conditions are at the slow manifold but with an impuls in the normal direction and a Gaussian initial uncertainty with variance 0.1. We let the experiment run for 200 time units. The setting we chose for this model can be found in table 2.

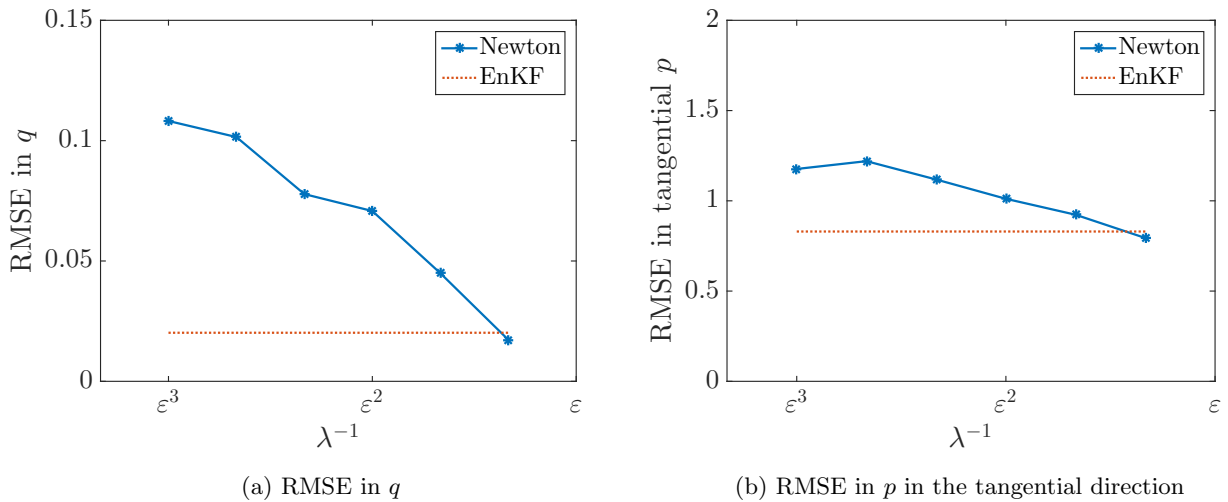(a) RMSE in q (b) RMSE in p in the tangential direction

Figure 12: Time averaged values of the RMSE for the penalty method (Newton) applied to the thermally embedded system, plotted against different values of λ. The dotted line (EnKF) shows the corresponding results obtained with the straight EnKF method including pseudo-observations from section 3.2.1. Errors obtained with the straight EnKF method without any balancing mechanism are larger by a factor of 25 in the positions and a factor of 20 in the tangential impulses than those shown.

Figure 12 shows that choosing λ too large in the penalty method yields nonoptimal results in this scenario. The influence of the parameter λ on the action variable (51) of the ensemble is displayed in Figure 13. When choosing λ too big, J will be too small and therefore the correction force (52) will be too small as well. That leads to correspondingly erroneous predictions of the positions and the tangential velocity. For $\lambda^{-1} > \frac{\varepsilon}{10}$ the action variable explodes and therefore the RMSE in q and p in the tangential direction is very large.

When using the EnKF with perturbed pseudo observations we obtain a RMSE in q of 0.02 and in p in the tangential direction of 0.83, which both approximately coincide with the smallest error we got when minimising (69) via the newton method. The time averaged value of J of the ensemble when using the EnKF is 7.45.

