Heat Transport in Reduced Order Convection Models

by

Matthew L. Olson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Applied and Interdisciplinary Mathematics)
in the University of Michigan
2020

Doctoral Committee:

Professor Charles R. Doering, Co-Chair
Professor William W. Schultz, Co-Chair
Professor Silas D. Alben
Professor Anthony M. Bloch
Professor Peter D. Miller

Matthew L. Olson

mlolson@umich.edu

ORCID iD: 0000-0003-0898-1993

Acknowledgments

I would first like to thank my advisors, Charles Doering and William Schultz, for their guidance, wisdom, and support. Throughout the past three and a half years, Charles Doering has provided a consistent positive influence and his insightful suggestions have been instrumental in shaping my research. William Schultz likewise has been a steadfast mentor, and has improved my understanding by challenging me to approach topics from different perspectives.

I would also like to thank Silas Alben, Anthony Bloch, and Peter Miller for taking the time to serve on my dissertation committee, and for their thought-provoking feedback on my research.

Next I would like to thank David Goluskin for his ideas that inspired this work, as well as the substantial advice he has given me along the way. I am also grateful for the many valuable discussions with my colleagues, Alex Vargo and Daniel Irvine (both mathematical and otherwise).

I am grateful for all of the teachers who have influenced me over the years. There are too many people to mention here but I would especially like to thank Stephen Gent, Peter Miller, Charles Doering, and William Schultz.

Finally, I want to express my appreciation to my parents for their support throughout my life. I would not be where I am today without their patience, guidance, and steady encouragement.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

ABSTRACT

Reduced order models (ROMs) are finite systems of ordinary differential equations (ODEs) that approximate the dynamics of the governing partial differential equations. This work considers ROMs for Rayleigh's 1916 mathematical model of thermal convection. A hierarchy of reduced models is developed that satisfy energy, temperature, and vorticity balance laws to promote agreement with the physics of thermal convection. These balance laws are generalized from idealized versions considered by previous authors, and new criteria are established for the vorticity balance in the general case. Each model in the hierarchy is an extension of the Lorenz equations and includes Fourier modes capable of producing zonal flow—horizontal mean flow that vertically shears the fluid.

Upper bounds on time-averaged heat transport are obtained for several models in the hierarchy. Bounds for the ODE models are derived by constructing auxiliary functions such that certain polynomial expressions are nonnegative. Nonnegativity is enforced by requiring these polynomial expressions to admit sum-of-squares representations. Polynomial auxiliary functions subject to such constraints can be optimized computationally with semidefinite programming, minimizing the resulting bound. Upper bounds are compared to particular solutions to the ODEs obtained using bifurcation analysis and numerical integration. An eight-ODE model in the hierarchy is explored in detail, revealing sharp or nearly sharp bounds on mean heat transport for numerous values of the model parameters, the Rayleigh and Prandtl numbers and the domain aspect ratio. In all cases where the Rayleigh number is small enough for the ODE models to closely approximate the physics of the governing equations, mean heat transport is maximized by the steady states that emerge from the first instability of the static state. These equilibria do not exhibit zonal flow, suggesting that this type of flow does not enhance heat transport. Analytical parameter-dependent bounds are derived for the eight-ODE model with quadratic auxiliary functions, and they are sharp for sufficiently small Rayleigh numbers.

CHAPTER 1

Rayleigh–Bénard Convection

1.1 Introduction

Thermal convection is a fundamental process in fluid dynamics, consisting of buoyancy-driven flow in the presence of an imposed temperature gradient. Convection occurs when warmer, less dense fluid lies below cooler, denser fluid and the resulting buoyancy forces are large enough to exceed the resistance provided by the fluid's viscosity. This process is realized in a wide variety of physical systems. Atmospheric convection, induced by radiative heating from the Sun, drives global wind and weather patterns [82]. Temperature gradients inside the Earth cause large-scale convection in the mantle, producing continental drift and generating the Earth's magnetic field [40]. Heat is transported towards the outer layers of stars in the convection zone [69]. Convective heat transport is also utilized in a variety of engineering applications.

The basic mathematical model describing natural thermal convection is Rayleigh–Bénard convection, and was first proposed by Lord Rayleigh [58]. In Rayleigh's original model, a fluid is contained between two infinite-length, horizontal, impermeable boundaries, and a temperature gradient is imposed within the fluid by fixing the temperatures along the top and bottom of the domain, with the higher temperature along the lower boundary. When the temperature difference is sufficiently small, the physics are simple: the fluid remains still, and thermal energy is transported only by conduction. At higher temperatures, buoyancy overcomes viscous damping, and the fluid develops a regular pattern of counter-rotating "convection cells" that actively transfer heat from the lower boundary. As the temperature is raised further, the flow in the bulk eventually becomes turbulent.

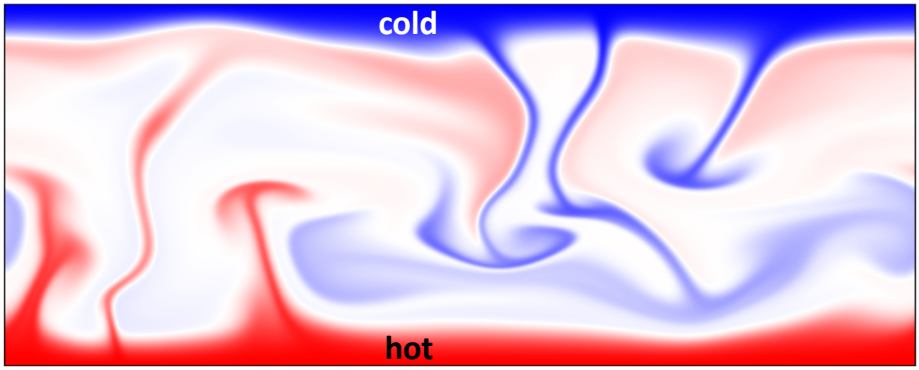


Figure 1.1: Temperature profiles for turbulent Rayleigh–Bénard convection at one instant in time obtained by numerical simulations performed by Hannah L. Swan, as depicted in [12]. At high temperatures, hot and cold plumes emerge from thermal boundary layers and enhance heat transport across the domain.

1.2 The Boussinesq equations

Consider a Newtonian fluid with density ρ and dynamic viscosity μ, subject to gravitational acceleration g aligned with the vertical (z) axis. The velocity, $\mathbf{v}(\mathbf{x}, t)$, and pressure, $p(\mathbf{x}, t)$, are governed by the Navier–Stokes equations [14]:

$$(\partial_t + \mathbf{v} \cdot \nabla)(\rho \mathbf{v}) = -\nabla p + \mu \nabla^2 \mathbf{v} - \rho g \hat{\mathbf{z}} \tag{1.1}$$

$$\partial_t \rho + \nabla \cdot (\rho \mathbf{v}) = 0. \tag{1.2}$$

Suppose the fluid is encased above and below by parallel impermeable walls. A temperature gradient is supplied by heating the lower boundary to a temperature T_b, while the upper boundary is held fixed at a temperature T_t. The temperature $T(\mathbf{x}, t)$ of the fluid is governed by an advection-diffusion equation [14], with diffusion modulated by the coefficient of thermal diffusivity, κ:

$$\partial_t T + \nabla \cdot (\mathbf{v} T) = \kappa \nabla^2 T. \tag{1.3}$$

Rayleigh considered fluid dynamics in the Boussinesq approximation[1], where density variations are considered negligible except in the term representing the buoyancy force, $\rho g \hat{\mathbf{z}}$. In this term, the density takes the form $\rho = \rho_0 (1 - \alpha(T - T_b))$ where ρ_0 is the reference density and α is the coefficient of thermal expansion. On the left-hand sides of (1.1) and (1.2), the density is replaced by the reference density ρ_0. Under the Boussinesq approximation,

[1]The Boussinesq approximation [70], or Oberbeck–Boussinesq approximation, is commonly employed in studies of Rayleigh–Bénard convection when density variations are small in comparison to the reference density of the fluid.

the material parameters ρ_0, μ, α and κ are treated as constants. Making these simplifications reduces the Navier–Stokes equations to

$$\partial_t \mathbf{v} + \mathbf{v} \cdot \nabla \mathbf{v} = -\frac{1}{\rho_0} \nabla p + \nu \nabla^2 \mathbf{v} - (1 - \alpha(T - T_b)) g \, \hat{\mathbf{z}}, \tag{1.4}$$

$$\nabla \cdot \mathbf{v} = 0, \tag{1.5}$$

$$\partial_t T + \mathbf{v} \cdot \nabla T = \kappa \nabla^2 T, \tag{1.6}$$

where $\nu := \mu / \rho_0$ is the kinematic viscosity.

Rayleigh–Bénard convection can be modeled in two or three spatial dimensions, but the focus of this thesis is on two-dimensional convection. This case is of interest in part due to its simplicity. As the temperature gradient of a fluid at rest is increased, the first instability that develops is two-dimensional [45]. Even when the flow becomes turbulent so that three-dimensional flow is expected, simplifying the problem to two dimensions may help to understand the mechanisms underlying turbulent convection. We therefore consider a fluid contained in a rectangular domain $(x, z) \in [0, \pi A d] \times [0, \pi d]$ of height πd and aspect ratio A. The velocity field in two dimensions is $\mathbf{v} = u\hat{\mathbf{x}} + w\hat{\mathbf{z}}$ and equation (1.5) implies that the flow is incompressible. Two-dimensional incompressible flows admit a stream function whose spatial derivatives are the components of the velocity field. Here we define the stream function as

$$\partial_z \psi := u, \quad \partial_x \psi := -w. \tag{1.7}$$

Assume that ψ has continuous second order partial derivatives so that the continuity equation (1.5) is automatically satisfied. The evolution of ψ is determined by taking the curl of the momentum equation (1.4). Letting the Jacobian of two functions be given by $\{f, g\} = \partial_x f \partial_z g - \partial_z f \partial_x g$, (1.4) becomes

$$\partial_t \nabla^2 \psi - \{\psi, \nabla^2 \psi\} = \nu \nabla^4 \psi - g\alpha \partial_x T. \tag{1.8}$$

The energy equation (1.6) may also be expressed in simpler form. Let $T_c(z)$ be the linear conduction profile satisfied by a fluid in the static state, defined in dimensional form as

$$T_c(z) := T_b - \frac{\Delta T}{\pi d} z, \tag{1.9}$$

and let θ be the negative deviation from the conduction state:

$$\theta := T_c(z) - T. \tag{1.10}$$

Hence (1.6) becomes:

$$\partial_t \theta - \{\psi, \theta\} = \kappa \nabla^2 \theta + \frac{\Delta T}{\pi d} \partial_x \psi. \tag{1.11}$$

To express the evolution equations for ψ and θ in dimensionless form, we scale length by d, time by d^2/κ, and the stream function by κ. The remaining dimensional quantities form two dimensionless groups. These are typically chosen as the Prandtl number, σ, and Rayleigh number, Ra, given by:

$$\sigma = \frac{\nu}{\kappa}, \qquad \mathrm{Ra} = \frac{g\alpha(\pi d)^3 \Delta T}{\nu \kappa}. \tag{1.12}$$

The Rayleigh number is a distinguished dimensionless group for thermal convection because it determines the stability of the conduction state. It is convenient to work with a modified Rayleigh number $\mathcal{R} = \mathrm{Ra}/\pi^4$ to avoid extra factors of π in the dimensionless equations, and to scale the temperature by $\Delta T/(\pi \mathcal{R})$. Therefore, the dimensionless equations governing Rayleigh–Bénard convection in two dimensions become [22]

$$\begin{aligned} \partial_t \nabla^2 \psi - \{\psi, \nabla^2 \psi\} &= \sigma \nabla^4 \psi + \sigma \partial_x \theta \\ \partial_t \theta - \{\psi, \theta\} &= \nabla^2 \theta + \mathcal{R} \partial_x \psi. \end{aligned} \tag{1.13}$$

These equations are often called the *Boussinesq equations* (or Oberbeck–Boussinesq equations) and we do so throughout this thesis.

