OPTIMAL BALANCE VIA ADIABATIC INVARIANCE OF APPROXIMATE SLOW MANIFOLDS

GEORG A. GOTTWALD, HAIDAR MOHAMAD, AND MARCEL OLIVER

Abstract. We analyze the method of optimal balance which was introduced by Viúdez and Dritschel (J. Fluid Mech. **521**, 2004, pp. 343–352) to provide balanced initializations for two-dimensional and three-dimensional geophysical flows, here in the simpler context of a finite dimensional Hamiltonian two-scale system with strong gyroscopic forces. It is well known that when the potential is analytic, such systems have an approximate slow manifold that is defined up to terms that are exponentially small with respect to the scale separation parameter. The method of optimal balance relies on the observation that the approximate slow manifold remains an adiabatic invariant under slow deformations of the nonlinear interactions. The method is formulated as a boundary value problem for a homotopic deformation of the system from a linear regime, where the slow-fast splitting is known exactly, to the full nonlinear regime. We show that, providing the ramp function which defines the homotopy is of Gevrey class 2 and satisfies vanishing conditions to all orders at the temporal end points, the solution of the optimal balance boundary value problem yields a point on the approximate slow manifold that is exponentially close to the approximation to the slow manifold via exponential asymptotics, albeit with a smaller power of the small parameter in the exponent. In general, the order of accuracy of optimal balance is limited by the order of vanishing of derivatives of the ramp function at the temporal end points. We also give a numerical demonstration of the efficacy of optimal balance, showing the dependence of accuracy on the ramp time and the ramp function.

1. Introduction

Nonlinear Hamiltonian two-scale systems with a single fast frequency possess an approximate slow manifold: a region in phase space characterized by smallness of an adiabatically invariant "fast energy". A trajectory near the approximate slow manifold will stay near it for a long period of time—often exponentially long with respect to the scale separation parameter under suitable assumptions (see, e.g., [15, 20]). It is important to stress that, despite the language used, this phase space region is not a manifold in any rigorous sense (except in trivial cases such as linear ODEs). Rather, it is described by a generally diverging asymptotic series [25].

An explicit description of an approximate slow manifold is usually only practical to a low fixed order of asymptotics because the number of terms grows exponentially with order. Optimal truncation, a powerful theoretical tool, e.g., for proving almost-invariance over exponentially long times, cannot be implemented in a computational model. It is, however, possible to numerically compute single points on the approximate manifold with an accuracy that is nearly as good as optimal

Date: October 16, 2017.

truncation. This procedure, which we refer to as *optimal balance*, is the subject of this paper.

The idea underlying optimal balance is that adiabatic invariants of the unperturbed dynamics remain adiabatic under slowly varying perturbations. If a homotopy varying in slow time perturbs the system from linear to fully nonlinear, trajectories that emerge from the known slow subspace at the linear end will connect to an approximately slow fully nonlinear state at the other end. Computationally, this amounts to solving a boundary value problem where the boundary condition at the linear end constrains to the slow linear subspace and the boundary condition at the fully nonlinear end constrains to the slow base-point coordinate of the approximate manifold.

Our motivation comes from studying balance in geophysical fluid flow. On large scales in the mid-latitudes, the Coriolis force nearly balances the pressure gradient force while inertial forces are subdominant. As a result, the flow approximately splits into a slow *balanced* component which evolves nonlinearly and interacts only weakly with the fast components which are approximately described by linear waves. A precise characterization of this splitting is a perennial theme in geophysical fluid dynamics; we refer the reader to the reviews of Vanneste [25] and McIntyre [18] for a more comprehensive background.

A computational procedure for describing balance is of considerable practical importance. First, unphysically imbalanced initial conditions may require unnecessary large amounts of artificial viscocity to ensure stability in a numerial simulation; thus, accurate balancing can improve numerical accuracy, in particular when frontogenesis is important [4]. Second, in studies of the role of inertial-gravity waves in the energy budget of the ocean, accurate diagnostics are currently lacking; optimal balance may provide a way to diagnose small imbalanced components in large unsteady flows with minimal ambiguity [27]. Third, enforcing balance is a practical necessity when assimilating noisy observations to initialize a weather forecast; not doing so results in spectacular failure (see, e.g., the wonderful historical account in [13]). To ensure that the assimilated state is consistently balanced, the analysis output is typically post-processed, e.g. using a digital filter [14]. Dynamical information about imbalance and approximate slow manifolds has only recently become part of the actual data assimilation procedure [19, 7]. Cotter [2], in particular, demonstrates that optimal balance can be used as a constraint when assimilating balanced states in a simple two-scale Hamiltonian model problem.

The method of optimal balance for rotating fluid flow was first proposed by Viúdez and Dritschel [26]. In their work, they coin the term "optimal potential vorticity balance" which reflects that rather than deforming the equations of motion, they ramp up the vorticity anomaly in the initial data. Mathematically, this is equivalent to homotopically turning on nonlinear interactions. In practical terms, this is feasible only when using a potential-vorticity-based fully Lagrangian code. In their work, they suggest a simple iterative scheme to solve the resulting boundary value problem and report good behavior both in terms of convergence of the algorithm and in terms of quality of balance as measured by independent diagnostics.

Cotter [2] studies optimal balance for data assimilation using a simple finite-dimensional Hamiltonian system which has been used as a prototype model for balance in a number of previous studies [23, 3, 8]. In particular, Cotter points out that their earlier results [3] imply rigorous exponential estimates for analytic ramp functions with exponentially decaying tails.

In the present paper, we consider optimal balance in the same finite-dimensional setting on a fixed finite interval in slow time. In this setting, the asymptotic behavior of the method is determined not only by the smoothness of the potential and of the ramp function, but also by the order of vanishing of the derivatives of the ramp function at the temporal end points. When the derivatives of the ramp function vanish only up to some finite order k at the initial and at the final time of the ramp, the rate of convergence of optimal balance is limited to $O(\varepsilon^{k+1})$, where ε denotes the time-scale separation parameter. Correspondingly, beyond-all-order accuracy requires that *all* derivatives of the ramp function vanish at the end points. However, the ramp function cannot simultaneously be uniformly analytic and satisfy the correct end-point conditions. Here, we show that exponential estimates can still be achieved when the ramp function is not analytic but of Gevrey class 2.

When the potential in this model is analytic, classical Hamiltonian normal form theory states that there exists a constant c and a symplectic transformation which approximately splits the system into fast and slow variables such that when the fast variable is initially zero, it remains $O(\exp(-c/\varepsilon))$ over times of $O(\exp(c/\varepsilon))$ as $\varepsilon \to 0$ [3]. Here we prove that optimal balance using the ramp functions described above yields a state that, if used as initial data for the original system, corresponds to a normal-form fast variable that remains $O(\exp(-c/\varepsilon^{1/3}))$-small over times of $O(\exp(c/\varepsilon))$. We present numerical results that indicate that the exponent $\frac{1}{3}$ is not sharp, but that an exponent 1 as in the classical normal form result cannot be achieved.

This result provides a strong justification of the method of optimal balance: the algorithm yields a point on the approximate slow manifold that is exponentially close to what could be obtained from an optimally truncated asymptotic expansion.

The method of proof has a long history. A concise mathematical treatment of adiabatic invariance for linear systems is given by Leung and Meyer [11], we refer the reader to this paper for some of the early history. Exponential estimates for nonlinear systems are due to Nekhoroshev [22] and Neĭshtadt [20, 21]. Cotter and Reich [3] apply this theory to the model problem under consideration here. In our proof, we do not use their Hamiltonian setting because the recursive step is easy only when applying a Cauchy estimate at each iteration. When the ramp function is not analytic but only of Gevrey class 2, Cauchy estimates are not available and the iteration does not directly close up. As we do not need estimates over times longer than $O(1)$ in slow time, we resort to more direct estimates on an explicit construction of the fast-slow splitting as used in [8].

The paper is organized as follows. In Section 2, we detail the finite-dimensional model for balance and review the direct construction of the slow vector field. In Section 3, we describe the method of optimal balance applied to this model. We state and prove our main theorems on optimal balance in Sections 4 and 5 for C^k potentials and for analytic potentials, respectively. Section 6 presents numerical simulations corroborating our analytical results. Section 7 concludes with a discussion and open questions.