The results of applying the blending method to the thermally embedded system are provided in Figure 14. Even though the blending approach is not designed to address the scenario B with stochastic perturbations, the errors are much smaller than those obtained with the EnKF method without pseudo-observations and tend to decrease with the size of the blending window. Yet, for this model they remain larger than those generated by the penalty method and especially the EnKF method with pseudo-observations. Again the biggest window (= 2000) coincides with the whole length of the forecast interval. Figure 15 depicts the dependence of the action variable (51) on the window size.
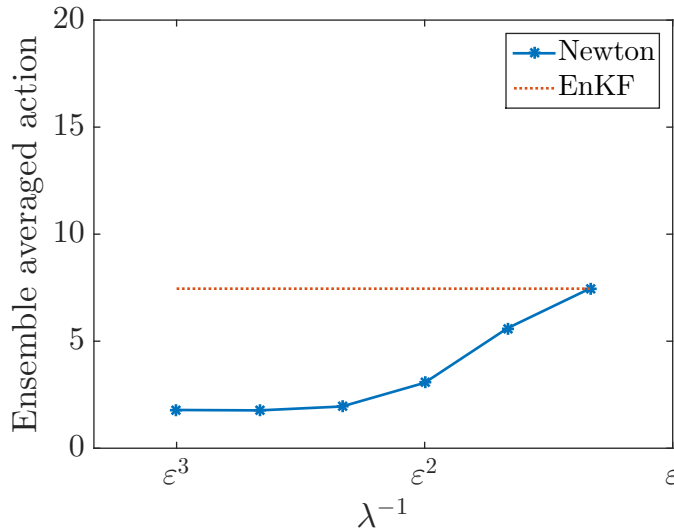
Figure 13: Dependence of the action variable J of the thermally embedded system on the balance parameter λ of the penalty method

(a) RMSE in q
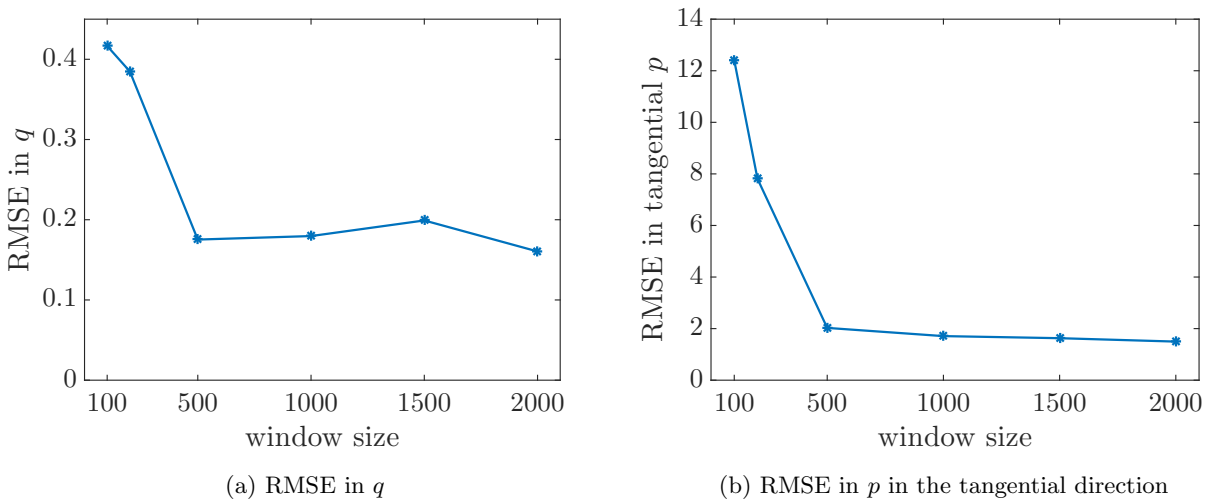

(b) RMSE in p in the tangential direction

Figure 14: Time averaged values of the RMSE for the blending method applied to the thermally embedded system, plotted against different sizes of the blending window