1.3 Boundary conditions

Boundary conditions must be specified for (1.13) so that the system of partial differential equations (PDEs) is well-posed. The vertical component of velocity vanishes along the impermeable horizontal walls at $z = 0$ and $z = \pi$. Another boundary condition is required to fully specify the velocity field; the most common choices are *stress-free* and *no-slip* conditions. In the former, the fluid does not impart a shear stress on the wall, meaning that the normal derivative of the horizontal velocity must vanish. Together, the stress-free and impermeability conditions imply

$$\psi, \, \partial_{zz} \psi = 0, \;\; \text{at } \; z = 0, \pi. \tag{1.14}$$

The impermeability condition only requires ψ to be constant along the boundaries, but without loss of generality we demand that this constant is zero along each boundary by choosing the constant of integration implied in the definition of ψ.[2] Stress-free conditions are employed in this thesis in part because it is straightforward to construct Fourier series whose basis elements satisfy these boundary conditions. This is advantageous for the construction of reduced order models presented in Chapter 2 that are studied throughout this thesis. If the stress-free conditions are replaced by no-slip boundary conditions, this requires that the fluid is at rest along the boundary. The stream function then satisfies

$$\psi,\ \partial_z \psi = 0, \quad \text{at } z = 0, \pi. \tag{1.15}$$

There are a few options of boundary conditions for the temperature as well. We employ fixed-temperature (isothermal) boundary conditions, meaning that the temperature is constant along the walls. Since θ is the deviation from the linear temperature profile, it vanishes along the isothermal horizontal boundaries:

$$\theta = 0,\ z = 0, \pi. \tag{1.16}$$

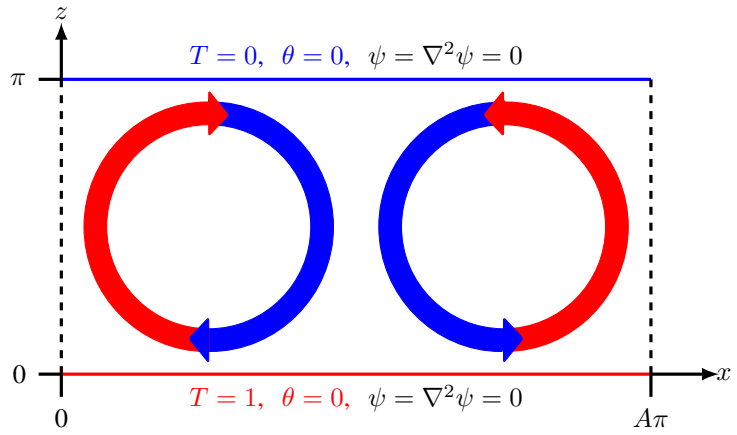


Figure 1.2: Schematic of the Rayleigh–Bénard domain with isothermal, stress-free walls at the top and bottom boundaries. All variables are periodic in x with period $A\pi$.

Another option is to impose fixed-flux conditions that specify the derivative of temperature normal to the boundary, but these conditions will not be explored in this thesis.

[2]If ψ takes different constant values along the two horizontal walls, (1.7) implies that the fluid has a net horizontal flux through any vertical strip. In an inertial reference frame, this flux is zero and so without loss of generality, ψ takes the same value along both boundaries.

It is common in studies of Rayleigh–Bénard convection to define periodic boundary conditions in the horizontal direction. We do so for all models constructed in this thesis, letting ψ and θ be periodic in x with period $A\pi$. Periodic domains are physically relevant in a variety of circumstances, including toric geometries and domains where the aspect ratio is sufficiently large. Other considerations can be made for fully enclosed domains, but we do not explore this type of geometry.

1.4 Heat transport and the Nusselt number

Rayleigh–Bénard convection is a mechanism of heat transport, and the dependence of the heat flux through the layer on the applied temperature gradient is of interest. To deduce this relationship, one typically seeks a relationship between the temperature difference, given in dimensionless form by the Rayleigh number, and the time-averaged rate of heat transport for solutions to (1.13). The dimensionless parameter that expresses heat transport is the Nusselt number, Nu, defined as the ratio of total heat transport, averaged over volume and infinite time, to the heat transport in the purely conductive state. Define the volume average as

$$\langle f \rangle := \frac{1}{A\pi^2} \int_0^{\pi} \int_0^{A\pi} f(x,z) \, \mathrm{d}x \, \mathrm{d}z, \tag{1.17}$$

the horizontal average as

$$\langle f \rangle_x := \frac{1}{A\pi} \int_0^{A\pi} f(x,z,t) \, \mathrm{d}x, \tag{1.18}$$

and the infinite-time average as

$$\overline{f} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^{\tau} f(\tau) \, \mathrm{d}\tau, \tag{1.19}$$

assuming the limit exists. In the variables ψ and θ, the Nusselt number is

$$\mathrm{Nu} = 1 + \tfrac{1}{\mathcal{R}} \overline{\langle \theta \partial_x \psi \rangle}. \tag{1.20}$$

Therefore, $\mathrm{Nu} = 1$ when the fluid is in the purely conductive state. The Nusselt number can be equivalently expressed with various other spatial integrals that have the same value as (1.20) when averaged over infinite time. One such alternative is the ratio of total heat transport at any fixed height $z \in [0, \pi]$, averaged horizontally and over time, to conductive

transport. This version of Nu is expressed as [22]:

$$\mathrm{Nu}(z) = 1 + \tfrac{1}{\mathcal{R}} \left[\overline{\partial_z \langle \theta \rangle_x}(z) + \overline{\langle \theta \partial_x \psi \rangle_x}(z) \right]. \tag{1.21}$$

For any statistically steady flow, this definition of Nu is independent of z and equivalent to (1.20). In either case, the Nusselt number generally depends on the choice of initial condition. Therefore one may consider the Nusselt number along particular solutions to (1.13), or seek its maximum or minimum among all flows satisfying the boundary conditions.

A major theoretical challenge is to determine the relationship between the Rayleigh number and Nusselt number for a fluid with Prandtl number σ, in a domain with aspect ratio A. It is often presumed that in the large-Ra limit, the Nusselt number along any statistically steady flow is asymptotic to a power function of Ra, that is

$$\mathrm{Nu} \sim \sigma^p \mathrm{Ra}^q, \quad \mathrm{Ra} \to \infty, \tag{1.22}$$

for some scaling exponents p and q. A major unsolved problem in the theory of thermal convection is to determine the value of these exponents (if such a relationship exists). Numerous investigators have addressed this problem, and a brief history of such works is presented below.

Malkus [43] and Priestley [56] each proposed that $\mathrm{Nu} \sim \mathrm{Ra}^{1/3}$ as $\mathcal{R} \to \infty$. This follows from the scaling argument that for fully developed, high-Ra flows at large aspect ratio, the time-averaged vertical heat flux is independent of the layer height. Another dimensional argument was offered by Spiegel [68, 69], who hypothesized $\mathrm{Nu} \sim (\sigma \mathrm{Ra})^{1/2}$. Spiegel postulated that at large-$\mathcal{R}$, hot and cold plumes originating from the boundary layer would approach the fluid's free fall velocity. In this so-called "ultimate" regime, transport across the bulk, rather than across thermal boundary layers, is the rate-limiting process.

Numerous studies have been conducted to determine the scaling law at large $\mathcal{R}$, including laboratory experiments [64, 50, 51, 78] and direct numerical simulations [71, 34, 33] . Recent results suggest that the $1/3$ scaling law persists over at least five orders of magnitude (from $\mathrm{Ra} \approx 10^{10}$ to 10^{15}) [33, 12]. Whether a transition to the ultimate regime postulated by Spiegel occurs is yet to be determined.

Both simulations and experiments introduce uncertainties that cannot be completely controlled. An alternative theoretical approach is to derive upper or lower bounds on $\mathrm{Nu}(\mathrm{Ra}, \sigma, A)$ directly from the equations of motion. Variational methods are used in [81, 79], where the maximal heat transport is rigorously bounded for two dimensional Rayleigh–

Bénard convection with stress-free isothermal boundaries. The tightest variational bounds establish that $q \leq 5/12$ in (1.22) [79]. This does not rule out the $q = 1/2$ scaling in the general case, but it implies that if the ultimate regime of convection exists, it must either occur for domains with no-slip boundaries or include fluid motions that are fully three-dimensional.

1.5 Outline of the dissertation

The focus of this dissertation is to study the heat transport properties of reduced order models of Rayleigh–Bénard convection. Reduced order models are discussed in Chapter 2, including the derivation of such models and a discussion of the properties that result in distinguished convection models. In Chapter 3, we study particular solutions of a distinguished eight-ODE model of Rayleigh-Bénard convection first studied by Gluhosvky, *et al.* [21]. Chapter 4 outlines a recently developed technique that provides bounds on time-averaged quantities for dynamical systems. This method is applied to produce upper bounds on heat transport for Gluhovsky's eight-ODE model in Chapter 5. Numerical bounds are constructed using computational techniques in §5.1, and an analytical approach is employed in §5.2 to establish rigorous upper bounds. The results in Chapters 3 and 5, along with Appendices A, B and C are adapted from [52]. Chapter 6 explores presents upper bounds on the Nusselt number for some reduced models with dimension larger than eight, and compares these bounds to various particular solutions. The work of Appendix D is tangential to this work, and explores a technique for diagnosing the accuracy of numerical simulations of the incompressible Euler equations.

Most of the work in this thesis was performed under the guidance of Charles Doering, William Schultz, and David Goluskin, who also provided editing throughout. This research was supported in part through computational resources and services provided by Advanced Research Computing at the University of Michigan, Ann Arbor.

CHAPTER 2

Reduced Order Convection Models

2.1 Introduction

A substantial body of research has been devoted to studying the dependence of the heat transport on the Rayleigh number in Rayleigh–Bénard convection. Even so, there remains a gap between the rigorous upper bounds on the Nusselt number derived from the equations of motion, and the maximal heat transport obtained from laboratory experiments and numerical simulations. As a complementary approach, one may construct *reduced order models*, or ROMs. These are finite systems of ordinary differential equations (ODEs) derived from the governing equations. Ideally, ROMs should approximate the dynamics of fluid convection. Various ODE models have been derived for Rayleigh–Bénard convection, beginning with the atmospheric model of Saltzman [60] that inspired the seminal study by Lorenz of a three-dimensional ROM now known as the Lorenz equations [41]. Following the work of Lorenz, many other ROMs have been studied as simplified convection models [32, 73, 74, 29, 21]. This chapter provides details on the construction of reduced order models for 2D Rayleigh–Bénard convection with stress-free isothermal boundaries in a horizontally periodic domain.

2.2 Galerkin expansions and low-order models

The Lorenz equations are constructed using a three-term Fourier expansion [41]:

$$\psi(x,z,t) = \sqrt{2}\,\tfrac{k^2+1}{k} X(\tau)\,\sin(kx)\sin(z), \tag{2.1}$$

$$\theta(x,z,t) = \sqrt{2}\tfrac{(k^2+1)^3}{k^2}\,Y(\tau)\,\cos(kx)\sin(z) + \tfrac{(k^2+1)^3}{k^2} Z(\tau)\sin(2z), \tag{2.2}$$

where $k := 2/A$ is the fundamental horizontal wavenumber for a domain of aspect ratio A, and $\tau = (k^2+1)t$. Substituting (2.1)–(2.2) into the Boussinesq equations and projecting the result onto the three Fourier modes in the expansions for ψ and θ produces a system of ODEs

for the evolution of the coefficient functions X, Y and Z. In doing so, some terms outside the span of the original Fourier modes are discarded, and therefore this type of system is sometimes called a truncated model. Letting $\beta = 4/(k^2+1)$ and $\rho = \mathcal{R}k^2/(k^2+1)^3$ so that the onset of convection corresponds to $\rho = 1$, we obtain the well-known system

$$\begin{aligned} \dot{X} &= -\sigma X + \sigma Y, \\ \dot{Y} &= -Y + X(\rho - Z), \\ \dot{Z} &= -\beta Z + XY. \end{aligned} \tag{2.3}$$

Although the Lorenz equations are a simplified model of Rayleigh's PDE, the ODE system correctly predicts the minimal Rayleigh number where convection can occur and accurately models the physics of 2D Rayleigh–Bénard convection near the onset of convection.[1] Lorenz discovered that under certain conditions, solutions of the Lorenz equations approach a chaotic attractor that is now the prototype for chaotic behavior in continuous dynamical systems.

The derivation of the Lorenz equations is an example of a general technique known as *Galerkin expansion*, where dependent variables are expanded in terms of an orthogonal set of basis functions that each satisfy the boundary conditions, producing an ROM. Stress-free boundaries are chosen for models constructed in this work in part because this allows expansion in terms of the Fourier basis. Square-integrable functions ψ and θ satisfying the boundary conditions are given by the series

$$\begin{aligned} \psi(x,z,t) &= \sum_{m=0}^{\infty}\sum_{n=1}^{\infty} \big(a_{mn}(t)\cos(mkx) + b_{mn}(t)\sin(mkx)\big)\sin(nz), \\ \theta(x,z,t) &= \sum_{m=0}^{\infty}\sum_{n=1}^{\infty} \big(c_{mn}(t)\cos(mkx) + d_{mn}(t)\sin(mkx)\big)\sin(nz). \end{aligned} \tag{2.4}$$

The subscripts on the coefficients in the above expansions correspond to the indices on the horizontal and vertical mode numbers of the associated Fourier modes. Inserting the expansions (2.4) into the Boussinesq equations (1.13) and projecting the resulting expression onto each basis element results in a system of ordinary differential equations describing the time evolution of the coefficient functions. Practical applications require that the series expansions be truncated to some finite number of terms in (2.4), resulting in a system of

[1] The onset of convection corresponds to the first instability of the zero equilibrium as the Rayleigh number is raised.