2. The model

We consider the Hamiltonian system of differential equations

$$\dot{q} = p\,, \tag{1a}$$

$$\dot{p} = Jp - \varepsilon\,\nabla V(q)\,, \tag{1b}$$

where $q\colon [0,T] \to \mathbb{R}^{2d}$ is the vector of positions, p is the vector of corresponding momenta, J is the canonical symplectic matrix in $2d$ dimensions, V is a smooth potential and ε is a small parameter.

When $d=1$, this system can be interpreted as describing the motion of a single Lagrangian particle in the rotating shallow water equations with frozen height field [23, 3, 6]. In this interpretation, $J\dot{q}$ represents the Coriolis force and $\varepsilon \to 0$ describes the limit of rapid rotation. Alternatively, (1) can be seen as describing the motion of a single charged particle in a planar potential V under the influence of a magnetic field normal to the plane of motion. In this interpretation, $J\dot{q}$ represents the Lorentz force and $\varepsilon \to 0$ corresponds to the mass of the particle going to zero while its charge remains constant.

The system (1) is Hamiltonian, albeit with a non-canonical symplectic structure. To leading order, the splitting into slow and fast degrees of freedom can be determined by inspection. Indeed, rescaling to slow time $\tau = \varepsilon t$, introducing a slow momentum $\pi = p/\varepsilon$, and setting $\varepsilon = 0$, we see that the leading order slow dynamics is given by

$$\frac{\mathrm{d}q}{\mathrm{d}\tau} = -J\nabla V(q)\,, \tag{2}$$

so that the corresponding leading order fast variable is $\omega = \pi + J\nabla V(q)$. This splitting can be iteratively refined by adding higher order terms, which gives an explicit formula for the nth-order slow vector field $G_n(q)$ which is needed as a reference for the optimal balance vector field to compare against and which is stated here in terms of the original fast time variables.

Theorem 1 ([8]). *For $n \in \mathbb{N}$, suppose $V \in C^{n+2}$, and set*

$$G_n(q) = \varepsilon \sum_{i=0}^{n} g_i(q)\,\varepsilon^i \tag{3}$$

with coefficient functions g_i recursively defined via

$$g_0(q) = -J\nabla V(q)\,, \tag{4a}$$

$$g_k(q) = -J \sum_{i+j=k-1} \mathrm{D}g_i(q)\,g_j(q)\,. \tag{4b}$$

For fixed $q_0 \in \mathbb{R}^{2d}$ and $a>0$, let $q(t)$ denote a solution to

$$\dot{q} = G_n(q) \tag{5}$$

with $q(0) = q_0$. Let $q_\varepsilon(t)$ solve the full parent dynamics (1) consistently initialized via $q_\varepsilon(0) = q_0$ and $p_\varepsilon(0) = G_n(q_0)$. Then there exist $\varepsilon_0 > 0$ and $c = c(q_0, a, V)$ such that

$$\sup_{t\in[0,a/\varepsilon]} \|q_\varepsilon(t) - q(t)\| \le c\,\varepsilon^{n+2} \tag{6}$$

for all $0 < \varepsilon \le \varepsilon_0$.

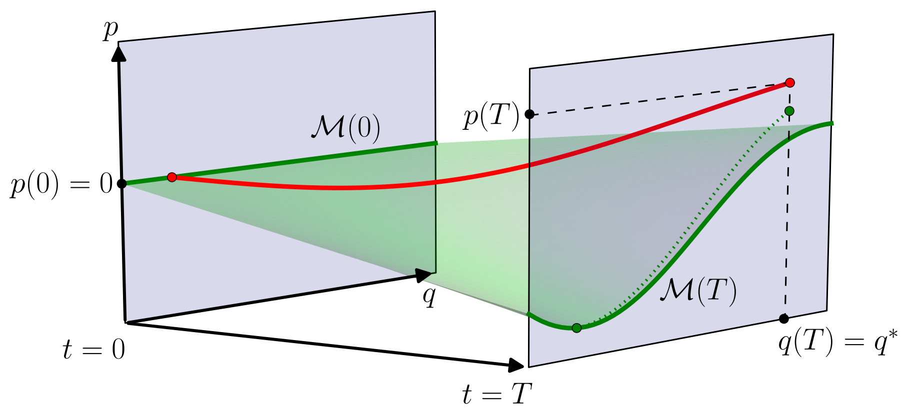

FIGURE 1. Sketch of the geometry of optimal balance in extended phase space.

We note that this result does not fully use the Hamiltonian structure; it relies only on the anti-symmetry of J. Thus, the resulting estimate is valid only over slow times of order one. Hamiltonian normal form theory will yield estimates that remain valid for much longer times [3]. For our purposes this is not required, but we make use of the explicit form of the slow vector field.

3. Optimal balance

On a conceptual level, optimal balance works by homotopically deforming the system into a simpler, e.g. linear system where the slow manifold is trivial to compute. Figure 1 provides a sketch in which the slow manifold at $t = 0$ is described by $p = 0$. The homotopy generates a surface of approximate slow manifolds in the extended phase space, illustrated by the green shaded surface. In general, for a frozen value of the homotopy parameter, the "manifold" $\mathcal{M}$ is only approximately invariant: trajectories drift away exponentially slowly with respect to the scale separation parameter ε. This is indicated by the dotted green line.

In the optimal balance, we identify the homotopy parameter with slowly varying time. In this case, the approximate slow manifold is an adiabatic invariant: a trajectory initially on the slow manifold will stay near it for a very long time while the manifold deforms slowly. Such a trajectory is shown in red in Figure 1. In this case, there are two sources of drift: on the one hand the drift already present for a frozen homotopy parameter. On the other hand, the drift due to the deformation of the manifold in time. In the following, we shall estimate both sources of drift.

Our task is to specify a single point on the approximate slow manifold $\mathcal{M}(T)$ by computing the fiber coordinate p^* for a given base-point coordinate q^*. In the extended phase space picture of Figure 1, this corresponds to specifying two boundary conditions: $q(T) = q^*$ and $p(0) = 0$. The first condition fixes the base-point. The second condition ensures that the entire trajectory remains near $\mathcal{M}(t)$

for all $t \in [0, T]$. We then define $p^* = p(T)$ as the computational approximation of the fiber coordinate.

For the prototype model (1), the procedure can be stated as follows. Take a smooth monotonic ramp function $\rho\colon [0,1] \to [0,1]$ with $\rho(0) = 0$ and $\rho(1) = 1$. For given $q^* \in \mathbb{R}^{2d}$, fix a ramp time $T > 0$ and solve the boundary value problem

$$\dot{q} = p\,, \tag{7a}$$

$$\dot{p} = Jp - \varepsilon\, \rho(t/T)\, \nabla V(q)\,, \tag{7b}$$

with boundary conditions

$$p(0) = 0 \qquad \text{and} \qquad q(T) = q^*\,. \tag{7c}$$

Then set $p^* = p(T)$.

We note that when the ramp parameter is frozen at $t = 0$, (7) reduces to the trivial linear fast-slow system $\dot{q} = p$ and $\dot{p} = Jp$, where p is fast and q is slow. This justifies the initial-time boundary condition $p(0) = 0$. We note that the boundary value $q(0)$ is not used explicitly in this setup.

In the following two sections, we analyze the accuracy of optimal balance by comparing against the slow vector field G_n associated with the original dynamical system (1), given by Theorem 1. We shall see that the asymptotic construction of the slow vector field for the ramped system (7) contains additional terms at $O(\varepsilon^{k+1})$ unless all derivatives of ρ up to order k vanish at the final time. Similarly, the description of the trivial slow manifold $p = 0$ differs from the description of the slow manifold for the ramped system (7) at $O(\varepsilon^{k+1})$ unless all derivatives of ρ up to order k vanish at the initial time. Thus, the order of accuracy of optimal balance is limited by the order of vanishing of derivatives of the ramp function at the end points.