5   Conclusions

We have proposed two modifications of the standard EnKF when applied to highly oscillatory systems; namely blended time-stepping and ensemble-based penalization, respectively. Blended time-stepping has been demonstrated to be efficient when the oscillatory energy of the Hamiltonian system is small. The method can also be used for stabilizing the EnKF for thermally embedded Hamiltonian systems when the oscillatory energy is of order one and can give rise to a correcting force term along the slow manifold. However, the resulting estimation errors are generally found to be larger than those obtained from the ensemble-based penalty method, which does a better job in maintaining the oscillatory energy of the system in this case. We have also discussed the practical choice of the tuneable parameters such as the penalty parameter λ and the background covariance matrix B for the penalty method, and the number of blending time-steps k for the blending scheme. The broader application areas of the two proposed stabilization techniques can, for example, be found in geophysics and molecular dynamics. Those application areas have motivated the scenarios A (small oscillatory energy, conservative motion) and B (oscillatory energy of order one, thermally embedded dynamics), respectively.
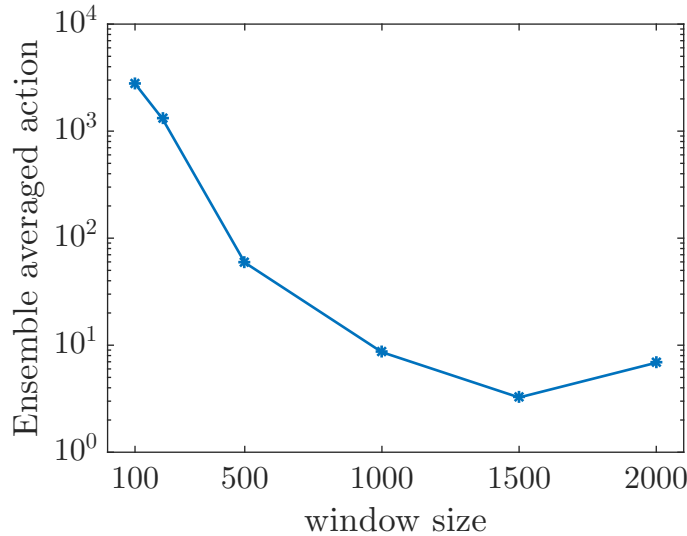
Figure 15: Dependence of the action variable J of the thermally embedded system on the window size when applying the blending method

Acknowledgments

References

[1] H. C. Andersen. Rattle - a velocity version of the shake algorithm for molecular-dynamics calculations. *Journal of Computational Physics*, 52(1):24–34, 1983.

[2] T. Benacchio, W. P. O'Neill, and R. Klein. A blended soundproof-to-compressible numerical model for small- to mesoscale atmospheric dynamics. *Monthly Weather Review*, 142(12):4416–4438, 2014.

[3] K. Bergemann and S. Reich. A mollified ensemble Kalman filter. *Quarterly Journal of the Royal Meteorological Society*, 136:1636–1643, 2010.

[4] S.C. Bloom, L. L. Takacs, A.M. Da Silva, and D. Ledvina. Data assimilation using incremental analysis updates. *Monthly Weather Review*, 124:1256–1271, 1996.

[5] F. A. Bornemann and C. Schütte. Homogenization of Hamiltonian systems with a strong constraining potential. *Physica D*, 102(1-2):57–77, 1997.

[6] A. J. Chorin. The numerical solution of Navier-Stokes equations for an imcompressible fluid. *Bulletin of the American Mathematical Society*, 73(6):928–931, 1967.

[7] C. Cotter. Data assimilation on the exponentially accurate slow manifold. *Philosophical Transactions of the Royal Society A*, 371(1991), 2013.

[8] G. Evensen. The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53:343, 2003.

[9] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations.* Springer series in computational mathematics. Springer, Berlin u.a., 2. edition, 2010.

[10] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35, 1960.

[11] P. Lynch and X.-Y. Huang. Initialization of the HIRLAM model using a digital filter. *Monthly Weather Review*, 120:1019–1034, 1992.

[12] S. Reich. Smoothed dynamics of highly oscillatory Hamiltonian systems. *Physica D*, 89:28—42, 1995.

[13] S. Reich. Smoothed Langevin dynamics of highly oscillatory systems. *Physica D*, 138:210–224, 2000.

[14] S. Reich and C. Cotter. Probabilistic forecasting and Bayesian data assimilation: a tutorial. *Cambridge University Press*, 2015.