ODEs that approximates the full PDE. Nonlinear interactions between the Fourier modes produce terms outside the span of the modes in the truncated model; these excess terms are discarded when projecting onto only the included modes.

2.3 Model construction

The series (2.4) for ψ and θ represent a class of functions that are equivalent up to a horizontal phase shift. Without loss of generality, we consider solutions of fixed horizontal phase by setting $a_{11} \equiv 0$, as in the derivation of the Lorenz equations. This restricts the frame of reference such that the pair of steady convection rolls present at the onset of convection is symmetric about the center of the domain. Making this choice determines the phase of all terms in (2.4) to maintain consistency upon substitution into the governing equations. For example, it immediately implies that $d_{11} \equiv 0$ since this would appear on the right-hand side of the a_{11} equation. As a result of this convention, exactly one coefficient remains in each term of the Fourier expansions for ψ and θ. Letting ψ_{mn} and θ_{mn} be the remaining coefficients under our convention, these new coefficients are related to those in (2.4) as shown in Table 2.1. One consequence of imposing this phase convention is that all modes of the form ψ_{0n} (also called shear modes) must have odd vertical wavenumber, while modes of the form θ_{0n} must have even vertical wavenumber.

Table 2.1: Coefficients in the expansions for ψ and θ satisfying our phase convention.

	$m+n$ even	$m+n$ odd
ψ_{mn}	b_{mn}	a_{mn}
θ_{mn}	c_{mn}	d_{mn}

The modes included in (2.4) are capable of capturing the dynamics of *zonal flow*, where mean horizontal flows near the top and bottom boundaries vertically shear the fluid [22]. Zonal flow has been observed in experiments of turbulent convection [35] and occurs in toroidal plasmas [11] and planetary atmospheres [5]. Interest in such mean horizontal flows motivated Howard and Krishnamurti [32] to augment the Lorenz equations by including the ψ_{01}, ψ_{12} and θ_{12} modes. That model helped illuminate a mean-flow instability but is not suitable for studying heat transport even as a low-order model because some of its trajectories are unbounded. Another issue with this model is that expressions for time-averaged heat transport such as (1.20) and (1.21) that are equivalent in the PDE dynamics give expressions

that generally differ in the ODE dynamics once projected onto the chosen set of six modes. Thiffeault and Horton [73, 74] found that adding the θ_{04} mode restores boundedness of trajectories and equality between the two truncated versions of Nu, (1.20) and (1.21), as well as conservation of mechanical energy in the dissipationless limit $\nu, \kappa \to 0$. Separately, Hermiz *et al.* [29] found that adding the ψ_{03} mode results in an ODE system whose solutions conserve the truncated version of total vorticity in the inviscid limit. These ideas culminated in an eight-dimensional model introduced by Gluhovsky *et al.* [21], who confirmed that including the θ_{04} mode added by Thiffeault and Horton [73, 74] and the ψ_{03} mode added by Hermiz *et al.* indeed combines the benefits of both. This system is called the HK8 model because it is the minimal extension of the six-mode model that restores these basic integral identities of the PDE. It is presented in §2.5, and studied in detail in Chapters 3 and 5.

We now describe the general form of reduced order convection models satisfying the above properties, presented previously in [76, 73]. Suppose a low-order model is defined by selecting a finite number of modes from the expansions of ψ and θ. Let M_ψ and M_θ be the sets of ordered pairs (m, n) corresponding to the indices of modes included in the truncated expansions for ψ and θ, respectively. Galerkin expansion yields the following system of ODEs [73]:

$$\dot{\psi}_{mn} = -\sigma\rho_{mn}\psi_{mn} + (-1)^{m+n}\sigma\frac{mk}{\rho_{mn}}\theta_{mn} + \frac{k}{\rho_{mn}}Q_{mn}, \tag{2.5}$$

$$\dot{\theta}_{mn} = -\rho_{mn}\theta_{mn} + (-1)^{m+n}\mathcal{R}(mk)\psi_{mn} + k\widetilde{Q}_{mn}, \tag{2.6}$$

where $\rho_{mn} := (mk)^2 + n^2$ are the eigenvalues of $-\nabla^2$, and Q_{mn}, $\widetilde{Q}_{mn}$ are given by the sum of all quadratic terms in the reduced order model. The quadratic terms arise from the nonlinear terms of (1.13) and their exact expressions are provided below. For fixed (m, n), each term in Q_{mn} is proportional to $\psi_{pq}\,\psi_{rs}$ whose modal pairs lie in the set

$$S_\psi(m,n) = \{(p,q),(r,s) \in M_\psi : m = |p \pm r|, n = |q \pm s|, (p,q) > (r,s)\}, \tag{2.7}$$

where $(p, q) > (r, s)$ is the lexicographical ordering, defined by

$$(p,q) > (r,s) \iff p > r \text{ or } (p = r \text{ and } q > s). \tag{2.8}$$

Writing S_ψ in this way ensures each term in Q_{mn} is associated with a unique choice of the four indices in S_ψ. Similarly, the terms in $\widetilde{Q}_m$ are proportional to $\psi_{pq}\,\theta_{rs}$, represented by the

set

$$S_\theta(m,n) = \{(p,q),(r,s) \in M_\theta : m = |p \pm r|, n = |q \pm s|\}. \tag{2.9}$$

No ordering is needed on the pairs in S_θ since commuting the modal pairs yields a distinct term in the sum. The quadratic terms are then expressed as

$$Q_{mn} = \sum_{S_\psi(m,n)} \frac{\mu_1}{d}\big[B_{pmr}B_{snq}(ps) - B_{qns}(qr)\big](\rho_{pq} - \rho_{rs})\psi_{pq}\psi_{rs}, \tag{2.10}$$

$$\widetilde{Q}_{mn} = \sum_{S_\theta(m,n)} \frac{\mu_2}{d}\big[B_{pmr}B_{snq}(ps) - \mu_3 B_{qns}B_{rpm}(qr)\big]\psi_{pq}\theta_{rs}, \tag{2.11}$$

where B, μ_1, μ_2, μ_3 and d are defined by

$$B_{ijk} = \begin{cases} -1, & i = j + k, \\ 1, & \text{else}, \end{cases} \tag{2.12}$$

$$\mu_1 = \begin{cases} B_{pmr}, & (m+n) \text{ even}, (r+s) \text{ odd}, \\ -B_{pmr}, & (m+n) \text{ odd}, (r+s) \text{ odd}, \\ -1, & \text{else}, \end{cases} \tag{2.13}$$

$$\mu_2 = \begin{cases} \mu_3\, B_{rpm}, & (m+n) \text{ even}, (r+s) \text{ odd}, \\ B_{mpr}, & (m+n) \text{ odd}, (r+s) \text{ even}, \\ B_{pmr}, & (m+n) \text{ even}, (r+s) \text{ even}, \\ 1 & \text{else}, \end{cases} \tag{2.14}$$

$$\mu_3 = \begin{cases} -1, & m = 0, \\ 1, & \text{else}, \end{cases} \tag{2.15}$$

$$d = \begin{cases} 2, & p = 0 \text{ or } r = 0, \\ 4, & \text{else}. \end{cases} \tag{2.16}$$

Given an ROM of the form (2.5)–(2.6), a version of the Nusselt number can be defined by projecting either version of Nu described in §1.4 onto $M_\psi \cup M_\theta$. We denote the Nusselt number of each reduced model by N to distinguish it from the PDE quantity Nu that it approximates. When the truncated Fourier series are inserted into the volume-averaged

definition of Nu (1.20), orthogonality reduces the expression to

$$N = 1 + \frac{1}{4\mathcal{R}} \sum_{(m,n)\in M_\psi \cap M_\theta} (-1)^{m+n} (mk)\, \overline{\psi_{mn}\theta_{mn}}. \tag{2.17}$$

Alternatively, deriving the expression for N from (1.21) yields

$$N = 1 + \frac{1}{\mathcal{R}} \sum_{(0,2n)\in M_\theta} (2n)\, \overline{\theta_{0,2n}}. \tag{2.18}$$

Whether the two definitions of N are equivalent in the long-time average depends on the choice of modes in the truncation model. Equivalence of these expressions is a desirable property for ROMs because an analogous result holds for the expressions (1.20) and (1.21) in the Boussinesq equations. Motivated by the study of optimal heat transport in Rayleigh–Bénard convection, we seek to determine the maximum Nusselt number, N^*, given by

$$N^* := \max_{\mathbf{x}(t)} N, \tag{2.19}$$

where the maximization is over all solutions $\mathbf{x}(t)$ for a given reduced order model, and generally depends on the parameters $\mathcal{R}, \sigma$ and k.

2.4 Conservation properties

Constructing a Galerkin-truncated model of the Boussinesq equations requires selecting a finite set of modes in the Fourier expansions for ψ and θ. There is no universally accepted way to choose which modes to include, although a few guidelines have been established to promote consistency with the Boussinesq equations. Authors in previous studies of low-order models [76, 73, 29, 21] have suggested imposing criteria based on certain conservation laws derived from the Boussinesq equations. In §2.4.1 we examine the restrictions on mode selection imposed by the conservation of energy, temperature and vorticity in the dissipationless limit $\nu, \kappa \to 0$, and in §2.4.2 we show that these restrictions do not change if one considers the analogous integral balance laws derived from the full Boussinesq equations.

2.4.1 Conservation laws in the dissipationless limit

The dimensionless form of the Boussinesq equations (1.13) is not amenable to taking the dissipationless limit, since some variable scalings depend on ν and κ. Instead, we nondimensionalize the equations by scaling length by d, time by $\sqrt{d/g\alpha\Delta T}$, the stream function by $\sqrt{g\alpha\Delta T d^3}$ and temperature by ΔT as in [73]. As $\nu, \kappa \to 0$, the governing equations become

$$\partial_t \nabla^2 \psi - \{\psi, \nabla^2 \psi\} = \partial_x \theta, \tag{2.20}$$

$$\partial_t \theta - \{\psi, \theta\} = \partial_x \psi. \tag{2.21}$$

The dissipationless version of the reduced equations is similar to (2.5)–(2.6), except that the first term on the right-hand side of each equation is removed, and factors of σ and $\mathcal{R}$ are absent. The equations (2.20)–(2.21) admit a number of conserved quantities [73, 21]. One such quantity is the energy, $E = K + U$, where K and U satisfy

$$K = \frac{1}{2}\big\langle\, |\nabla\psi|^2 \,\big\rangle, \tag{2.22}$$

$$U = \big\langle z\theta \big\rangle, \tag{2.23}$$

with the volume average $\langle\cdot\rangle$ defined by (1.17). To verify that this is conserved by the dissipationless Boussinesq equations, multiply (2.20) by ψ and average over the domain, imposing boundary conditions and integrating by parts when necessary, to obtain

$$\frac{1}{2}\partial_t \big\langle\, |\nabla\psi|^2 \,\big\rangle = -\big\langle \psi \partial_x \theta \big\rangle. \tag{2.24}$$

Likewise, multiplying (2.21) by z and taking the volume average produces

$$\partial_t \big\langle z\theta \big\rangle = \big\langle \psi \partial_x \theta \big\rangle. \tag{2.25}$$

Adding these two expressions provides the desired result $\partial_t E = 0$. Thiffeault and Horton [73, 74] found that all Galerkin-truncated models conserving energy in the dissipationless limit must satisfy:

Criterion 1 (Energy balance)**.** If $(m, n) \in M_\psi \cap M_\theta$, then $(0, 2n) \in M_\theta$.

The Lorenz equations have $M_\psi \cap M_\theta = \{(1, 1)\}$ and $(0, 2) \in M_\theta$, satisfying Criterion 1. On the other hand, the model of Howard and Krishnamurti includes $(1, 2) \in M_\psi \cap M_\theta$

but is missing $(0,4) \in M_\theta$; adding θ_{04} restores the energy balance [73]. There are a few important consequences of selecting reduced models that satisfy the energy criterion. All trajectories of such models remain bounded, even in the presence of dissipation [73]. This is significant because unbounded trajectories have been observed for certain ODE models [32], marking a significant divergence from the physics of Rayleigh–Bénard convection. Also, the two definitions of N, (2.17) and (2.18), are equivalent along all solutions of models satisfying the energy criterion [73]. Finally, these models also satisfy the conservation of total temperature,

$$\partial_t \big\langle \theta \big\rangle = 0, \tag{2.26}$$

derived by taking the volume average of (2.21).