4. Algebraic optimal balance

In this section, we consider the case when V or the ramp function ρ are only finitely differentiable. Then the best we can expect is an algebraic rate of convergence of optimal balance. The proof is a straightforward generalization of the classical fast-slow splitting used to prove Theorem 1 in [8].

Theorem 2. *For $n \in \mathbb{N}$, suppose $\rho \in C^{n+1}([0,1])$ with $\rho(0) = 0$ and $\rho(1) = 1$ satisfying the algebraic order condition*

$$\rho^{(i)}(0) = \rho^{(i)}(1) = 0 \tag{8}$$

for $i = 1, \dots, n$. Suppose further that $V \in C^{n+2}$. Fix $a > 0$ and consider a sequence of ramp times $T = a/\varepsilon$ and a sequence of solutions (q, p), implicitly parameterized by ε, to the boundary value problem (7). Then there exists a constant $c = c(\rho, a, n, V)$ such that

$$\|p(T) - G_n(q^*)\| \leq c\, \varepsilon^{n+2}\,. \tag{9}$$

Proof. By choosing appropriate units of time, we can take $a = 1$ without loss of generality. We then introduce the fast variable $w = p - \varepsilon\, F_{n+1}$, where

$$F_n(q, t) = \sum_{i=0}^{n} f_i(q, t)\, \varepsilon^i \tag{10}$$

with coefficients f_i to be determined. Then

$$\dot{q} = \varepsilon\, F_{n+1} + w\,, \tag{11a}$$

$$\dot{w} = (J - \varepsilon\, \mathrm{D}F_{n+1})\, w + \varepsilon\, (J F_{n+1} - \rho\, \nabla V) - \varepsilon^2\, \partial_\tau F_{n+1} - \varepsilon^2\, \mathrm{D}F_{n+1}\, F_{n+1} \tag{11b}$$

where, as before, $\tau = \varepsilon t$ so that $\partial_t = \varepsilon \partial_\tau$. We now eliminate the inhomogeneous term on the right of (11b) order by order up to an $O(\varepsilon^{n+2})$ remainder. This leads to the recursive expression

$$f_0 = -\rho\, J \nabla V(q)\,, \tag{12a}$$

$$f_k = -J\, \partial_\tau f_{k-1} - J \sum_{i+j=k-1} \mathrm{D}f_i(q)\, f_j(q) \tag{12b}$$

for $k = 1, \ldots, n+1$. We remark that for $\rho \equiv 1$, F_n reduces to G_n introduced in Theorem 1. Thus,

$$\dot{w} = (J - \varepsilon\, \mathrm{D}F_{n+1})\, w + O(\varepsilon^{n+2}) \tag{13}$$

so that left-multiplying by w implies

$$\frac{\mathrm{d}}{\mathrm{d}t} \|w\| \le \varepsilon\, \|\mathrm{D}F_{n+1}\|\, \|w\| + O(\varepsilon^{n+2})\,. \tag{14}$$

By assumption, $\partial_\tau^i \rho(0) = \rho^{(i)}(0) = 0$ so that $f_i(q,0) = 0$ for $i = 1, \ldots, n$. Recalling that $p(0) = 0$, we obtain

$$w(0) = p(0) - \varepsilon\, F_n(q(0), 0) - \varepsilon^{n+2}\, f_{n+1}(q(0), 0) = O(\varepsilon^{n+2})\,. \tag{15}$$

Hence, applying the Gronwall lemma to (14), we find that there exists $c = c(\rho, n, V)$ such that

$$\sup_{t \le T} \|p(t) - \varepsilon\, F_n(q(t), t)\| \le c\, \varepsilon^{n+2}\,. \tag{16}$$

Comparing with (4) and noting that $\rho^{(i)}(1) = 0$ for $i = 1, \ldots, n$, we see that

$$G_n(q^*) = \varepsilon\, F_n(q(T), T)\,. \tag{17}$$

The required estimate (9) follows. □

Corollary 3. *Suppose that, in the setting of Theorem 2, V is analytic and asymptotically strictly convex. Then for every desired order $n \in \mathbb{N}$ there exists a ramp function so that the method of optimal balance generates a state which remains balanced to $O(\varepsilon^{n+2})$ over times of $O(\exp(c/\varepsilon))$ under the dynamics of system* (1).

This result is a consequence of the uniqueness of the asymptotic expansion. More specifically, the fast variable $p - G_n$ in our construction and the fast variable p_ε in the Hamiltonian normal form setting of [3] coincide up to terms of $O(\varepsilon^{n+2})$. Thus, optimal balance at $O(\varepsilon^{n+2})$ in the sense of Theorem 2 is equivalent to $p_\varepsilon = O(\varepsilon^{n+2})$ in the notation of [3]. Since V is assumed analytic and asymptotic strict convexity of V implies that trajectories remain in a compact subset of phase space for all times, [3, Theorem 2.1] applies and yields persistent $O(\varepsilon^{n+2})$ smallness of the fast variable over exponentially long times.

5. Exponential optimal balance

In this section, we refine the result of Section 3 for the case when V is analytic and ρ is of Gevrey class 2.

Let us first recall that a function $f \in C^\infty(U)$ for $U \subset \mathbb{R}$ open is of Gevrey class s if there exist constants C and β such that

$$\sup_{x\in U} |f^{(n)}(x)| \leq C \, \frac{n!^s}{\beta^n} \tag{18}$$

for all $n \in \mathbb{N}$; see, e.g., [9]. We write $f \in G^s(U)$. Then the following is true.

Theorem 4. *Suppose $\rho \in G^2(0,1)$ with $\rho(0) = 0$ and $\rho(1) = 1$ satisfying the exponential order condition*

$$\rho^{(i)}(0) = \rho^{(i)}(1) = 0 \tag{19}$$

for all $i \in \mathbb{N}^*$. Fix $a > 0$, and consider a sequence of ramp times $T = a/\varepsilon$ and a sequence of solutions (q,p), implicitly parameterized by $\varepsilon \leq 1$, to the boundary value problem (7). Now suppose there exists a compact subset of phase space $\mathcal{K} \subset \mathbb{R}^{2d}$ containing this sequence of solution trajectories and that there exist $R > 0$ and $z_0 \in \mathbb{R}^{2d}$ with $\mathcal{K} \subset B_{R/2}(z_0)$ such that V is analytic on $B_R(z_0)$. Then there exist $n = n(\rho, a, V, \varepsilon) \in \mathbb{N}$ and positive constants $c = c(\rho, a, V)$ and $d = d(\rho, a, V)$ such that

$$\|p(T) - G_n(q^*)\| \leq d \, \mathrm{e}^{-c\varepsilon^{-\frac{1}{3}}} \, . \tag{20}$$

To prove this theorem, we proceed as in the proof of Theorem 2, albeit with a more careful estimate on the remainder term. Defining w as before, we write (13) in the form

$$\dot{w} = (J - \varepsilon \, \mathrm{D}F_{n+1}) \, w - R_{n+1} \, , \tag{21}$$

with explicit remainder

$$R_{n+1} = \varepsilon^{n+3} \, \partial_\tau f_{n+1} + \sum_{k=n+1}^{2(n+1)} \varepsilon^{k+2} \sum_{\substack{i+j=k \\ i,j \leq n+1}} \mathrm{D}f_i \, f_j \, . \tag{22}$$

The key observation is that each of the terms appearing in the expression for f_k, and each of the terms appearing in the expression for the remainders R_k, is a product of functions which depend only on ρ and functions which depend only on V. Hence, they can be written as inner products of coefficient vectors encoding all ρ-dependence with coefficient vectors encoding all V-dependence. A Hölder-like inequality will separate the two, so that we can estimate each class of coefficients separately in its respective norm.