[15] S. Reich and T. Hundertmark. On the use of constraints in molecular and geophysical fluid dynamics. *European Physical Journal Special Topics*, 200:259—270, 2011.

[16] H. Rubin and P. Ungar. Motion under a Strong Constraining Force. *Communications on Pure and Applied Mathematics*, 10(1):65–87, 1957.

[17] F. Takens. *Motion under the influence of a strong constraining force.* Springer Berlin Heidelberg, 1980.

[18] J. Zhou, S. Reich, and B.R. Brooks. Elastic molecular dynamics with self-consistent flexible constraints. *J. Chem. Phys.*, 112:7919—7929, 2000.

A   Tangential momentum method

Motivated by the Störmer-Verlet method we propose a projection method for (36)–(38) which satisfies the hidden constraint

$$G(q)p = 0 \tag{82}$$

up to a given tolerance and the explicit constraint (38) up to a global error of order $o(h^2)$. Although usually one would like to bound the constraint (38) by an a priori given tolerance, we intentionally violate this principle. There are two reasons for this. Given non tangential initial values an update by e.g. SHAKE/RATTLE will map onto the constraint manifold, but unfortunately the momenta can change their direction drastically. In the context of data assimilation this leads to completely wrong forecasts. Despite this practical numerical problem, we also want to leverage the more accurate dynamics of the full model and therefore avoid the complete elimination of all unbalanced dynamics. Our proposed method reads

$$q_{n+\frac{1}{2}} = q_n + \frac{h}{2}p_n \tag{83}$$

$$p_{n+1} = p_n - h\nabla V(q_{n+\frac{1}{2}}) - hG^T(q_{n+\frac{1}{2}})\lambda_{n+\frac{1}{2}} \tag{84}$$

$$q_{n+1} = q_{n+\frac{1}{2}} + \frac{h}{2}p_{n+1} \tag{85}$$

$$G(q_{n+1})p_{n+1} = 0. \tag{86}$$

We immediately see that our method is a second order consistent method to (36)–(37).

Lemma A.1. *The flow ψ_h given by method (83)-(86) is symmetric and therefore of second order. In particular the global error of the constraint satisfies*

$$\|g(q(t_n)) - g(q_n)\| \leq Ch^2. \tag{87}$$

Proof. Since after one step the tangential condition (86) is satisfied, we can assume

$$G(q_n)p_n = 0 \tag{88}$$

without loss of generality. To prove symmetry we have to check $\psi_{-h}^{-1} = \psi$. To find the equations of ψ_{-h}^{-1} we consider (83) – (88) and exchange h by $-h$ and (q_n, p_n) by (q_{n+1}, p_{n+1}). After these transformations the equations become

$$G(q_{n+1})p_{n+1} = 0 \tag{89}$$

$$q_{n+\frac{1}{2}} = q_{n+1} - \frac{h}{2}p_n \tag{90}$$

$$p_n = p_{n+1} + h\nabla V(q_{n+\frac{1}{2}}) + hG(q_{n+\frac{1}{2}})\lambda_{n+\frac{1}{2}} \tag{91}$$

$$q_n = q_{n+\frac{1}{2}} - \frac{h}{2}p_n \tag{92}$$

$$G(q_n)p_n = 0, \tag{93}$$

which are exactly (83) – (88). Therefore the method is symmetric. Since any symmetric method has even order [9] and the proposed method is a composition of first order consistent methods, we conclude the statement. The bound for the constraint (87) follows then directly by expanding $g(q(t))$ around $g(q_n)$. ∎