The Boussinesq equations also satisfy conservation of the integral of vorticity [21] in the dissipationless limit:

$$\partial_t \Big\langle \nabla^2 \psi \Big\rangle = 0, \tag{2.27}$$

determined by taking the volume average of (2.20). Next we identify the criterion for reduced models to satisfy vorticity conservation in the dissipationless limit. Let the function $\mathrm{sc}(mkx)$ represent either $\sin(mkx)$ or $\cos(mkx)$ according to the convention established in Table 2.1. Projecting the integral of vorticity onto M_ψ yields

$$\big\langle \nabla^2 \psi \big\rangle = \left\langle -\sum_{(m,n)\in M_\psi} \rho_{mn} \psi_{mn} \,\mathrm{sc}(mkx) \sin(nz) \right\rangle \tag{2.28}$$

$$= -\sum_{(m,n)\in M_\psi} \rho_{mn} \psi_{mn} \big\langle \mathrm{sc}(mkx) \sin(nz) \big\rangle. \tag{2.29}$$

The volume average vanishes for each term with $m \neq 0$ and for all even n, so that only the shear modes ψ_{0n} remain. For these terms, $\rho_{0n} = n^2$, so after integrating, the volume-averaged vorticity reduces to

$$\big\langle \nabla^2 \psi \big\rangle = -\tfrac{1}{\pi} \sum_{(0,n)\in M_\psi} \tfrac{1-\cos(\pi n)}{n} n^2 \, \psi_{0n} \tag{2.30}$$

$$= -\tfrac{1}{\pi} \sum_{\substack{(0,n)\in M_\psi \\ n \text{ odd}}} 2n \, \psi_{0n}. \tag{2.31}$$

Taking the time derivative and using (2.5) in the special case $m = 0$, we obtain

$$\partial_t \langle \nabla^2 \psi \rangle = -\tfrac{1}{\pi} \sum_{(0,n)\in M_\psi} 2n\, \dot{\psi}_{0n} \tag{2.32}$$

$$= -\tfrac{1}{\pi} \sum_{(0,n)\in M_\psi} 2n\, \dot{\psi}_{0n} \tag{2.33}$$

$$= -\tfrac{1}{\pi} \sum_{(0,n)\in M_\psi} \tfrac{2k}{n} \sum_{S_\psi(0,n)} (-1)^{p+s} p\,(s + B_{qns}\, q)\,(q^2 - s^2) \psi_{pq} \psi_{ps}. \tag{2.34}$$

The set $S_\psi(0,n)$ (2.7) contains all pairs of modes contributing to the quadratic part of the ODE for the shear mode ψ_{0n}. These pairs take the form $\{(p,q),(p,s)\}$, where $p \neq 0$ and either $n = q + s$ or $n = |q - s|$, and correspond to terms proportional to $\psi_{pq}\, \psi_{ps}$. Suppose that M_ψ contains the modes (p,q), (p,s) and $(0,|q-s|)$. To ensure cancellation of the resulting term in (2.34), the mode $(0, q+s)$ must also be included in M_ψ. Then (2.34) has exactly two terms proportional to $\psi_{pq}\psi_{rs}$, whose sum is

$$(-1)^{p+s}\, p\,(q^2 - s^2)\, \tfrac{k}{2} \left(\tfrac{1}{(q+s)}(s+q) + \tfrac{1}{|q-s|}(-|s-q|) \right) \psi_{pq}\psi_{ps} = 0. \tag{2.35}$$

Criterion 2 (Vorticity balance)**.** If $(p,q) \in M_\psi$ and $(p,s) \in M_\psi$, then $(0,|q-s|) \in M_\psi$ if and only if $(0, q+s) \in M_\psi$.

The six-ODE and seven-ODE models discussed above include $(1,1),(1,2) \in M_\psi$ as well as the shear mode $(0,1) \in M_\psi$. Hence vorticity conservation is enforced by adding $(0,3) \in M_\psi$ [29].

2.4.2 Integral balances in the presence of dissipation

For each of the conservation laws of the dissipationless Boussinesq equations, there exists an analogous integral balance derived from the full PDE (1.13). The resulting energy, temperature and vorticity balance laws are

$$\partial_t \left[\tfrac{1}{2} \left\langle |\nabla \psi|^2 \right\rangle + \sigma \langle z\theta \rangle \right] = \sigma \left\langle z \nabla^2 \theta \right\rangle - \sigma \left\langle (\nabla^2 \psi)^2 \right\rangle, \tag{2.36}$$

$$\partial_t \langle \theta \rangle = \left\langle \nabla^2 \theta \right\rangle, \tag{2.37}$$

$$\partial_t \left\langle \nabla^2 \psi \right\rangle = \sigma \left\langle \nabla^4 \psi \right\rangle. \tag{2.38}$$

Here we show that if a reduced order model obeys conservation of energy, temperature and vorticity in the dissipationless limit, it also satisfies the corresponding balance equations for the PDE with dissipation.

If the truncated Fourier expansions for ψ and θ are substituted into (2.36), orthogonality reduces the left-hand side of the energy balance to

$$\partial_t E = \sum_{\substack{(m,n)\in M_\psi \\ m\neq 0}} \tfrac{1}{4}\rho_{mn}\psi_{mn}\dot{\psi}_{mn} + \sum_{(0,n)\in M_\psi} \tfrac{1}{2}n^2\psi_{0n}\dot{\psi}_{0n} - \sigma \sum_{(0,n)\in M_\theta} \tfrac{1}{2n}\dot{\theta}_{0,2n}. \tag{2.39}$$

The proof in [73] shows that the above expression vanishes in the absence of dissipation provided the modes are selected as specified in §2.4.1. Here, similar cancellation occurs, leaving only the terms resulting from the dissipative terms of (1.13):

$$\partial_t E = -\sum_{\substack{(m,n)\in M_\psi \\ m\neq 0}} \tfrac{1}{4}\rho_{mn}^2\sigma\psi_{mn}^2 - \sum_{(0,n)\in M_\psi} \tfrac{1}{2}n^4\sigma\psi_{0n}^2 + \sigma \sum_{(0,n)\in M_\theta} \tfrac{n}{2}\theta_{0,2n}. \tag{2.40}$$

The projection of the right-hand side of (2.36) onto any Fourier-truncated ψ and θ is identical to the above expression, so the general energy balance holds under the exact same conditions as its dissipationless version. The truncated versions of the integral balances for temperature (2.37) and vorticity (2.38) also follow directly from their conservation in the dissipationless case, and can be proved in a similar manner. We call low-order models that obey each of the conservation laws above *distinguished models*.

2.5 The HK8 model

The smallest distinguished model that includes nontrivial zonal flow is the HK8 model first considered by Gluhovsky, et al. [21]. We analyze particular solutions to the HK8 model in Chapter 3 and construct upper bounds on the heat transport in Chapter 5. The HK8 system is derived by projecting the Boussinesq equations onto the Fourier modes in the ansatz

$$\begin{aligned}\psi(x,z,t) &= \psi_{11}(t)\sin kx\sin z + \psi_{12}(t)\cos kx\sin 2z + \psi_{01}(t)\sin z + \psi_{03}(t)\sin 3z,\\ \theta(x,z,t) &= \theta_{11}(t)\cos kx\sin z + \theta_{12}(t)\sin kx\sin 2z + \theta_{02}(t)\sin 2z + \theta_{04}(t)\sin 4z,\end{aligned} \tag{2.41}$$

The chosen Galerkin truncation includes the triplet $\{\psi_{11},\theta_{11},\theta_{02}\}$ that alone produces a scaled version of the Lorenz equations (2.3). It also includes the analogous triplet with the

vertical mode numbers doubled, $\{\psi_{12}, \theta_{12}, \theta_{04}\}$, alone yielding a rescaled version of the Lorenz equations. Modes in each triplet are coupled together by the two remaining modes, ψ_{01} and ψ_{03}. The HK8 model obtained by projecting the PDEs (1.13) onto the modes in (2.41) is [22]

$$\begin{aligned}
\dot{\psi}_{11} &= -\sigma(k^2+1)\psi_{11} + \sigma\tfrac{k}{k^2+1}\theta_{11} + \tfrac{k}{2}\tfrac{k^2+3}{k^2+1}\psi_{01}\psi_{12} - \tfrac{3k}{2}\tfrac{k^2-5}{k^2+1}\psi_{12}\psi_{03},\\
\dot{\psi}_{01} &= -\sigma\,\psi_{01} - \tfrac{3k}{4}\psi_{11}\psi_{12},\\
\dot{\psi}_{12} &= -\sigma(k^2+4)\psi_{12} - \sigma\tfrac{k}{k^2+4}\theta_{12} - \tfrac{1}{2}\tfrac{k^3}{k^2+4}\psi_{11}\psi_{01} + \tfrac{3k}{2}\tfrac{k^2-8}{k^2+4}\psi_{11}\psi_{03},\\
\dot{\theta}_{11} &= -(k^2+1)\theta_{11} + \mathcal{R}k\psi_{11} - k\psi_{11}\theta_{02} - \tfrac{k}{2}\psi_{01}\theta_{12} + \tfrac{3k}{2}\theta_{12}\psi_{03},\\
\dot{\theta}_{02} &= -4\,\theta_{02} + \tfrac{k}{2}\psi_{11}\theta_{11},\\
\dot{\theta}_{12} &= -(k^2+4)\theta_{12} - \mathcal{R}k\psi_{12} + \tfrac{k}{2}\psi_{01}\theta_{11} - \tfrac{3k}{2}\psi_{03}\theta_{11} + 2k\psi_{12}\theta_{04},\\
\dot{\psi}_{03} &= -9\,\sigma\,\psi_{03} + \tfrac{k}{4}\psi_{11}\psi_{12},\\
\dot{\theta}_{04} &= -16\,\theta_{04} - k\psi_{12}\theta_{12}.
\end{aligned} \tag{2.42}$$

When projected onto the modes in the HK8 truncation, the volume-averaged expression (1.20) for the Nusselt number becomes

$$N = 1 + \tfrac{k}{4\mathcal{R}}(\overline{\psi_{11}\theta_{11} - \psi_{12}\theta_{12}}), \tag{2.43}$$

while the horizontally-averaged expression (1.21) becomes

$$N = 1 + \tfrac{1}{\mathcal{R}}(\overline{2\theta_{02} + 4\theta_{04}}), \tag{2.44}$$

where the bar denotes the time average integral, given by (1.19). It is shown in [73] that the infinite-time averages (2.43) and (2.44) must be equal for all solutions of the HK8 model. The truncated Nusselt number provides a measure of heat transport along solutions of the HK8 model, and is studied in detail in Chapters 3 and 5.

CHAPTER 3

Particular Solutions of the HK8 Model

The HK8 model is a distinguished reduced order model for Rayleigh–Bénard convection with stress-free isothermal boundary conditions in a horizontally periodic domain, and was first considered in [21]. We examine various particular solutions of the HK8 model (2.42), providing candidates for (and lower bounds on) the optimal heat transport. We begin by summarizing the bifurcation structure of steady states of the HK8 model reported in [22] and verified here. Then we examine the heat transport along some time-dependent solutions. The results of this chapter are originally reported in [52]. The maximum N among these particular solutions provides a candidate for the maximum among all solutions, and we use it to assess the sharpness of upper bounds on N reported in Chapter 5.