To formalize this idea, we need to introduce some notation. We define the Cartesian product $\mathcal{F} \oplus \mathcal{G}$ of two vectors $\mathcal{F} = (\mathcal{F}^1, \dots, \mathcal{F}^N)$ and $\mathcal{G} = (\mathcal{G}^1, \dots, \mathcal{G}^M)$ as

$$\mathcal{F} \oplus \mathcal{G} = (\mathcal{F}^1, \dots, \mathcal{F}^N, \mathcal{G}^1, \dots, \mathcal{G}^M) \tag{23}$$

and the tensor product $\mathcal{A} \otimes \mathcal{G}$ of a vector of linear operators $\mathcal{A} = (\mathcal{A}^1, \dots, \mathcal{A}^N)$ acting on a vector $\mathcal{G} = (\mathcal{G}^1, \dots, \mathcal{G}^M)$ as

$$\mathcal{A} \otimes \mathcal{G} = (\mathcal{A}^1\mathcal{G}^1, \dots, \mathcal{A}^1\mathcal{G}^M, \dots, \mathcal{A}^N\mathcal{G}^1, \dots, \mathcal{A}^N\mathcal{G}^M) \, . \tag{24}$$

Further, we define the vector family $\{\mathcal{R}_k\}$ as

$$\mathcal{R}_0 = \rho\,, \tag{25a}$$

$$\mathcal{R}_{k+1} = \partial_\tau \mathcal{R}_k \oplus \bigoplus_{i+j=k} \mathcal{R}_i \otimes \mathcal{R}_j \quad \text{for} \quad k = 0, \dots, n\,, \tag{25b}$$

$$\mathcal{R}_{k+1} = \bigoplus_{\substack{i+j=k\\ i,j\leq n+1}} \mathcal{R}_i \otimes \mathcal{R}_j \quad \text{for} \quad k = n+1, \dots, 2n+2 \tag{25c}$$

where the components of $\mathcal{R}_j$ are acting on the components of $\mathcal{R}_l$ by multiplication and the indexed Cartesian product can be performed in any order as long as the order convention remains fixed throughout. Analogously, we define the family $\{\mathcal{F}_k\}$ as

$$\mathcal{F}_0 = -J\nabla V\,, \tag{26a}$$

$$\mathcal{F}_{k+1} = -\Bigl(\mathcal{F}_k \oplus \bigoplus_{j+l=k} \mathrm{D}\mathcal{F}_j \otimes \mathcal{F}_l\Bigr)\,\mathbb{J}_{k+1} \quad \text{for} \quad k = 0, \dots, n+1\,, \tag{26b}$$

$$\mathcal{F}_{k+1} = -\Bigl(\bigoplus_{\substack{i+j=k\\ i,j\leq n+1}} \mathrm{D}\mathcal{F}_j \otimes \mathcal{F}_l\Bigr)\,\mathbb{J}_{k+1} \quad \text{for} \quad k = n+2, \dots, 2n+2\,, \tag{26c}$$

where $\mathbb{J}_{k+1}$ denotes the block-diagonal matrix of matching dimension with J on the main diagonal.

As the recursive structure of the coefficient vectors mirrors the recursive structure in the definition of the f_k by (12), we can write

$$f_k = \sum_{i=1}^{N} \mathcal{R}_k^i\, \mathcal{F}_k^i \equiv \langle \mathcal{R}_k, \mathcal{F}_k \rangle\,. \tag{27}$$

Likewise, the remainder (22) takes the form

$$R_{n+1} = J \sum_{k=n+1}^{2(n+1)} \varepsilon^{k+2}\, \langle \mathcal{R}_{k+1}, \mathcal{F}_{k+1} \rangle\,. \tag{28}$$

We first consider the family of coefficient vectors $\mathcal{R}_k$. For a general $\mathcal{R} \equiv (\mathcal{R}^1, \dots, \mathcal{R}^N)$, we define the norm

$$|\mathcal{R}| = \max_{i=1,\dots,N} |\mathcal{R}^i|\,. \tag{29}$$

We then have the following estimate with respect to this norm.

Lemma 5. *Let $\rho \in G^2(0,1)$ with $C = 1$ and $\beta \leq 1$ in (18). Then*

$$|\mathcal{R}_k| \leq \frac{(k+1)!^2}{\beta^{k+1}}\,. \tag{30}$$

Proof. We recursively define a family of function classes via $\Gamma_1 = \{\rho\}$ and $r \in \Gamma_k$ for $k \geq 2$ if there exists a nonnegative integer $m \in \mathbb{N}$, a multi-index of length $s \in \mathbb{N}^*$ of strictly positive integers $\alpha \in (\mathbb{N}^*)^s$, and functions $r_j \in \Gamma_{\alpha_j}$ for $j = 1, \dots, s$ such that $k = m + |\alpha|$ and

$$r = \partial_\tau^m \prod_{j=1}^{s} r_j\,. \tag{31}$$

We note that the components of $\mathcal{R}_{k-1}$ are of class Γ_k. We shall show that $r \in \Gamma_k$ satisfies

$$\sup_{\theta\in(0,1)} |r(\theta)| \leq \frac{k!^2}{\beta^k} \,. \tag{32}$$

Due to the definition of the norm (29), this implies (30).

To prove (32), we proceed by induction on k. For $k = 1$, the statement is obvious. Now suppose $k \geq 2$, so that r has a decomposition of the form (31). We can also assume, without loss of generality, that when $s = 1$, $|\alpha| = \alpha_1 = 1$ and $m = k-1$. In this case, the statement is a direct consequence of the Gevrey class property (18). Now suppose that $s \geq 2$. Then, by the Leibniz rule,

$$|r(\theta)| = \left| \sum_{|\beta|=m} \binom{m}{\beta} \prod_{j=1}^{s} \partial_\tau^{\beta_j} r_j(\theta) \right| \leq \sum_{|\beta|=m} \binom{m}{\beta} \frac{(\alpha+\beta)!^2}{\beta^{|\alpha+\beta|}} \leq \frac{m!}{\beta^k} \frac{k!^2}{m!} \,, \tag{33}$$

where the first inequality uses the induction hypothesis and the second inequality is based on the observation that $|\alpha+\beta| = m + |\alpha| = k$ and a combinatorial inequality which is stated and proved as Lemma 10 in the Appendix. $\square$

We now turn to the family $\mathcal{F}_k$. We define the corresponding norms as follows. For $z_0 \in \mathbb{R}^{2d}$ fixed and arbitrary $r > 0$, let $B_r(z_0)$ denote the closed ball of radius r centered at z_0. For a vector field f on $\mathbb{R}^{2d}$, we write

$$\|f\|_r = \sup_{z\in B_r(z_0)} \|f(z)\| \tag{34}$$

and define a norm for $\mathcal{F} \equiv (\mathcal{F}^1, \dots, \mathcal{F}^N)$ via

$$\|\mathcal{F}\|_r = \sum_{i=1}^{N} \|\mathcal{F}^i\|_r \,. \tag{35}$$

We now prove a variant of Cauchy's estimate in this setting.

Lemma 6. *Let $r > s > 0$, and suppose the components of $\mathcal{F}$ and $\mathcal{G}$ are analytic on $B_r(z_0)$. Then*

$$\|\mathrm{D}\mathcal{F} \otimes \mathcal{G}\|_s \leq \frac{1}{r-s} \, \|\mathcal{F}\|_r \, \|\mathcal{G}\|_s \,. \tag{36}$$

Proof. Let h be any component of $\mathrm{D}\mathcal{F} \otimes \mathcal{G}$, i.e., there are components f, g of $\mathcal{F}$ and $\mathcal{G}$, respectively, such that $h = \mathrm{D}f \, g$. For fixed $z \in B_s(z_0)$, the function $\phi(t) = f(z + t\, g(z))$ is analytic for $|t| \leq \delta \equiv (r-s)/\|g\|_s$. Since $\phi'(0) = h$, the classical Cauchy estimate implies

$$\|h(z)\| = \|\phi'(0)\| \leq \frac{1}{\delta} \sup_{|t|\leq\delta} \|\phi(t)\| \leq \frac{1}{r-s} \, \|f\|_r \, \|g\|_s \,. \tag{37}$$