For the implementation we need to solve the weakly nonlinear equation (95) in $\tilde{\lambda}_{n+\frac{1}{2}}$

$$p_{n+1} = p_n - h\nabla V(q_{n+\frac{1}{2}}) - G^T(q_{n+\frac{1}{2}})\tilde{\lambda}_{n+\frac{1}{2}} \tag{94}$$

$$G\left(q_{n+\frac{1}{2}} + \frac{h}{2}p_{n+1}\right)G^T(q_{n+\frac{1}{2}})\tilde{\lambda}_{n+\frac{1}{2}} = G\left(q_{n+\frac{1}{2}} + \frac{h}{2}p_{n+1}\right)\left(p_n - h\nabla V(q_{n+\frac{1}{2}})\right) \tag{95}$$

where $\tilde{\lambda}_{n+\frac{1}{2}} = h\lambda_{n+\frac{1}{2}}$ for the sake of numerical stability. To solve the equation we use e.g. a fixed point iteration over $\lambda_{n+\frac{1}{2}}$ i.e. in step k we first compute p_{n+1}^k by (94) and subsequently solve the linear equation (95) for $\lambda_{n+\frac{1}{2}}^k$.

B   Blending method

B.1   Harmonic oscillator

To understand the general linear behavior of our blending method we consider the classical harmonic oscillator

$$H(q,p) = \frac{1}{2}Kq^2 + \frac{1}{2}p^2 \tag{96}$$

$$\dot{q} = p \tag{97}$$

$$\dot{p} = -Kq \tag{98}$$

If we apply the blending method as given in (66) we can explicitly find the *Lagrange* multiplier $\lambda_{n+\frac{1}{2}}$ and therefore get the discrete flow map

$$\begin{pmatrix} q_{n+1} \\ p_{n+1} \end{pmatrix} = \psi_h^\alpha \left(\begin{pmatrix} q_n & p_n \end{pmatrix}^T \right) = \begin{pmatrix} 1 - \frac{K\alpha h^2}{2} & \frac{(1+\alpha)h}{2} + \frac{K\alpha h^3}{4} \\ -Kh\alpha & \alpha - \frac{K\alpha h^2}{2} \end{pmatrix} \begin{pmatrix} q_n \\ p_n \end{pmatrix} \tag{99}$$

To analyze the stability properties to compute the Eigenvalues of ψ_h^α which are given by

$$\Lambda_{1,2} = \frac{1+\alpha}{2} - \frac{Kh^2\alpha}{2} \pm \frac{1}{2}\sqrt{d}, \quad d := K^2 h^4 \alpha^2 - 2Kh^2\alpha^2 - 2Kh^2\alpha + \alpha^2 - 2\alpha + 1. \tag{100}$$

The absolut values of the Eigenvalues $\|\Lambda_{1,2}\| \leq 1$ are depicted in figure 16b. The Störmer-Verlet method ($\alpha = 1$) is stable for $Kh^2 < 2$ and so is the blending method for every $\alpha \in (0,1)$. Furthermore we conclude two important findings from the figure. First the method is a projection for $\alpha = 0$ since one of the Eigenvalues is zero in this case and second the blending method is dissipative for $\alpha \in (0,1)$. During this dissipative transition phase a bifurcation occurs which splits a conjugated pair of complex Eigenvalues to a pair of real ones. This happens at $\alpha_\pm$, introduced in (101), when the discriminant d changes sign.

$$\alpha_\pm = \begin{cases} \frac{1}{4} & Kh^2 = 1 \\ \frac{Kh^2+1\pm 2\sqrt{Kh^2}}{(Kh^2-1)^2} & \text{else} \end{cases} \tag{101}$$