3.1 Steady states

The HK8 system has three branches of nonzero equilibria that we call L_1, L_2, and TC in analogy with [32]. At sufficiently small $\mathcal{R}$ the unique asymptotic state is the zero equilibrium, corresponding to the purely conductive state in the PDE. When $\mathcal{R} = \mathcal{R}_{L_1}$, given by

$$\mathcal{R}_{L_1} := \frac{(k^2+1)^3}{k^2}, \tag{3.1}$$

the zero state undergoes a pitchfork bifurcation giving rise to L_1 equilibria that exist whenever $\mathcal{R} > \mathcal{R}_{L_1}$, so-named because the only nonzero modes are the first Lorenz triplet,

$$\psi_{11} = \pm\sqrt{8}\,\tfrac{1}{k^2+1}\sqrt{\mathcal{R}-\mathcal{R}_{L_1}}, \quad \theta_{11} = \pm\sqrt{8}\,\tfrac{k^2+1}{k}\sqrt{\mathcal{R}-\mathcal{R}_{L_1}}, \quad \theta_{02} = \mathcal{R}-\mathcal{R}_{L_1}. \tag{3.2}$$

As shown in Figure 3.1(a), the L_1 states are an approximation of a PDE steady state with two convection rolls. The Rayleigh number $\mathcal{R}_{L_1}$ reaches a minimum of $27/4$ when $k^2 = 1/2$, corresponding exactly to the onset of convection for 2D stress-free Rayleigh–

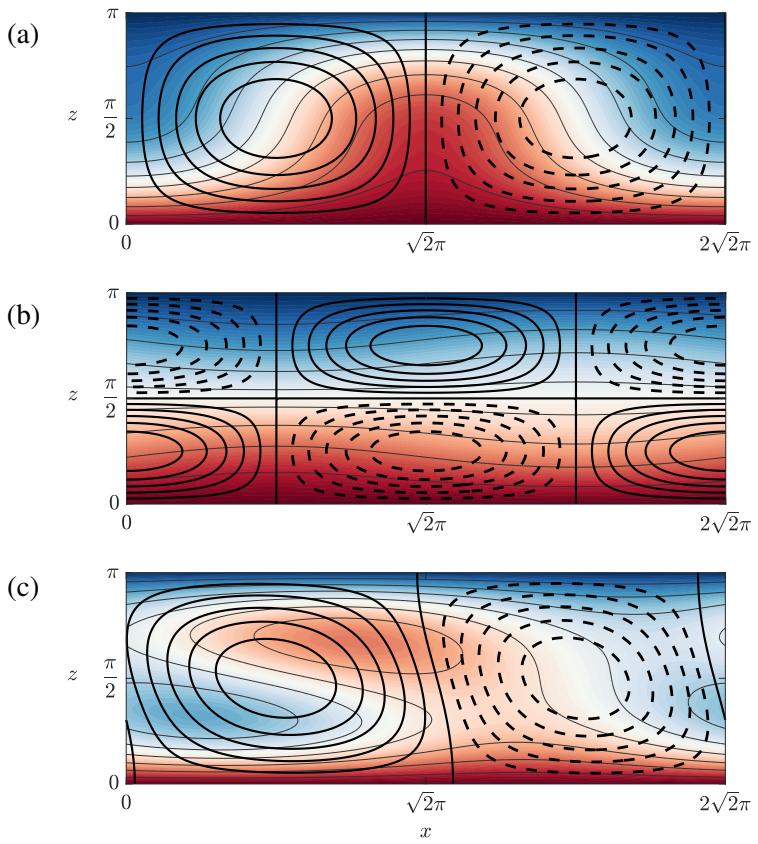


Figure 3.1: Streamlines overlaid on contours of total temperature (T) for approximations of steady convection states whose mode amplitudes in the truncated Galerkin expansion (2.41) are equilibria of the HK8 model with $(k^2, \sigma) = (1/2, 10)$. Each of the three types of equilibria is depicted near its onset: (a) an L_1 state at $\mathcal{R} = 10$, (b) an L_2 state at $\mathcal{R} = 185$, and (c) a TC state at $\mathcal{R} = 150$. The T scale ranges from 0 (dark) to 1 (light). Positive and negative vorticity is indicated by solid and dashed streamlines, respectively. The TC states, in particular, display unphysical behavior due to the truncation of the PDE, evidenced by the internal temperature maximum in (c).

Bénard convection. Therefore, we define the critical Rayleigh number $\mathcal{R}_c$ as

$$\mathcal{R}_c := \frac{27}{4}. \tag{3.3}$$

Similarly, when the Rayleigh number is $\mathcal{R}_{L_2}$, defined as

$$\mathcal{R}_{L_2} := \frac{(k^2+4)^3}{k^2}, \tag{3.4}$$

the zero solution undergoes a second pitchfork bifurcation, giving rise to L_2 equilibria when $\mathcal{R} > \mathcal{R}_{L_2}$. The nonzero modes are the second Lorenz triplet,

$$\psi_{12} = \pm\sqrt{8}\,\tfrac{1}{k^2+4}\sqrt{\mathcal{R}-\mathcal{R}_{L_2}}, \quad \theta_{12} = \mp\sqrt{8}\,\tfrac{k^2+4}{k}\sqrt{\mathcal{R}-\mathcal{R}_{L_2}}, \quad \theta_{04} = \tfrac{1}{2}(\mathcal{R}-\mathcal{R}_{L_2}). \tag{3.5}$$

As shown in Figure 3.1(b), the L_2 states are an approximation of a PDE steady state with a two-by-two array of convection rolls. In the full PDE there are counterparts to the L_1 and L_2 branches that bifurcate from the conductive state at the same $\mathcal{R}$ values. They agree asymptotically with the truncated states in the weakly nonlinear regime but this resemblance decreases as $\mathcal{R}$ grows.

The L_1 and L_2 equilibria are scaled versions of the nonzero equilibria of the Lorenz equations. In fact, if the HK8 equations are restricted to the three-dimensional subspaces spanned by the nonzero variables in either (3.2) or (3.5), the resulting dynamics are equivalent to those of the Lorenz equations. This implies that for any solution to the Lorenz equations, a corresponding solution to the HK8 system can be obtained by a suitable linear change of variables. In the HK8 system, these lower-dimensional dynamics appear to be unstable for all sufficiently large $\mathcal{R}$.

The third branch of equilibria found in the HK8 model is called the TC branch because it corresponds to so-called tilted cells [32]. As shown in Figure 3.1(c), a pair of steady convection rolls produces a mean horizontal flow whose direction breaks the symmetry of the L_1 and L_2 states. All eight modes are nonzero in the TC states, and here we compute them numerically using the numerical continuation software MATCONT [10]. The temperature field in Figure 3.1(c) appears unphysical, as evidenced by the existence of interior maxima, reflecting the fact that the truncated model is not capturing the full PDE dynamics at this $\mathcal{R}$.

Depending on the values of k^2 and σ as $\mathcal{R}$ is varied, there are five different bifurcation structures where the TC branch connects to the L_1 or L_2 branch, or both. Figure 3.2 shows an example of each possible bifurcation structure, along with the regimes in the k^2–σ parameter plane where each structure occurs. In regimes III–V, a pair of TC branches connects to each L_1 branch in a pitchfork bifurcation at $\mathcal{R}_{TC_1}$, where [22]

$$\frac{\mathcal{R}_{TC_1}}{\mathcal{R}_{L_1}} = 1 + \frac{27\sigma^2}{k^2+1}\frac{k^4+5k^2+7}{(10\sigma+3\sigma^2)(k^2+1)^2+2(k^2+4)(5k^2-4)}. \tag{3.6}$$

The above denominator is negative in regimes I and II, so the TC and L_1 branches do not connect. Similarly, in regimes I–IV, a pair of TC branches connects to each L_2 branch in a

pitchfork bifurcation at $\mathcal{R}_{TC_2}$, where

$$\frac{\mathcal{R}_{TC_2}}{\mathcal{R}_{L_2}} = 1 + \frac{27\sigma^2}{k^2+4}\frac{k^4+5k^2+7}{(10\sigma-3\sigma^2)(k^2+4)^2+2(k^2+1)(5k^2+11)}. \tag{3.7}$$

For parameter combinations in regime V the TC and L_2 branches do not connect, as the denominator of (3.7) is negative. The bifurcations connecting the branch of TC equilibria to the L_1 and L_2 branches are quantitatively accurate only in the $\sigma \to 0$ limit since this is when they occur in the weakly nonlinear regime. Counterparts to the TC branch have been observed for the full PDE, at least for some values of k^2 and σ [22].

The L_1, L_2, and TC states are the only nonzero steady states of the HK8 model [22], so finding the maximum N among them at a given parameter set yields the maximum heat transport by any steady state. Evaluating (2.43) or (2.44) to find N in the HK8 model gives the heat transport by the L_1 and L_2 equilibria:

$$N_{L_1} = 3 - 2\,\frac{\mathcal{R}_{L_1}}{\mathcal{R}}, \qquad\qquad N_{L_2} = 3 - 2\,\frac{\mathcal{R}_{L_2}}{\mathcal{R}}. \tag{3.8}$$

Both values approach 3 as $\mathcal{R} \to \infty$, but $N_{L_1} > N_{L_2}$ at any parameters where both states exist. We computed N_{TC} numerically at many parameter values in all five parameter regimes. In regimes I–IV we found $N_{L_1} > N_{TC}$ in all cases, meaning the L_1 branch maximizes heat transport among steady states. In regime V, at sufficiently large $\mathcal{R}$ the TC branch maximizes N among steady states. Whether these maximal steady N values are maximal among time-dependent solutions as well remains to be determined by the bounds computed in Chapter 5. We note that the results at large $\mathcal{R}$ are unlikely to be representative of the PDE: the values of N_{L_1} begin to deviate from the values of Nu for the analogous steady solutions of the Boussinesq equations (i.e., the primary branch of convection rolls that arises as the first instability of the conduction state) near $\mathcal{R} = 10$, and the steady states develop internal maxima. Mean horizontal flow has been observed to reduce heat transport in simulations of 2D Rayleigh–Bénard convection in a horizontally periodic domain with stress-free boundaries [25].

All three nonzero steady branches can undergo Hopf bifurcations. Determining k^2 and σ for the various pitchfork, saddle-node, and Hopf bifurcations to exist (with $\mathcal{R}$ as the bifurcation parameter) subdivides the five parameter regimes of Figure 3.2 into 16 regimes, as reported in [22]. The Hopf bifurcation on the L_1 branch involves only the three Lorenz modes $\{\psi_{11}, \theta_{11}, \theta_{02}\}$ and is precisely the Hopf bifurcation found in the Lorenz

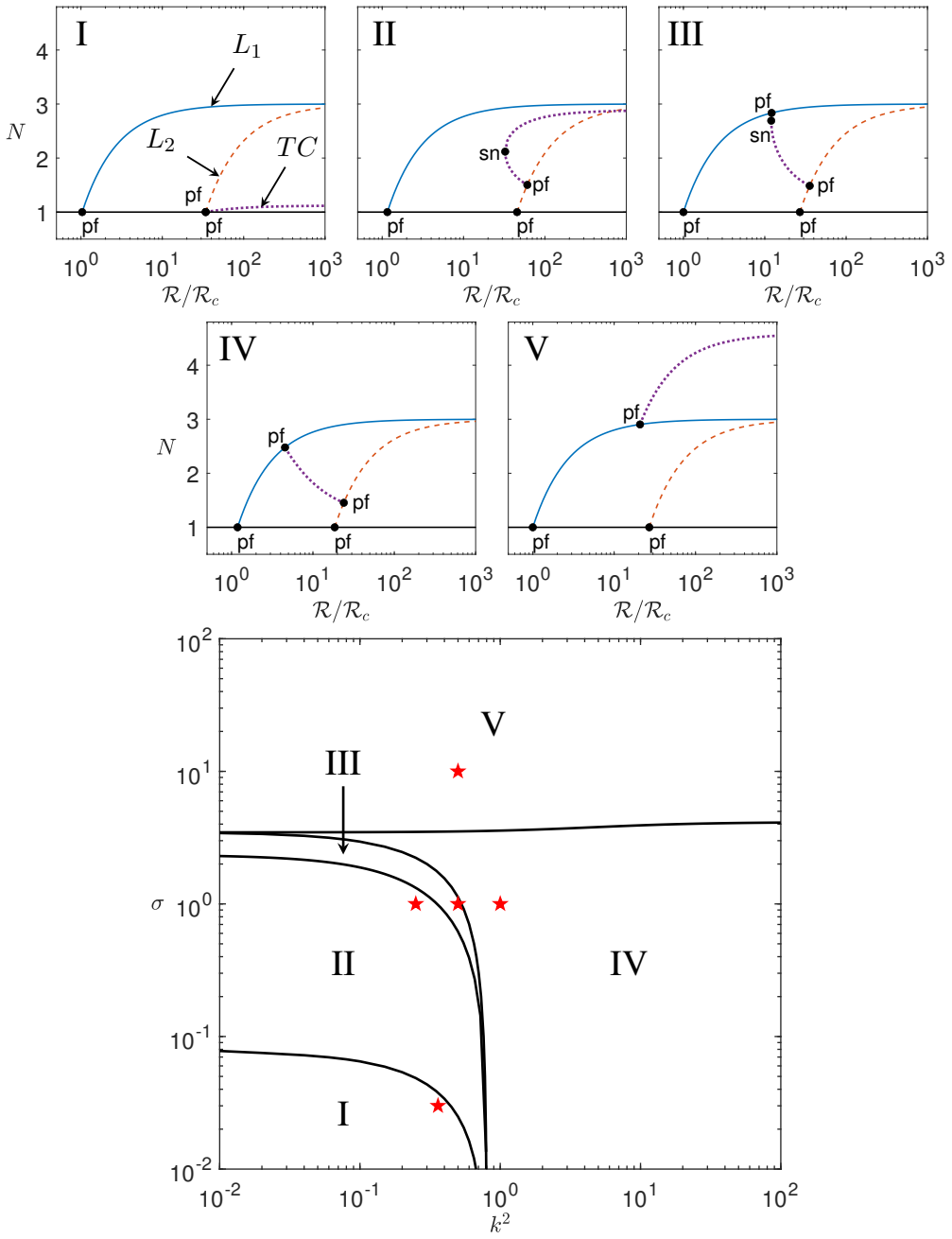


Figure 3.2: Examples of the five bifurcation structures of steady states of the HK8 model, with $\mathcal{R}$ as the bifurcation parameter (top). The k^2–σ parameter regimes where each bifurcation structure occurs are also shown and numbered correspondingly (bottom). Stars (★) in parameter space indicate the particular values for each example bifurcation diagram. Pitchfork ('pf') and saddle-node ('sn') bifurcations are labeled. Stability of steady states and locations of Hopf bifurcations are not indicated. In region I, the TC branch bifurcates from the L_2 branch at a Rayleigh number slightly larger than $\mathcal{R}_{L_2}$. The results in this figure were reported by [22] and independently verified here.

equations [41]. With the present variables (scaled differently from the Lorenz equations), the Hopf bifurcation exists when $\sigma > 1 + 4/(k^2+1)$ and occurs at $\mathcal{R}_{H_1}$, where

$$\frac{\mathcal{R}_{H_1}}{\mathcal{R}_{L_1}} = 1 + (\sigma+1)\frac{\sigma(k^2+1)+(k^2+5)}{\sigma(k^2+1)-(k^2+5)}. \tag{3.9}$$

Since the L_2 subspace is another rescaling of the Lorenz equations, its Hopf bifurcation is similar. It exists when $\sigma > 1 + 16/(k^2+4)$ and occurs at $\mathcal{R}_{H_2}$, where

$$\frac{\mathcal{R}_{H_2}}{\mathcal{R}_{L_2}} = 1 + (\sigma+1)\frac{\sigma(k^2+4)+(k^2+20)}{\sigma(k^2+4)-(k^2+20)}. \tag{3.10}$$

Additional Hopf bifurcations may occur on the TC branch, as detailed in [22].