Estimate (36) then follows from the definition of the norm (35). $\square$

Lemma 7. *Let $z_0 \in \mathbb{R}^{2d}$, $R > 0$, and V be analytic on $B_{R(z_0)}$. Then there exist constants $C > 0$ and $\gamma > 0$ such that for any $n \in \mathbb{N}^*$ and $k \in \{0, \dots, 2n+3\}$,*

$$\|\mathcal{F}_k\|_{R/2} \leq C \left(\frac{n}{\gamma}\right)^k \,. \tag{38}$$

Proof. We set

$$\delta = \frac{R}{4n+6} \quad \text{and} \quad M = \max\biggl\{ \sup_{z \in B_R(z_0)} |\nabla V(z)|, \delta \biggr\} \,, \tag{39}$$

and recursively define the sequence (S_k) via

$$S_0 = 1 \,, \qquad S_{k+1} = S_k + \sum_{i+j=k} S_i \, S_j \tag{40}$$

which has the asymptotic behavior [5, pp. 474–475]

$$S_k \sim \frac{(3 - 2\sqrt{2})^{-k-\frac{1}{2}}}{2\sqrt{\pi k^3}} \,. \tag{41}$$

We will proceed to show that

$$\|\mathcal{F}_k\|_{R-\delta k} \leq \frac{M^{k+1}}{\delta^k} \, S_k \,. \tag{42}$$

The claimed estimate (38) is then a direct consequence of (42), (41), (39) and $\delta k \leq R/2$. Indeed, when $M = \delta$ in (39), then (38) holds with $\gamma = 3 - 2\sqrt{2}$. Otherwise, $M = \sup_{z \in B_R(z_0)} |\nabla V(z)|$, so that choosing $\gamma = R \, (3 - 2\sqrt{2})/(10 \, M)$ will suffice. The minimum of both provides an n-independent choice of γ, with similar considerations for C.

To prove (42), we proceed by recursion on k. For $k = 0$, the statement is trivial. Now suppose the result is proved up to index k. Then, by Lemma 6,

$$\begin{aligned}
\|\mathcal{F}_k\|_{R-\delta(k+1)} &\leq \|\mathcal{F}_k\|_{R-\delta(k+1)} + \frac{1}{\delta} \sum_{i+j=k} \|\mathcal{F}_i\|_{R-\delta k} \, \|\mathcal{F}_j\|_{R-\delta(k+1)} \\
&\leq \|\mathcal{F}_k\|_{R-\delta k} + \frac{1}{\delta} \sum_{i+j=k} \|\mathcal{F}_i\|_{R-\delta i} \, \|\mathcal{F}_j\|_{R-\delta j} \\
&\leq \frac{M^{k+2}}{\delta^{k+1}} \biggl(\frac{\delta}{M} \, S_k + \sum_{i+j=k} S_i \, S_j \biggr) \\
&\leq \frac{M^{k+2}}{\delta^{k+1}} \, S_{k+1} \,,
\end{aligned} \tag{43}$$

where the second inequality is due to the nesting of the balls over which the supremum is taken, the third inequality is due to the recursion hypothesis, and the last inequality uses $M \geq \delta$. □

Proof of Theorem 4. Without loss of generality, we assume that $a = 1$. Recalling the expression for the remainder in the form (28), noting that

$$\|\langle \mathcal{R}, \mathcal{F} \rangle\|_r \leq |\mathcal{R}| \, \|\mathcal{F}\|_r \,, \tag{44}$$

and referring to Lemma 5 and Lemma 7, we estimate

$$
\begin{aligned}
\|R_{n+1}\|_{R/2} &\leq \sum_{k=n+2}^{2n+3} \varepsilon^{k+2} \, |\mathcal{R}_k| \, \|\mathcal{F}_k\|_{R/2} \\
&\leq C \sum_{k=n+2}^{2n+3} \varepsilon^{k+2} \, \frac{(k+1)!^2}{\beta^{k+1}} \left(\frac{n}{\gamma}\right)^k \\
&\leq C_1 \, \varepsilon^2 \sum_{k=n+2}^{2n+3} \varepsilon^k \, \frac{n^{3k}}{\alpha^k} \\
&\leq C_1 \, \frac{\delta^{n+3}}{1-\delta} \, . \qquad (45)
\end{aligned}
$$

The third step is based on Stirling's inequality in the form $m! < \mathrm{e}^{m-1} \, m^{m+1/2}$ for every $m \geq 2$, the inequality $k+1 \leq 2n+3 \leq 5n$, and the observation that factors growing algebraically in k can always be absorbed by lowering α and adjusting the overall multiplicative constant. In the final step in (45) we have estimated the sum by the corresponding infinite geometric series under the assumption that $\delta \equiv \varepsilon \, n^3/\alpha < 1$ and $\varepsilon \leq 1$. Let us now choose

$$
n = \left\lfloor \left(\frac{\alpha\delta}{\varepsilon} \right)^{\frac{1}{3}} \right\rfloor . \qquad (46)
$$

Then

$$
\|R_{n+1}\|_{R/2} \leq \frac{C_1}{1-\delta} \, \delta^{(\alpha\delta/\varepsilon)^{\frac{1}{3}}} \leq C_2 \, \mathrm{e}^{-c\varepsilon^{-\frac{1}{3}}} \, , \qquad (47)
$$

where, in the last inequality, we have fixed $\delta \in (0,1)$ so that $c > 0$. Now following the same steps as in the proof of Theorem 2 and using assumption (19) at $t = 0$, we observe that $w(0) = 0$ so that there exists a constant $C_3(T)$ such that

$$
\sup_{t \leq T} \|p(t) - \varepsilon \, F_n(q(t),t)\| \leq C_3(T) \, \mathrm{e}^{-c\varepsilon^{-\frac{1}{3}}} \, . \qquad (48)
$$

Using assumption (19) now at $t = T$, we verify that

$$
G_n(q^*) = \varepsilon \, F_n(q(T),T) \qquad (49)
$$

and the required estimate follows. □

Remark 1. While it is possible to find ramp functions in G^s for any $s > 1$ satisfying the exponential order condition (19), the proof as stated will generalize only to Gevrey classes $s > 2$, as the required generalization of Lemma 9,

$$
\sum_{m=0}^{n} (m+\ell)!^{s-1} \, (n+k-m-\ell)!^{s-1} \leq (n+k)!^{s-1} \qquad (50)
$$

fails for $s < 2$. For $s > 2$, the final estimate reads

$$
\|p(T) - G_n(q^*)\| \leq d \, \mathrm{e}^{-c\varepsilon^{-\frac{1}{s+1}}} \, . \qquad (51)
$$

As this is weaker than (20), this generalization is of little interest, in particular since a suitable ramp function in G^2 is easily available; see (55) below.

As in Section 4, we can combine our result with the long-time invariance of approximate balance provided by [3, Theorem 2.1] as follows.

Corollary 8. *In the setting of Theorem 4, suppose V is analytic and strictly convex. Then the method of optimal balance generates a state which remains balanced to $O(\exp(-c/\varepsilon^{1/3}))$ over times of $O(\exp(c/\varepsilon))$ under the dynamics of system* (1).

6. Numerical Tests

A direct numerical demonstration of Theorems 2 and 4 is impossible as we do not have direct access to the reference slow vector field G_n. We thus resort to computing the following proxy for the balance error.

(1) Given q^*, compute the corresponding p^* via optimal balance.
(2) Evolve the full system (1), initialized with $q(0) = q^*$ and $p(0) = p^*$, forward up to some time t_1 which is fixed independent of ε on the slow time scale. (For the simulations shown below, $\varepsilon t_1 = 0.5$.)
(3) "Rebalance" the evolved state, i.e., find a p_1^* via optimal balance for the given $q_1^* = q(t_1)$.
(4) Define the diagnosed imbalance as $I = \varepsilon^{-1} \, \|p(t_1) - p_1^*\|$.

We note that the diagnosed imbalance I is not a direct measure of the imbalance error $\|p(T) - G_n(q^*)\|$. On the one hand, I may be overestimating imbalance because during the forward simulation of model (1), there is a slow drift off the slow manifold. However, since t_1 is taken to be small, this contribution is small as well as asymptotically subdominant. A more serious question is whether I may underestimate the imbalance, because re-balancing may simply reproduce the same bias committed during the initial balancing. Since imbalanced motion is oscillatory on the fast time scale, we would expect that the diagnosed imbalance oscillates on the fast time scale as a function of t_1, so that the amplitude of this oscillation can be taken as a measure of imbalance. However, we did a careful pre-study which showed that I depends almost monotonically on t_1. Thus, simply looking at the imbalance for fixed t_1 already gives robust results. Moreover, as we shall see, the diagnosed imbalance reproduces the predictions of Theorem 2 accurately. This gives strong empirical support to the idea that I is a useful proxy for imbalance which could also be used in more complex situations, e.g., for the study of rotating fluids.