Therefore the red and the white area in figure 16a depict the parameter configurations for a (stable) node and a (stable) spiral respectively and we still have to expect oscillatory behavior as long as α is close enough to 1.
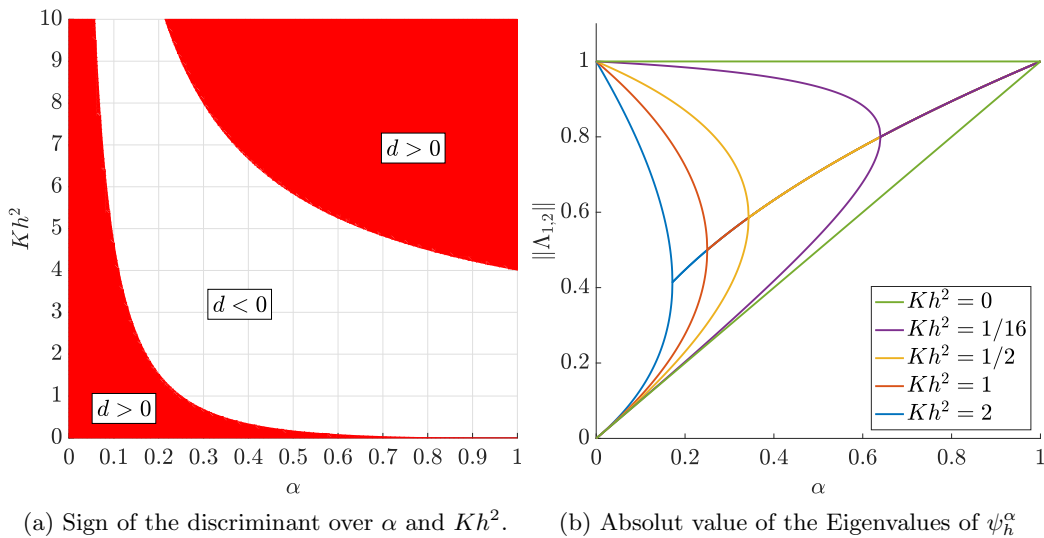

(a) Sign of the discriminant over α and Kh^2.   (b) Absolut value of the Eigenvalues of ψ_h^α

Figure 16: In (a) we depict the areas for negative or positive discriminant of the Eigenvalues for ψ_h^α and in (b) we depict the absolute values of the Eigenvalues as functions of α for different Kh^2.

B.2 Coupled harmonic oscillator

So far we omitted an explicit second time scale in our discussion. To understand the scale interaction we now consider a coupled harmonic oscillator given by

$$H(q,p) = \frac{1}{2}(p_1^2 + p_2^2) + \frac{K}{2}(q_1 - q_2)^2 + \frac{q_2^2}{2} \tag{102}$$

$$\dot{q} = p \tag{103}$$

$$\dot{p} = -K \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} q - \begin{pmatrix} 0 \\ q_2 \end{pmatrix}. \tag{104}$$

This system has the steady state $q_{1,2} = 0, p_{1,2} = 0$ and a slow manifold $q_1 = q_2$ with $p_1 = p_2$. As stated before, we are interested in the slow dynamics. In this simple case we can actually express the fast and the slow variables, given by x and y respectively, explicitly by the following transformation

$$x = \frac{q_1 - q_2}{2} \tag{105}$$

$$y = \frac{q_1 + q_2}{2}. \tag{106}$$

Using this we can rewrite (103)–(104) as

$$H(x,y,p_x,p_y) = p_x^2 + p_y^2 + 2Kx^2 + \frac{(x-y)^2}{2} \tag{107}$$

$$\dot{x} = \frac{p_x}{2} \tag{108}$$

$$\dot{y} = \frac{p_y}{2} \tag{109}$$

$$\dot{p_x} = -4Kx - (x - y) \tag{110}$$

$$\dot{p_y} = (x - y), \tag{111}$$

where we can see the slow manifold directly. Keeping that in mind we apply the blending method given by (66) to the equations of motion (103)-(104) again. Since our system is linear we can also explicitly calculate the *Lagrange* multiplier

$$\lambda_{n+\frac{1}{2}} = \frac{hp_{2,n}}{4} + \frac{q_{2,n}}{2} + \frac{p_{1,n}}{2h} - \frac{p_{2,n}}{2h}. \tag{112}$$