Periodic states emerging from Hopf bifurcations of the L_1 and L_2 branches cannot produce larger heat transport than N_{L_1}. This is because these periodic states remain in their respective subspaces of Lorenz triplets, and in the Lorenz equations the truncated Nusselt number is maximized on the nonzero equilibria [23, 44]. It remains possible that time-dependent states involving all eight modes can have larger N than all steady states; we examine such states in the next section.

3.2 Time-dependent states

We numerically integrated (2.42) to search for attracting time-dependent solutions of the HK8 model for $(k^2, \sigma) = (1/2, 10)$. These parameter values lie in regime V of Figure 3.2, and they correspond to the standard choice $(\beta, \sigma) = (8/3, 10)$ in the Lorenz equations. This k value minimizes the Rayleigh number $\mathcal{R}_{L_1}$ of convective instability in both the HK8 model and the PDE. Numerical integration was carried out using MATLAB's `ode45` function with absolute and relative tolerances of 10^{-12} and 10^{-9}, respectively, and all other settings at their default values. The time-averaged Nusselt number (2.44) was computed by averaging periodic trajectories over several full periods and by averaging nonperiodic trajectories over 10^4 to 10^5 time units (after initial transients).

When $(k^2, \sigma) = (1/2, 10)$, the TC branch has subcritical Hopf bifurcations at $\mathcal{R} \approx 21.8\,\mathcal{R}_c$ and $\mathcal{R} \approx 999\,\mathcal{R}_c$. At values of $\mathcal{R}$ slightly beyond the first bifurcation, numerical integration with a variety of randomly generated initial conditions gives trajectories where all eight modes appear to be chaotic. The system exhibits period doubling bifurcations as $\mathcal{R}$ is increased from the first Hopf bifurcation of the TC branch and as $\mathcal{R}$

is decreased from $\mathcal{R} \approx 46\,\mathcal{R}_c$ into the nonperiodic regime, but the possibility of a period doubling cascade was not explored in detail. Figure 3.3 (top row) shows part of an apparently chaotic trajectory at $\mathcal{R} = 250$. We were able to find such trajectories for $\mathcal{R}/\mathcal{R}_c \in (21.8, 45)$ and again at $\mathcal{R} \gtrsim 1.8 \times 10^3\,\mathcal{R}_c$; between these two intervals, the only states we found using numerical integration are periodic. An example at $\mathcal{R} = 500$ is shown in Figure 3.3 (bottom row). Bistability was detected for $\mathcal{R} \gtrsim 1.8 \times 10^3\,\mathcal{R}_c$, where some initial conditions produced solutions that approached periodic trajectories, while others tended towards a nonperiodic attractor similar to those depicted in the top row of Figure 3.3. Such behavior was not identified for any smaller values of $\mathcal{R}$.

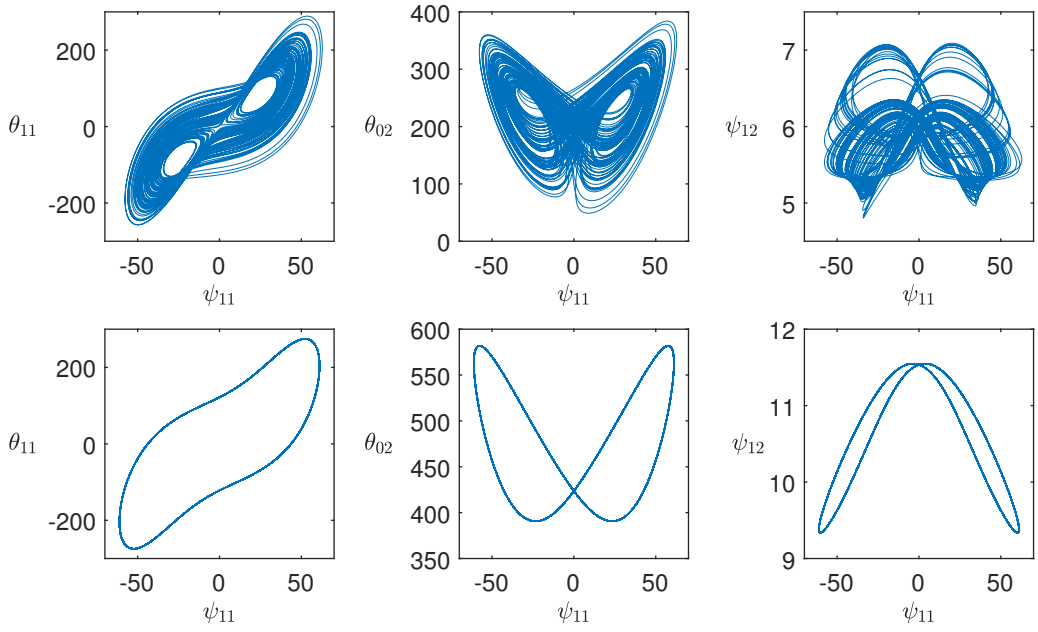


Figure 3.3: Evolution of the Lorenz modes $(\psi_{11}, \theta_{11}, \theta_{02})$ and the ψ_{12} mode in phase space for trajectories obtained by direct time integration of the HK8 system. The top row displays orbits of an apparently chaotic trajectory at $\mathcal{R} = 250$, while the bottom row depicts a stable periodic trajectory at $\mathcal{R} = 500$. In each case, all eight variables are generically nonzero along orbits.

Figure 3.4 shows N versus $\mathcal{R}$ for all steady states and time-dependent states found using time integration. The N values in the nonperiodic regime lie below the steady maximum, N_{TC}. As $\mathcal{R}$ is raised, the N values of the periodic states surpasses N_{TC}, meaning that heat transport is *not* maximized by a steady state at large $\mathcal{R}$. At such large $\mathcal{R}$, however, the HK8 model is not expected to closely reflect behavior of the PDE. At smaller $\mathcal{R}$ we did not find any time-dependent states with N larger than the steady state maximum. For $\mathcal{R} \lesssim 71\,\mathcal{R}_c$ the steady states indeed maximize heat transport, as follows from our sharp upper bounds

on N in §5.1 that are equal to $\max\{N_{L_1}, N_{TC}\}$.

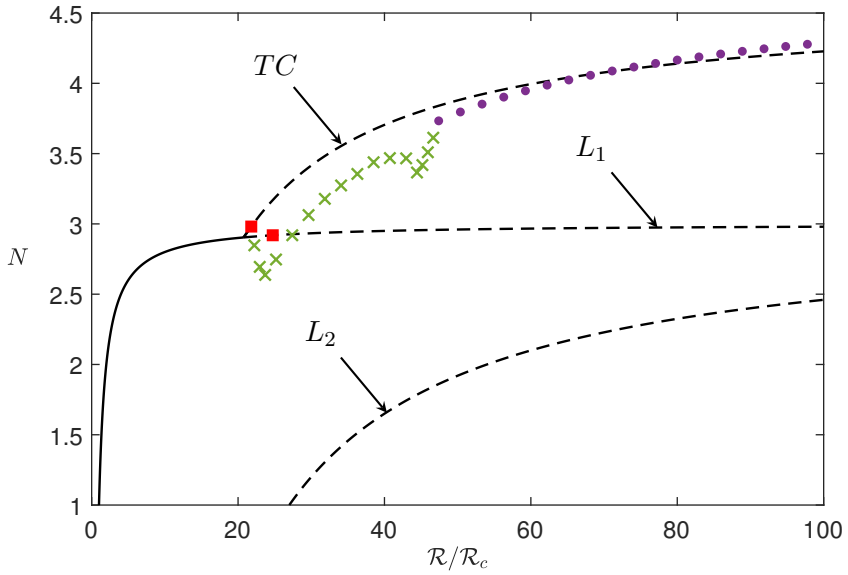


Figure 3.4: Nusselt number of steady and time-dependent solutions of the HK8 model for $(k^2, \sigma) = (1/2, 10)$. Solid lines denote linearly stable equilibria while dashed lines denote unstable equilibria. Locations of Hopf bifurcations are indicated by red squares (■). Symbols denote time averages over time-dependent states that are periodic (●) or appear to be chaotic (×).

CHAPTER 4

Auxiliary Functions, Convex Optimization, and Sums-of-Squares

4.1 Introduction

Time averages of functions of dynamical variables are often of interest for nonlinear dynamical systems, more so than the value of the function at any particular instant in time. In recent years, a general technique has been developed to determine upper or lower bounds on time-averaged quantities for ordinary differential equations [7]. Such results are global in the sense that they provide a bound on a given quantity over all solution trajectories with initial conditions in a given domain. These methods involve choosing an *auxiliary function* defined on the state space of the ODE that facilitate proving the desired bound. Using auxiliary functions to prove bounds on time averages is reminiscent of the common technique of using Lyapunov functions to determine global stability properties for ODEs. Like the functions used in Lyapunov's method, auxiliary functions need not have any particular relationship to the system dynamics. The auxiliary function method has been applied for a variety of systems [23, 24, 17], including modal equations of PDEs such as the models described in Chapter 2. We present an overview of the auxiliary function method in §4.2. The application to polynomial dynamical systems is discussed in §4.3, where bounds are computed numerically with techniques of convex optimization (§4.4). Finally, an application of the auxiliary function method to the Lorenz equations is provided as an example of the technique.

4.2 Maximal time averages for ODEs

Consider a well-posed autonomous ODE $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$, and $\dot{\mathbf{x}}$ denotes the time derivative of $\mathbf{x}$. Let $\mathbf{x}(t)$ be a trajectory of the ODE with initial condition $\mathbf{x}(0) = \mathbf{x}_0$. The long-time average of a scalar quantity $\Phi(\mathbf{x})$ along $\mathbf{x}(t)$ is given

by

$$\overline{\Phi}(\mathbf{x}_0) := \lim_{\tau\to\infty} \frac{1}{\tau} \int_0^\tau \Phi(\mathbf{x}(\tau))\,\mathrm{d}\tau. \tag{4.1}$$

In this chapter, we define $\Phi = 1 + 1/\mathcal{R}(2\theta_{02} + 4\theta_{04})$ so that $\overline{\Phi} = N$ (2.44), the finite-dimensional analogue of the horizontally averaged Nusselt number for Rayleigh–Bénard convection.

Computing (4.1) exactly is only possible when trajectories of the ODE are known exactly, such as at the equilibria. In practice, when formulae for the relevant trajectories are not explicitly known, the time average may be estimated by numerically integrating the ODE over a sufficiently large time and using the result to approximate the limit in (4.1). Estimates obtained in this way may converge slowly, and are restricted to only sufficiently stable trajectories stemming from a set of chosen initial conditions. It is generally impossible to perform this computation over all relevant initial conditions, especially in systems exhibiting chaotic behavior.

Auxiliary functions [7] allow another way to obtain information about time-averaged quantities for dynamical systems, without selecting a particular trajectory. The auxiliary function method provides bounds that are global in the sense that the bounds on (4.1) hold over all trajectories (or equivalently, over all initial conditions) of the ODE. Accordingly, define the maximal time average, $\overline{\Phi}^*$, by

$$\overline{\Phi}^* := \max_{\mathbf{x}_0\in\Omega} \overline{\Phi}(\mathbf{x}_0), \tag{4.2}$$

where $\Omega \subset \mathbb{R}^n$ is a compact domain containing the attracting region of the ODE, and seek a global upper bound $U \in \mathbb{R}$ so that $U \geq \overline{\Phi}^*$. While U must be independent of the choice of trajectory, it may depend on the values of various model parameters. For example, upper bounds on N will generally depend on $\mathcal{R}$, σ, and k.