In our proof-of-concept implementation, we use the potential

$$V(q) = \tfrac{3}{4}\, q_1^4 + \tfrac{1}{4}\, q_2^4 \,. \tag{52}$$

and solve the boundary value problem (7) by simple shooting with an off-the-shelf ODE solver and root finder. More efficient implementations would use multiple shooting and possibly a symplectic time-discretization. The ramp functions are of the form

$$\rho(\theta) = \frac{f(\theta)}{f(\theta) + f(1-\theta)} \,, \tag{53}$$

where

$$f(\theta) = \theta^k \tag{54}$$

for different exponents k, or $f(\theta) = \exp(-1/\theta)$, so that

$$\rho(\theta) = \frac{e^{-1/\theta}}{e^{-1/\theta} + e^{-1/(1-\theta)}} \,. \tag{55}$$

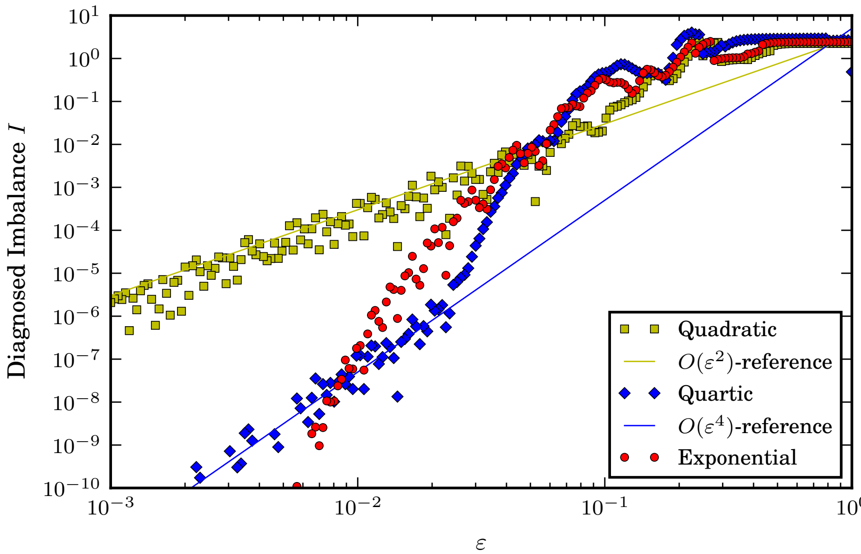

FIGURE 2. Diagnosed imbalance I as a function of ε for different ramp functions. "Quadratic," "quartic," and "exponential" refer to the ramp functions (53) with $f(\theta) = x^2$, $f(\theta) = x^4$, and $f(\theta) = \exp(-1/\theta)$, respectively. The ramp time is $T = 2/\varepsilon$.

The ramp function (55) satisfies the exponential order condition and is of Gevrey class 2, thus it satisfies the assumptions of Theorem 4.[1]

In Figure 2, we compare the performance of ramp functions satisfying different order conditions. For two algebraic ramp functions with $k = 2$ and $k = 4$ in (54) corresponding to $n = 1$ and $n = 3$ in the algebraic order condition of Theorem 2, the predicted respective quadratic and quartic decay of imbalance is clearly visible. The super-algebraic decay of imbalance for the ramp function with exponential order condition is seen as a convex-shaped curve in the log-log plot of I vs. ε.

In Figure 3, we explore the dependence of the diagnosed imbalance I on the ramp time T for the exponential ramp function case. For a given value of ε, longer ramp times yield smaller diagnosed imbalances. A rigorous study goes beyond the Theorems proved here.

Figure 4 shows the same data as Figure 3, but with a doubly logarithmic vertical axis. Assuming that the diagnosed imbalance behaves in the general form suggested by Theorem 4, i.e., if

$$I = d\,\mathrm{e}^{-c\varepsilon^{-\alpha}}\,, \tag{56}$$

then

$$\ln(\ln d - \ln I) = \ln c - \alpha\,\ln\varepsilon\,. \tag{57}$$

Then, plotting $\ln(-\ln I)$ vs. $\ln\varepsilon$ should asymptote to a line of slope $-\alpha$. The data points show such behavior for a good range of small values of ε before the accuracy

[1]Indeed, each of the terms appearing in (55) are of class G^2, see Lemma 11. As Gevrey classes are vector spaces, the denominator of (55) is also of class G^2. Finally, nonsingular quotients of G^2-functions are again of class G^2, see Lemma 12.

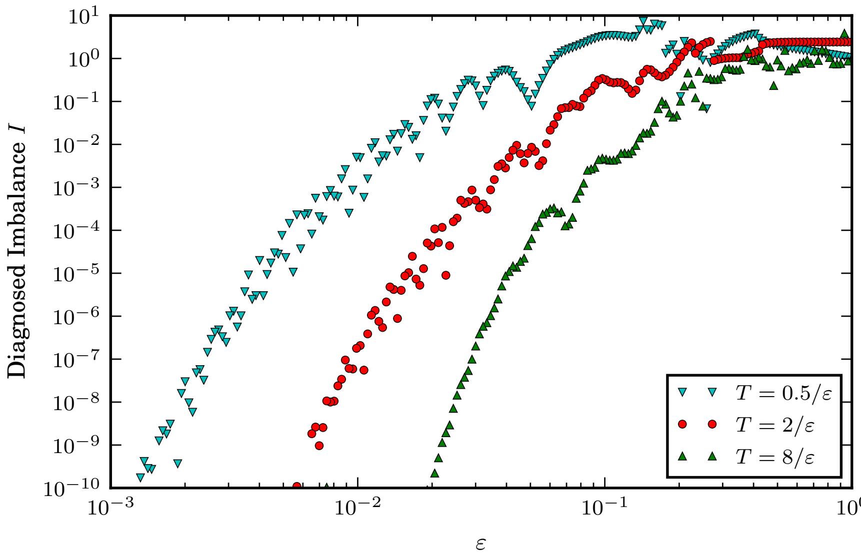

FIGURE 3. Diagnosed imbalance I as a function of ε when taking the exponential ramp function (53) with $f(\theta) = \exp(-1/\theta)$ for three different ramp times.

of the time integrator and root solver, controlled to be at least 10^{-10}, breaks down. The observed behavior is better than $\alpha = 1/3$ obtained in the bounds of Theorem 4, but depends on the ramp time. For large ramp times, the error is dominated by the derivatives of the potential V, and the estimated exponent comes close to $\alpha = 1$, which is the exponent expected from the usual exponential asymptotics [3]. For shorter ramp times, the influence of the ramp function becomes more important so that the estimated exponent decreases, but not to as low as the theoretical bounds.

7. Discussion

Our results show, in the context of a simple finite dimensional Hamiltonian model problem, that the method of optimal balance yields a state which is exponentially close to a balanced state obtained by optimal truncation of an asymptotic series describing the approximate slow manifold. We believe that similar results will apply to more general Hamiltonian fast-slow systems in the absence of resonances.

The results give strong support to the notion that optimal balance may in fact be the best practically available characterization of a slow manifold in this context. As optimal truncation of an asymptotic series is not computationally feasible, optimal balance could therefore be used as a computable definition of a balanced state (this idea has in fact been proposed earlier by McIntyre [17]).