Finally our discretization reads

$$\begin{pmatrix} q_{n+1} \\ p_{n+1} \end{pmatrix} = A \begin{pmatrix} q_n \\ p_n \end{pmatrix} \tag{113}$$

$$\tag{114}$$

with

$$A := \begin{pmatrix} \frac{K\alpha}{2}h^2 + 1 & \frac{K\alpha}{2}h^2 + \frac{\alpha h^2}{4} - \frac{h^2}{4} & -\frac{K\alpha}{4}h^3 + \frac{\alpha h}{4} + \frac{3h}{4} & \frac{K\alpha}{4}h^3 + \frac{\alpha h^3}{8} - \frac{\alpha h}{4} - \frac{h^3}{8} + \frac{h}{4} \\ \frac{K\alpha}{2}h^2 & -\frac{K\alpha}{2}h^2 - \frac{\alpha h^2}{4} - \frac{h^2}{4} + 1 & \frac{K\alpha}{4}h^3 - \frac{\alpha h}{4} + \frac{h}{4} & -\frac{K\alpha}{4}h^3 - \frac{\alpha h^3}{8} + \frac{\alpha h}{4} - \frac{h^3}{8} + \frac{3h}{4} \\ -K\alpha h & K\alpha h + \frac{\alpha h}{2} - \frac{h}{2} & -\frac{K\alpha}{2}h^2 + \frac{\alpha}{2} + \frac{1}{2} & \frac{K\alpha}{2}h^2 + \frac{\alpha h^2}{4} - \frac{\alpha}{2} - \frac{h^2}{4} + \frac{1}{2} \\ K\alpha h & -K\alpha h - \frac{\alpha h}{2} - \frac{h}{2} & \frac{K\alpha}{2}h^2 - \frac{\alpha}{2} + \frac{1}{2} & -\frac{K\alpha}{2}h^2 - \frac{\alpha h^2}{4} + \frac{\alpha}{2} - \frac{h^2}{4} + \frac{1}{2} \end{pmatrix} \tag{115}$$

where the stepwidth is given by h. As in the previous case of the harmonic oscillator, we now want to check the stability i.e. the spectral norm and also the other absolute values of the Eigenvalues. Since the algebraic expressions are rather lengthy, we restrict ourselves to the numerical evaluation in figure 17a and figure 17b. In contrast to the simpler example in section B.1, this time the Eigenvalues not only depend on Kh^2, but on K too. So K can be interpreted as scale separation parameter and indeed figure 17a shows the desired dissipative behavior as for the harmonic oscillator. For smaller K we can already see deformations in 17b which can lead to instabilities in more severe cases. To finally argue why the blending method is a reasonable approximation,
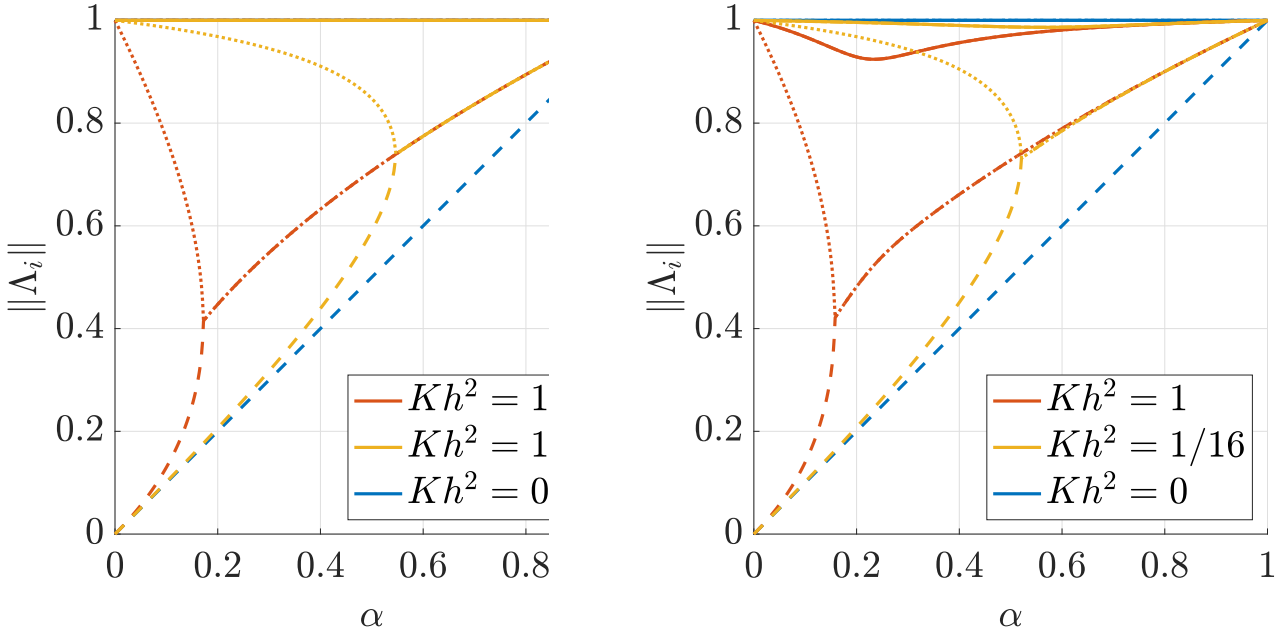
(a) Absolute values of the Eigenvalues for $K = 10^8$ (b) Absolute values of the Eigenvalues for $K = 1$