A global upper bound could be constructed simply by maximizing Φ pointwise [75]:

$$\overline{\Phi}^* \leq \max_{\mathbf{x}\in\Omega} \Phi(\mathbf{x}). \tag{4.3}$$

However, in practice this will almost always produce bounds that are much larger than $\overline{\Phi}^*$, and will not produce a meaningful result unless Φ is bounded on Ω. More useful bounds can be determined by introducing an auxiliary function $V : \mathbb{R}^n \to \mathbb{R}$ in the class C^1 of continuously differentiable functions. Any such V remains bounded along bounded trajectories. This implies $\overline{\mathbf{f}\cdot\nabla V} = 0$ on every trajectory, where the gradient is with respect

to $\mathbf{x}$, since [23]

$$\overline{\mathbf{f}(\mathbf{x}(t)) \cdot \nabla V(\mathbf{x}(t))} = \overline{\tfrac{\mathrm{d}}{\mathrm{d}t} V(\mathbf{x}(t))} \tag{4.4}$$

$$= \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau \tfrac{\mathrm{d}}{\mathrm{d}t} V(\mathbf{x}(t)) \, \mathrm{d}t \tag{4.5}$$

$$= \lim_{\tau \to \infty} \frac{1}{\tau} \Big[V(\mathbf{x}(\tau)) - V(\mathbf{x}(0)) \Big] = 0. \tag{4.6}$$

This identity implies that given any initial condition $\mathbf{x}_0 \in \Omega$ and $V \in C^1$,

$$\overline{\Phi} = \overline{\Phi + \mathbf{f} \cdot \nabla V} \leq \max_{\mathbf{x} \in \Omega} \left[\Phi(\mathbf{x}) + \mathbf{f}(\mathbf{x}) \cdot \nabla V(\mathbf{x}) \right]. \tag{4.7}$$

This is useful because computing or estimating the right-hand maximum requires no knowledge of trajectories. While $\Phi(\mathbf{x})$ may be unbounded over Ω, a judicious choice of $V(\mathbf{x})$ makes the above maximum finite. Since (4.7) applies to all trajectories in bounded systems and for all $V \in C^1$, it remains true when we maximize $\overline{\Phi}$ over initial conditions and minimize the upper bound over V to find

$$\overline{\Phi}^* \leq \inf_{V \in C^1} \max_{\mathbf{x} \in \Omega} \left[\Phi(\mathbf{x}) + \mathbf{f}(\mathbf{x}) \cdot \nabla V(\mathbf{x}) \right]. \tag{4.8}$$

It was proved in [75] that for all bounded well-posed ODEs and continuous $\Phi(\mathbf{x})$, the inequality in (4.8) is an equality. Their result guarantees the existence of an auxiliary function (or sequence thereof) that yield arbitrarily sharp bounds on $\overline{\Phi}^*$. In practice, the infimum in (4.8) can often be attained, and we do so for the Lorenz equations in §4.6, and for the HK8 model in Chapter 5. A convenient way to express (4.8) is to define a function $S(\mathbf{x})$ as

$$S(\mathbf{x}) := U - \Phi(\mathbf{x}) - \mathbf{f}(\mathbf{x}) \cdot \nabla V(\mathbf{x}). \tag{4.9}$$

Then, an upper bound is implied by the nonnegativity of S, and the optimization problem (4.8) can be expressed as [75]

$$\overline{\Phi}^* = \inf_{\substack{V \in C^1 \\ S \geq 0}} U, \tag{4.10}$$

where $S \geq 0$ must hold for all $\mathbf{x} \in \Omega$. The equality (4.10) means that for every U that is a valid upper bound on $\overline{\Phi}^*$, there exists a sequence of auxiliary functions certifying this bound. The challenge is to construct such a V and verify that indeed $S \geq 0$.

4.3 Polynomial dynamical systems and sum-of-squares optimization

The right-hand side of (4.10) is an optimization problem over the infinite-dimensional space C^1, and determining the optimal auxiliary function is intractable in general. When the right-hand side of the ODE, $\mathbf{f}$, as well as the quantity Φ are each polynomial in $\mathbf{x}$, the problem can be made tractable by restricting the class of auxiliary functions [23]. The first step is to let V be a polynomial of degree no larger than d, giving an optimization problem over the finite-dimensional set $\mathbb{P}_{n,d}$ of such polynomials in n variables. Optimization over this smaller set sometimes gives a bound strictly larger than $\overline{\Phi}^*$, but if so the bound improves as d is raised.

Deciding whether a polynomial is non-negative over $\Omega \subset \mathbb{R}^n$ is generally NP-hard [48] in both the degree and dimension. This means that as either d or n are increased the computation time asymptotically increases faster than any polynomial. Therefore a modification of this constraint is necessary to make the problem tractable. We thus use a relaxation that has become standard for polynomial optimization since its introduction two decades ago [37, 49, 53]: nonnegativity of S is ensured by the stronger requirement that S admits a representation as a sum-of-squares (SOS) of other polynomials.

Let $\mathbb{P}^+_{n,d}$ be the space of non-negative, real-valued, degree-d polynomials in n variables, and $\Sigma_{n,d}$ be the analogous space of SOS polynomials. Clearly the degree d must be even for either set to be nonempty, and $\Sigma_{n,d} \subseteq \mathbb{P}^+_{n,d}$ for all n, d. Equivalence of these two sets holds in only a few cases [59]: for polynomials of at most two variables ($n \leq 2$), quadratic polynomials ($d \leq 2$), or quartic polynomials of three variables ($n = 3, d = 4$). In all other cases, there exist non-negative polynomials that do not admit an SOS representation. Techniques of polynomial optimization have been applied to prove global stability results by constructing Lyapunov functions [53], to identify the region of attraction for ODEs [54], and to determine global bounds on time-averaged quantities [7]. One benefit of using an SOS constraint is that deciding whether a polynomial is in $\Sigma_{n,d}$ can be performed in polynomial time in both n and d. An efficient algorithm for this purpose was developed in [55], based on theoretical work on SOS polynomials by Shor [62, 63].

If S is assumed to be SOS, then for each fixed degree d, the upper bound from the resulting polynomial optimization problem is [7, 17, 23]

$$\overline{\Phi}^* \leq U_d^* := \min_{V \in \mathbb{P}_{n,d}} U \quad \text{s.t.} \quad \mathrm{S} \in \Sigma_{\mathrm{n,d}}. \tag{4.11}$$

The SOS-constrained polynomial optimization problem on the right-hand side of (4.11) is computationally tractable if d and n are not too large. If a chosen degree d does not yield a sharp upper bound, raising d improves bounds at the expense of computational cost. In practice bounds often converge rapidly as the degree of V is raised [17, 23].

The usual computational approach to solving an SOS-constrained optimization problem of the form (4.11) is to reformulate it as a semidefinite program (SDP), a standard type of conic optimization problem. Most modern SOS algorithms utilize the Gram matrix form [8], wherein the polynomial S is represented as:

$$S = \mathbf{b}^T \mathcal{Q} \mathbf{b}, \tag{4.12}$$

where $\mathbf{b}$ is a vector of polynomial basis functions. The vector $\mathbf{b}$ must be chosen so that S is in the span of the scalar polynomial terms in $\mathbf{b}\mathbf{b}^T$, so that at least one $\mathcal{Q}$ satisfies (4.12). It can be shown that $S \in \Sigma_{n,d}$ if and only if there exists a basis vector $\mathbf{b}$ such that the resulting Gram matrix is symmetric positive semidefinite, written $\mathcal{Q} \succeq 0$ [55]. This reduces the task of deciding whether a polynomial is SOS to a convex optimization problem with a matrix constraint (an SDP). Efficient algorithms exist for solving SDPs, as discussed in §4.4.

Given a basis $\mathbf{b}$, the polynomial optimization (4.11) can be formulated equivalently as

$$\overline{\Phi}^* \leq U_d^* := \min_{V \in \mathbb{P}_{n,d}} U \quad \text{s.t.} \quad \begin{array}{l} S = \mathbf{b}^\mathsf{T} \mathcal{Q} \mathbf{b} \\ \mathcal{Q} \succeq 0. \end{array} \tag{4.13}$$

In the above optimization problem, the bound U and the coefficients of the polynomial V are tunable. The equality $S = \mathbf{b}^\mathsf{T} \mathcal{Q} \mathbf{b}$ is enforced by expanding the right-hand product and matching coefficients on each monomial term, amounting to affine constraints on the entries of $\mathcal{Q}$. Thus the optimization is over symmetric matrices $\mathcal{Q}$ subject to affine and semidefinite constraints. The tunable variables—U and the coefficients of the ansatz for V—appear linearly in the constraints on $\mathcal{Q}$. These two types of constraints on a semidefinite matrix are what define an SDP [4]. Various software is available to solve such SDPs computationally, and such results are reported in Chapter 4. Analytical solutions are possible in cases leading to very small SDPs, as in [23, 55].

4.4 Convex optimization and semidefinite programming

A semidefinite program (SDP) is a standard type of conic optimization problem with a matrix decision variable, where the feasible region is given by the intersection of a convex cone ($\mathcal{Q} \succeq 0$) and a hyperplane describing linear constraints ($S = \mathbf{b}^T \mathcal{Q} \mathbf{b}$). The objective function depends on the entries in $\mathcal{Q}$—in the case of (4.13), it is just the quantity U that appears in the constant term of $\mathbf{b}^T \mathcal{Q} \mathbf{b}$.

Various software is available to solve SDPs computationally, including the solver MOSEK [47]. In MOSEK, SDPs are solved using a primal-dual interior point method, as described below. An SDP can be written in the general form [54]:

$$\begin{aligned} \min \quad & \langle C, \mathcal{Q} \rangle, \\ \text{s.t.} \quad & \langle A_i, \mathcal{Q} \rangle = p_i, \\ & \mathcal{Q} \succeq 0, \end{aligned} \tag{4.14}$$

where $\mathcal{Q} \in \mathbb{R}^{n \times n}$ is the matrix decision variable, $A_i, C \in \mathbb{R}^{n \times n}$ and $\mathbf{p} \in \mathbb{R}^m$ are given, and the angled brackets represent the matrix inner product $\langle A, B \rangle = \sum_{i,j} A_{ij} B_{ij}$. The dual problem associated with (4.14) is given by

$$\begin{aligned} \max \quad & \mathbf{b}^T \mathbf{y}, \\ \text{s.t.} \quad & \sum_{i=1}^{m} y_i A_i \succeq C, \end{aligned} \tag{4.15}$$

for $\mathbf{y} \in \mathbb{R}^n$. The feasible sets for these two optimization problems are given by the set of all decision variables satisfying the constraints. In the typical case, where both feasible sets are nonempty, the SDP has the strong duality property: (4.14) and (4.15) have the same optimal solution [54]. Primal-dual algorithms take advantage of this property by solving both primal and dual problems simultaneously. Convergence can be obtained by solving the problem iteratively, terminating when the "duality gap" between the two solutions falls below a pre-determined threshold. Sum-of-squares optimization problems can be expressed in the form (4.14) with the MATLAB package YALMIP [38].

4.5 Scaling and monomial reduction

If the SDP corresponding to (4.11) is solved without simplification, computational cost and numerical ill-conditioning quickly become prohibitive as the polynomial degree is raised. Both aspects may be improved by restricting the ansatz for the auxiliary function V. Numerical conditioning can be improved by scaling the phase space variables in the governing ODE system.

Symmetry conditions can be used to restrict the ansatz for V. Suppose that both Φ and the ODE are invariant under a symmetry given by the linear transformation Λ, so that $\Phi(\Lambda\mathbf{x}) = \Phi(\mathbf{x})$ and $\mathbf{f}(\Lambda\mathbf{x}) = \Lambda\mathbf{f}(\mathbf{x})$. Then any bound proved using the auxiliary function method can be proved with symmetric V, so that $V(\Lambda\mathbf{x}) = V(\mathbf{x})$ [24, 36]. Symmetry reductions are convenient to implement for sign-symmetries of the variables in $\mathbf{x}$, where Λ is a diagonal matrix such that each diagonal entry is ± 1. Given Φ and $\mathbf{f}$, let monomials be represented in vector form by multi-indices $\boldsymbol{\alpha} \in \mathbb{Z}^n$, where

$$\mathbf{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}. \tag{4.16}$$

Next, let a symmetry Λ of the form described above be represented by $\mathbf{s} \in \mathbb{Z}_2^n$, with $s_i = 1$ if x_i is reflected under Λ, and $s_i = 0$ otherwise. For example, the vector $(1, 1, 0)$ corresponds to the symmetry $(x, y, z) \mapsto (-x, -y, z)$. Within this framework, a monomial is invariant under the symmetry Λ if and only if $\mathbf{s} \cdot \boldsymbol{\alpha} \equiv 0 \pmod 2$. This provides a computationally efficient way to determine if a candidate monomial in the general ansatz for V is symmetric. The set of symmetries for a given problem can be determined in much the same way. First, construct the matrix $\mathbf{A}$ whose rows are the multi-indices of each term in the polynomials $x_i f_i(\mathbf{x})$ and $\Phi(\mathbf{x})$. Any symmetry must then satisfy $\mathbf{A}\mathbf{s} \equiv 0 \pmod 2$.