However, a number of questions remain open. An obvious question is the sharpness of the analysis, both in terms of the current restriction to ramp functions in Gevrey classes G^s for $s \geq 2$, and in terms of the exponent α in the imbalance scaling (56). A more practical concern is the best choice of ramp time T for fixed ε. Our analysis concerns only the scaling with respect to ε, but the structure of the estimates, along with the numerical results, suggests that at least initially the

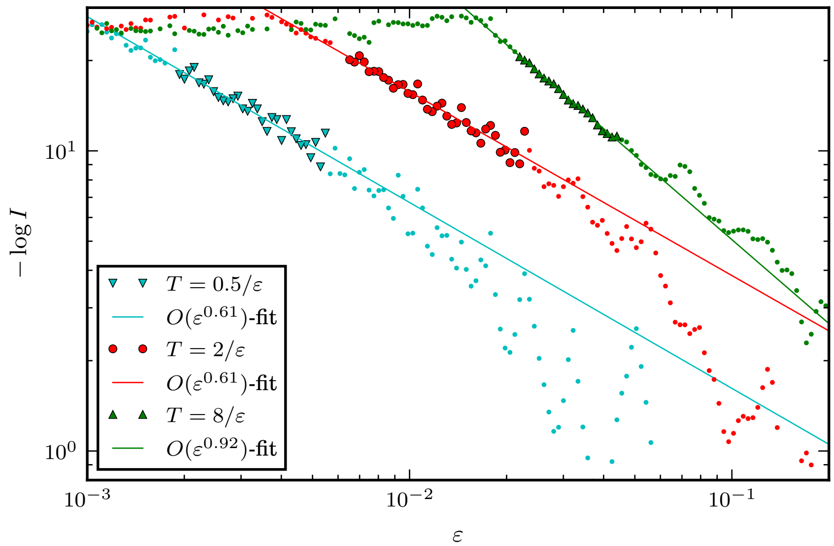

FIGURE 4. The same data as in Figure 3, shown on a doubly logarithmic vertical axis. This allows a least squares fit to determine the power of ε in the exponent of the expression for the exponential convergence rate, see equation (57). The linear least squares fit was performed over a finite interval in ε, indicated by the larger marker symbols.

results improve when the ramp time is increased. This, however, cannot go on forever because beyond some ramp time T_{opt}, the imbalance due to the drift off the approximate manifold will dominate and imbalance will increase as T is increased further. How to design an adaptive algorithm which chooses an optimal ramp time automatically is entirely open.

Whereas optimal balance has been successfully used in geophysical fluid dynamics, the theory presented here was developed only for finite-dimensional Hamiltonian systems. It is therefore a natural question how our results translate to infinite-dimensional Hamiltonian systems. A direct generalization of the model (1) is the semilinear Klein–Gordon equation in the non-relativistic limit (e.g. [24]). In general, obtaining results on approximate slow manifolds for infinite dimensional Hamiltonian systems is difficult since unbounded operators may destroy the scale separation and slow-fast or fast-fast resonances may emerge. Existing results either apply to special solutions (e.g. [12]), bounded slow subsystems (e.g. [10]), or require spatial analyticity of solutions (e.g. [16]). Finding the right analytical setting for the semi-linear Klein–Gordon equation is a subject of ongoing research.

The question of justification of optimal balance in geophysical flow problems is even more difficult, although our main motivation and reported successful implementations come from this area. Further, the question of efficient implementation, in terms of run-time and in terms of coding effort, is of considerable practical relevance. For the toy model considered here, we were able to solve the optimal balance system problem (7) by simple shooting. However, this might fail or become excessively expensive in higher dimensions.

Sophisticated boundary value solvers may be needed but are hard to implement and computationally costly. We remark that we have provided only an approximate iterative solution of the boundary-value-problem (7) while the issue of well-posedness of this boundary-value problem was not addressed. Viúdez and Dritschel [26] suggest an iterative procedure in which one integrates back and forth, resetting to the correct boundary condition at each end. Empirically, their approach converges well in the geophysical fluid dynamics context of their study. The iterative back-and-forth integrations can be understood as nudging toward the boundary-values. For linear systems, back-and-forth nudging can be rigorously proven to converge to the true solution [1]; the problem considered here is, in our understanding, not directly covered by these results but we expect that a proof could be obtained with reasonable effort. In our concrete simulations, shooting was slightly more efficient than back-and-forth nudging and converged for a moderately larger set of parameters. Finding the best strategy is an open problem.

Appendix A. Combinatorial estimates

In the following, we prove an estimate on the combinatorial constants which appear in the proof of Theorem 4. This result is stated as Lemma 10 below. We begin with a special case which is needed in the proof of the general result.

Lemma 9. *Let $n \in \mathbb{N}$ and $k, \ell \in \mathbb{N}^*$ with $1 \leq \ell < k$. Then*

$$\sum_{m=0}^{n} (m+\ell)!\,(n+k-m-\ell)! \leq (n+k)!\,. \tag{58}$$

Proof. We proceed by induction on n. For $n = 0$, the statement is obvious. Now suppose the statement is true up to step $n-1$. Then

$$\begin{aligned}\sum_{m=0}^{n} (\ell+m)!\,(n+k-\ell-m)! &= (\ell+n)!\,(k-\ell)! + \sum_{m=0}^{n-1} (\ell+m)!\,(n+k-\ell-m)! \\ &\leq (n+k-1)! + (n+k-1) \sum_{m=0}^{n-1} (\ell+m)!\,(n-1+k-\ell-m)! \\ &\leq (n+k-1)! + (n+k-1)\,(n-1+k)! = (n+k)! \end{aligned} \tag{59}$$

where the first inequality is due to $1 \leq \ell < k$ and the second inequality uses the induction hypothesis. □

Lemma 10. *Let $n \in \mathbb{N}$, and $s \in \mathbb{N}^*$. Then for a multi-index of strictly positive integers $\alpha \in (\mathbb{N}^*)^s$ with $|\alpha| = k$,*

$$\sum_{|\beta|=n} \frac{(\alpha+\beta)!^2}{\beta!} \leq \frac{(n+k)!^2}{n!}\,, \tag{60}$$

where the sum is over multi-indices β of length s.

Proof. We proceed by induction on s. For $s = 1$, the two sides of (60) are trivially equal. Now suppose the statement holds true up to step $s-1$. We write $\alpha = (\alpha', \ell)$

and $\beta = (\beta', m)$, where α' and β' are multi-indices of length $s-1$, and ℓ and k are integers satisfying $1 \leq \ell < k$. Then

$$
\begin{aligned}
\sum_{|\beta|=n} \frac{(\alpha+\beta)!^2}{\beta!} &= \sum_{m=0}^{n} \frac{(\ell+m)!^2}{m!} \sum_{|\beta'|=n-m} \frac{(\alpha'+\beta')!^2}{\beta'!} \\
&\leq \sum_{m=0}^{n} \frac{(\ell+m)!^2}{m!} \, \frac{(n-m+k-\ell)!^2}{(n-m)!} \\
&= \frac{(n+k)!}{n!} \sum_{m=0}^{n} \binom{n}{m} \binom{n+k}{m+\ell}^{-1} (m+\ell)! \, (n+k-m-\ell)! \, . \qquad (61)
\end{aligned}
$$

As the ratio of the binomial coefficients that appear in the right hand sum is always bounded above by 1, the proof is achieved by Lemma 9. □

Appendix B. G^2-estimates on the exponential ramp function

The following two results are necessary to show that the exponential ramp function (55) used in the numerical experiments above is of Gevrey class 2. We believe that the results are classical; Lemma 11, for example, is stated without proof in [9, p. 218]. In this appendix, we give complete proofs for the convenience of the reader.

Lemma 11. *The function*

$$
f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \exp(-1/x) & \text{for } x > 0 \end{cases} \qquad (62)
$$

is of Gevrey class 2 *uniformly in* $\mathbb{R}$.