Figure 17: The trend of the Eigenvalues over varying α. In every case there is at least a pair of Eigenvalues, of which the absolute values are close to unity. For $Kh^2 = 0$ the splitting point of the conjugate pair moves up to the origin and therefore there is a third one on top and only a single one going from one to zero. For the less well separated case (b) dissipation is introduced through all Eigenvalues, although this defect vanishes as $K \to \infty$.

we compare solutions given by the blending method applied to (103)–(104) with unbalanced solutions, but both under the transformation (105) – (106). In figure 18 we immediately see that the fast component is quickly dissipated and the slow component is resolved rather well. Nevertheless a small phase error due to the blending is still present. In simulations that are not shown, we find the quality of the blending to depend on the scale separation, so that reasonably large K are required for the strategy to work (c.f. the Eigenvalues whose absolute value are close to one in figure 17a, although they are clearly in the dissipative regime in the case of figure 17b). This is expected, though, since the blending approach has been designed to handle systems involving a small scale separation parameter which only in the asymptotic limit leads to a system with constrained dynamics. It is not specifically designed to handle systems that support particular solutions which are constrained to a manifold in the absence of a time scale separation.
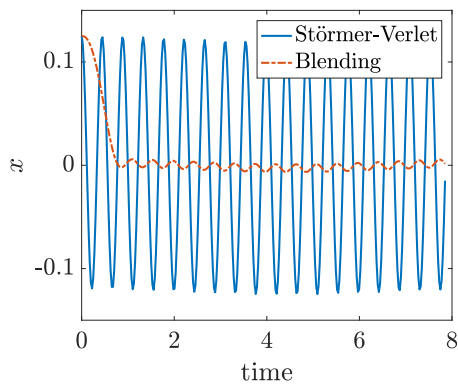
(a) Fast component over time
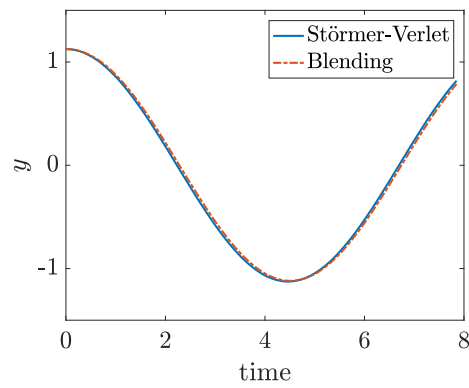
(b) Slow component over time

Figure 18: The figures show the fast and slow coordinates of the coupled harmonic oscillator with stiffness $K = 10^2$. The solutions are generated by the Störmer-Verlet scheme and the blending scheme using stepwidth $h = 2.5 \, 10^{-2}$. For the blending parameter we use linear interpolation from 0 to 1 over a blending window of $k = 40$ steps.