Proof. The function f is holomorphic in the right complex half-plane. Fixing $\lambda \in (0, \frac{1}{2})$, the Cauchy integral formula

$$
f^{(n)}(x) = \frac{n!}{2\pi i} \int_\gamma \frac{f(z)}{(z-x)^{n+1}} \, \mathrm{d}z \qquad (63)
$$

applies in particular when γ is a circle of radius λx centered at x. We parameterize γ writing $z(\theta) = x + \lambda x \, w(\theta)$ where $w(\theta)$ is an arc-length parameterization of the unit circle. Then

$$
\begin{aligned}
|f^{(n)}(x)| &\leq \frac{n!}{2\pi \, (\lambda x)^{n+1}} \int_0^{2\pi} |f(z(\theta))| \, \mathrm{d}\theta \\
&\leq \frac{n!}{(\lambda x)^{n+1}} \sup_{\theta \in [0,2\pi]} \left| \exp\biggl(-\frac{1 + \lambda \, \overline{w}(\theta)}{x \, |1 + \lambda \, w(\theta)|^2} \biggr) \right| \\
&\leq \frac{n!}{(\lambda x)^{n+1}} \exp\biggl(-\frac{1-\lambda}{x \, |1+\lambda|^2} \biggr) \, . \qquad (64)
\end{aligned}
$$

Maximizing the right hand side with respect to x and using Sterling's inequality in the form $m^m \, \mathrm{e}^{-m} \leq m!$, we obtain

$$
\sup_{x \in \mathbb{R}} |f^{(n)}(x)| \leq \frac{(n+1)!^2}{\eta^{n+1}} \qquad (65)
$$

with $\eta = \lambda(1-\lambda)/(1+\lambda)^2$. This proves that f is of Gevrey class 2. □

Lemma 12. *Let $U \subset \mathbb{R}$ be open and suppose $f, g \in G^2(U)$ with $g \geq c > 0$ for some constant c. Then $h = f/g \in G^2(U)$.*

Proof. Without loss of generality, assume that $f \leq 1$ and $g \geq 1$ on U, so that $h \leq 1$. Further, let α denote the smaller of the two parameters appearing in the denominator of the Gevrey class estimates (18) of f and g. Set $\beta = \alpha/3$. Using the Leibniz rule for the nth derivative of the product gh and rearranging terms, we have

$$h^{(n)} = \frac{1}{g} \left(f^{(n)} - \sum_{j=0}^{n-1} \binom{n}{j} g^{(n-j)} \, h^{(j)} \right). \tag{66}$$

We now proceed by induction on n. For $n = 0$, the statement is obvious. Now suppose that h satisfies a Gevrey class estimate of the form (18) with parameter β up to order $n - 1$. Then

$$|h^{(n)}(x)| \leq \frac{n!^2}{\alpha^n} + n! \sum_{j=0}^{n-1} \frac{(n-j)!}{\alpha^{n-j}} \, \frac{j!}{\beta^j} \leq \frac{n!^2}{\alpha^n} \left(1 + \frac{3^{n-1}}{n!} \sum_{j=0}^{n-1} (n-j)! \, j! \right) \leq \frac{n!^2}{\beta^n} \,, \tag{67}$$

where the last inequality is based on the observation that

$$\sum_{j=0}^{n-1} (n-j)! \, j! = n! + \sum_{j=1}^{n-1} (n-j)! \, j! \leq n! + (n-1) \, (n-1)! \leq 2 \, n! \tag{68}$$

and further that $1 + 2 \cdot 3^{n-1} \leq 3^n$. □

Acknowledgments

We thank David Dritschel for many useful discussions on balance and optimal potential vorticity balance. The numerical study is based on prior work done by Zekun Yang as part of her Bachelor thesis at Jacobs University. This paper contributes to the project "The interior energy pathway: internal wave emission by quasi-balanced flows" of the Collaborative Research Center TRR 181 "Energy Transfers in Atmosphere and Ocean" funded by the German Research Foundation. Funding through the TRR 181 is gratefully acknowledged. HM and MO further acknowledge funding by German Research Foundation grant OL-155/6-1.

References

[1] D. Auroux and J. Blum, *Back and forth nudging algorithm for data assimilation problems*, C. R. Acad. Sci. Paris, Ser. I, 340 (2005), pp. 873–878.

[2] C. Cotter, *Data assimilation on the exponentially accurate slow manifold*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 371 (2013), pp. 20120300, 11.

[3] C. J. Cotter and S. Reich, *Semigeostrophic particle motion and exponentially accurate normal forms*, Multiscale Model. Simul., 5 (2006), pp. 476–496 (electronic).

[4] M. J. P. Cullen, *A comparison of numerical solutions to the eady frontogenesis problem*, Quart. J. R. Meteorol. Soc., 134 (2008), pp. 2143–2155.

[5] P. Flajolet and R. Sedgewick, *Analytic combinatorics*, Cambridge University Press, 2009.

[6] J. E. Frank and G. A. Gottwald, *Stochastic homogenization for an energy conserving multi-scale toy model of the atmosphere*, Phys. D, 254 (2013), pp. 46–56.

[7] G. A. Gottwald, *Controlling balance in an ensemble Kalman filter*, Nonlinear Process. Geophys., 21 (2014), pp. 417–426.

[8] G. A. Gottwald and M. Oliver, *Slow dynamics via degenerate variational asymptotics*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 470 (2014), p. 20140460.

[9] K. Itô, *Encyclopedic Dictionary of Mathematics*, vol. 1, MIT Press, 1993.

[10] K. U. Kristiansen and C. Wulff, *Exponential estimates of symplectic slow manifolds*, J. Differential Equations, 261 (2016), pp. 56–101.
[11] A. Leung and K. Meyer, *Adiabatic invariants for linear Hamiltonian systems*, J. Differential Equations, 17 (1975), pp. 32–43.
[12] N. Lu, *Small generalized breathers with exponentially small tails for Klein-Gordon equations*, J. Differential Equations, 256 (2014), pp. 745–770.
[13] P. Lynch, *The Emergence of Numerical Weather Prediction: Richardson's Dream*, Cambridge University Press, 2006.
[14] P. Lynch and X.-Y. Huang, *Initialization of the HIRLAM model using a digital filter*, Mon. Weather Rev., 120 (1992), pp. 1019–1034.
[15] R. S. MacKay, *Slow manifolds*, in Energy Localisation and Transfer, T. Dauxois, A. Litvak-Hinenzon, R. S. MacKay, and A. Spanoudaki, eds., World Scientific, 2004, pp. 149–192.
[16] K. Matthies and A. Scheel, *Exponential averaging for Hamiltonian evolution equations*, Trans. Amer. Math. Soc., 355 (2003), pp. 747–773.
[17] M. E. McIntyre, *Spontaneous imbalance and hybrid vortex-gravity structures*, J. Atmos. Sci., 66 (2009), pp. 1315–1326.
[18] M. E. McIntyre, *Balanced flow*, in Encyclopedia of Atmospheric Sciences, J. Pyle and F. Zhang, eds., Academic Press, second ed., 2015, pp. 298–303.
[19] L. J. Neef, S. M. Polavarapu, and T. G. Shepherd, *Four-dimensional data assimilation and balanced dynamics*, J. Atmos. Sci., 63 (2006), pp. 1840–1858.
[20] A. I. Neĭshtadt, *On the accuracy of conservation of the adiabatic invariant*, J. Appl. Math. Mech., 45 (1981), pp. 58–63.
[21] ———, *The separation of motions in systems with rapidly rotating phase*, Prikl. Mat. Mekh., 48 (1984), pp. 197–204.
[22] N. Nekhoroshev, *An exponential estimate of the time of stability of a nearly-integrable Hamiltonian system*, Russ. Math. Surv., 32 (1977), pp. 1–65.
[23] M. Oliver, *Variational asymptotics for rotating shallow water near geostrophy: a transformational approach*, J. Fluid Mech., 551 (2006), pp. 197–234.
[24] M. Tsutsumi, *Nonrelativistic approximation of nonlinear Klein-Gordon equations in two space dimensions*, Nonlinear Anal., 8 (1984), pp. 637–643.
[25] J. Vanneste, *Balance and spontaneous wave generation in geophysical flows*, Ann. Rev. Fluid Mech., 45 (2013), pp. 147–172.
[26] Á. Viúdez and D. G. Dritschel, *Optimal potential vorticity balance of geophysical flows*, J. Fluid Mech., 521 (2004), pp. 343–352.
[27] J.-S. von Storch, G. Badin, and M. Oliver, *The interior energy pathway: inertial gravity wave emission by oceanic flows*, in Energy Transfers in Atmosphere and Ocean, C. Eden and A. Iske, eds., Springer-Verlag. to appear.

(G.A. Gottwald) School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia

(H. Mohamad) School of Engineering and Science, Jacobs University, 28759 Bremen, Germany

(M. Oliver) School of Engineering and Science, Jacobs University, 28759 Bremen, Germany