The V ansatz can be further restricted by observing that the SOS constraint on S requires the highest-degree monomials in S to be of even degree. Such monomials generally come from the $\mathbf{f} \cdot \nabla V$ term in (4.9). If the highest-degree monomials in V are of even degree d, then for the reduced order models studied in this thesis, the polynomial $\mathbf{f} \cdot \nabla V$ generally includes terms of odd degree $d + 1$. Hence the SOS constraint can be satisfied only if the leading terms in V are constrained such that the highest-degree terms in $\mathbf{f} \cdot \nabla V$ cancel. This condition amounts to linear constraints on the coefficients of the highest-degree terms of V. In Chapters 5 and 6 these conditions are applied, leading to a significant reduction in the monomial basis. The smaller V ansatz leads to the same bounds and leads to smaller numerical error and decreased computational complexity.

Scaling the ODE variables has a significant impact on the numerical conditioning of the SDP computations. A common heuristic when using SOS optimization to study dynamical systems is to scale the state variables so that the relevant trajectories lie within the region $[-1, 1]^n$, as in [23, 28]. In practice, doing so often improves numerical conditioning of the SDP.

4.6 Example: Sum-of-squares bounds on the Lorenz equations

As a simple example of the auxiliary function method, we consider the Lorenz equations (2.3), and seek upper global upper bounds on the truncated analogue of the Nusselt number, given by $\Phi = 1 + \frac{2}{\rho}Z$. For $\rho \leq 1$, the origin is globally asymptotically stable, so we obtain the trivial bound $\Phi^* = 1$. When $\rho > 1$, the bound $\overline{\Phi} \leq 3 - \frac{2}{\rho}$ is proved using the sum-of-squares method with

$$V = \frac{1}{\rho\beta}\left[\frac{1}{\rho - 1}\left(\frac{X^2}{\sigma} + Y^2 + Z^2\right) - 2Z\right]. \tag{4.17}$$

To obtain the desired bound, notice that after some simplification, the polynomial S in (4.9) constructed from the above function V admits the sum-of-squares factorization

$$S = \left(3 - \frac{2}{\rho}\right) - \Phi - f \cdot \nabla V = \frac{2}{\beta\rho(\rho - 1)}(x - y)^2 + \frac{2}{\rho(\rho - 1)}(z - (\rho - 1))^2, \tag{4.18}$$

where $3 - \frac{2}{\rho}$ is the minimal value of U such that the polynomial S is SOS if and only if $\rho > 1$. In fact, this bound is achieved by the equilibria of the Lorenz equations,

$$(X, Y, Z) = \left(\pm\sqrt{\beta(\rho - 1)}, \pm\sqrt{\beta(\rho - 1)}, \rho - 1\right), \tag{4.19}$$

meaning that the equilibria maximize Φ for all $\rho > 1$. Both analytical and numerical sum-of-squares computations are performed in [23] for the Lorenz equations with various choices of Φ.

CHAPTER 5

Application of SOS Optimization to the HK8 Model

In this chapter, upper bounds on heat transport in the HK8 model (2.42) are constructed with the sum-of-squares method. This is done numerically in §5.1 with semidefinite programming techniques. Rigorous upper bounds are established in §5.2 by constructing quadratic auxiliary polynomials by hand that certify an upper bound on N^*. Results in this chapter were originally reported in [52].

5.1 Numerical upper bounds

To compute upper bounds on N^*—the maximum of N among all trajectories in the HK8 model—we numerically solved polynomial optimization problems of the form (4.13). In the definition (4.9) of S, the vector $\mathbf{f}$ is the right-hand side of the HK8 model (2.42), and we choose

$$\Phi = 1 + \tfrac{1}{\mathcal{R}} \left(2\theta_{02} + 4\theta_{04}\right), \tag{5.1}$$

so that $\overline{\Phi} = N$ according to (2.44). As discussed in Chapter 3, the two expressions for the Nusselt number, (2.43) and (2.44), are equivalent along all time-averaged trajectories. Upper bounds on the two quantities proved using the auxiliary function method are also identical. To see this, notice that if $V_0 := \theta_{02}/2 + \theta_{04}/4$, then

$$\tfrac{k}{4} \left(\psi_{11}\theta_{11} - \psi_{12}\theta_{12}\right) = 2\theta_{02} + 4\theta_{04} + \mathbf{f} \cdot \nabla V_0, \tag{5.2}$$

where the quantity on the left-hand side is the function whose time average corresponds to (2.43). Therefore, if an upper bound on (2.43) is obtained with the auxiliary function V, the same bound on (2.44) can be established with the auxiliary function $V + V_0$.

Following the notation established in Chapter 4, the resulting upper bound is $U_d^* \geq N^*$, with auxiliary functions V of polynomial degree $d = 2$, 4, 6, and 8. We first discuss details

on the implementation of the auxiliary function method, and discussion of the results begins in §5.1.2.

5.1.1 Numerical procedure

We first detail the specific monomial reduction and scaling implemented to produce the results of §5.1. When solving the optimization problem in (4.13) to find the bound U_d^*, we need not apply a fully general polynomial ansatz for V because some structure of V can be anticipated by examining the structure of the HK8 model. Restricting the V ansatz accordingly improves numerical conditioning and reduces computational cost. One source of structure that may be exploited is symmetry. The ODE (2.42) and the quantity to be bounded (5.1) are each invariant under both of the following symmetries:

$$\begin{aligned}(\psi_{11},\psi_{01},\psi_{12},\theta_{11},\theta_{02},\theta_{12},\psi_{03},\theta_{04}) &\mapsto (-\psi_{11},\psi_{01},-\psi_{12},-\theta_{11},\theta_{02},-\theta_{12},\psi_{03},\theta_{04}),\\ (\psi_{11},\psi_{01},\psi_{12},\theta_{11},\theta_{02},\theta_{12},\psi_{03},\theta_{04}) &\mapsto (\psi_{11},-\psi_{01},-\psi_{12},\theta_{11},\theta_{02},-\theta_{12},-\psi_{03},\theta_{04}).\end{aligned} \tag{5.3}$$

Without loss of generality, we impose these same symmetries on the V ansatz, since this does not change the optimal bounds U_d^* [24, 36]. The first symmetry in (5.3) dictates that the degrees of $\psi_{11}, \theta_{11}, \psi_{01}$, and ψ_{03} have an even sum in each monomial of the V ansatz, and the second symmetry in (5.3) dictates the same for $\psi_{12}, \theta_{12}, \psi_{01}$, and ψ_{03}. The symmetric monomials take the form

$$(\psi_{01}\psi_{03})^{d_1}(\psi_{11}\theta_{11})^{d_2}(\psi_{12}\theta_{12})^{d_3}\,\theta_{02}^{d_4}\,\theta_{04}^{d_5}\,(p(\mathbf{x}))^2, \tag{5.4}$$

where $d_1, \ldots d_5$ are non-negative integers and $p(\mathbf{x})$ is a polynomial of the HK8 variables. Since Φ, V, and the ODE share the same symmetries, the polynomial S defined by (4.9) does also. The Gram matrix representing S therefore can be written in block diagonal form, for a properly ordered polynomial basis vector, without changing the optimum of the SDP. This is automated by YALMIP, and in §5.2 it is illustrated explicitly for the case of quadratic V. During computation, the blocks are treated separately, often resulting in a significant decrease in both time and memory cost.

The second structural constraint on V comes from the requirement that the highest-degree terms in the polynomial $\mathbf{f} \cdot \nabla V$ be of even degree—a necessary condition for the SOS constraint in (4.13) to be satisfied. In general one expects an odd maximum degree of $d+1$ since $\mathbf{f}$ is quadratic. To avoid this we require that the highest-degree terms cancel in

$\mathbf{f} \cdot \nabla V$. This imposes linear constraints on the highest-degree terms in V that we encode into the V ansatz. In the present application to the HK8 model, these linear constraints imply that the degree d terms take the form

$$(\psi_{01}\psi_{03})^a (q(\mathbf{x}))^2, \tag{5.5}$$

where a is a non-negative integer and $q(\mathbf{x})$ is a polynomial of degree $(d-a)/2$. This condition on V, along with the imposed symmetry, gives an ansatz with far fewer than the $\binom{8+d}{d}$ monomials in a general degree-d polynomial in eight variables, as summarized in Table 5.1. In theory the smaller V ansatz leads to the same bounds, but in practice it approximates these bounds with less numerical error, as well as lower computational cost.

Table 5.1: Number of monomials in the ansatz for the auxiliary function V of degree d before and after reducing the ansatz using the structure of the HK8 model. The number of monomials before reduction is $\binom{8+d}{d}$. Average computation time is reported when the memory cost is not prohibitive. In the unreduced case, some monomial reduction is automated by the solver.

	Unreduced		Reduced	
d	Monomials	Time (s)	Monomials	Time (s)
2	45	0.3	11	0.3
4	495	2	88	0.5
6	3003	180	488	12
8	12870	-	2084	4300
10	43758	-	7251	-

Scaling the ODE variables has a significant impact on the numerical conditioning of the SDP computations. A common heuristic applied in SOS optimizations of dynamical systems is to scale the state variables so that the relevant trajectories lie within the region $[-1,1]^n$ [23, 28]. The variable scalings for the HK8 system were determined empirically using a combination of time integration and SOS bounds on the time average of each state variable. To achieve the desired scaling across a wide range of parameter values, the two Lorenz triplets $\{\psi_{11}, \theta_{11}, \theta_{02}\}$ and $\{\psi_{12}, \theta_{12}, \theta_{04}\}$ were scaled by their values at the L_1 (3.2) and L_2 (3.5) steady states, respectively. The remaining variables, ψ_{01} and ψ_{03}, were scaled by $\sqrt{\mathcal{R}}$ and $\sqrt{\mathcal{R}}/27$, respectively, motivated by their values at the TC equilibria. For many computations, all variables were scaled down further, typically by a factor of two, to ensure

that trajectories remained within $[-1,1]^8$.

After defining the ansatz for V with the above conditions and scaling the ODE variables as described above, we formulated the SOS problems of the form (4.11) as SDPs with the MATLAB software YALMIP [38, 39] (version R20190425). The resulting optimization problems were then solved using the optimizer MOSEK (version 9.0.98) [47]. Most computations were performed on a 3.0 GHz Intel Xeon processor, with some smaller problems solved on a laptop with a 2.2 GHz Intel i5 processor.

Several criteria were used to assess the accuracy of the SDP results. One such criterion is that the residual r is small, where r is defined as the largest coefficient in absolute value of any monomial in the difference $S - \mathbf{b}^{\mathsf{T}}\mathcal{Q}\mathbf{b}$. While the Gram matrix returned by the solver must be semidefinite at each iteration, in practice the polynomial $\mathbf{b}^T\mathcal{Q}\mathbf{b}$ does not exactly match the polynomial S, violating the equality constraint in (4.13). As a result, computed upper bounds will often slightly underestimate the solution to the SOS problem. Sufficiently small residuals indicate that the Gram matrix provides an accurate approximation to the polynomial S, implying this infeasiblity error is small. Often, larger residuals indicate poor numerical conditioning due to improperly scaled state variables. Residuals for a few selected cases are given by the quantity r in Table 5.2.

One way to directly assess the accuracy is to compare the numerical upper bounds to the maximal Nusselt number N^* in cases where the bound is sharp. For instance, analytical bounds obtained in §5.2 provide an interval of $\mathcal{R}$ (e.g. at the standard parameters with $\mathcal{R} \leq 14\,\mathcal{R}_c$) where all bounds constructed with V of degree at least two must be sharp. In such cases, a measure of the accuracy of the bounds is given by

$$\delta_1 := \frac{|U - L|}{L}, \tag{5.6}$$

where L is the lower bound on N^* determined by the largest N among all known particular solutions. In the optimal case, $L = N^*$ and small δ_1 implies sharpness of the bounds. When the bounds are not sharp, δ_1 simply provides the gap between upper and lower bounds on N^*. Another way to assess the accuracy of the SDP is to compute the difference between bounds computed using the two equivalent definitions for N at the same set of parameters. If U_1 and U_2 denote bounds on Φ corresponding to (2.43) and (2.44), respectively, then the relative difference is

$$\delta_2 := \frac{|U_2 - U_1|}{U_1}. \tag{5.7}$$

We used each of the three quantities defined above—r, δ_1, and δ_2—as diagnostic tools

to assess the validity of the bounds obtained via SDP. These quantities are displayed in Table 5.2.

Table 5.2: Upper bounds (U) on the Nusselt number computed using SOS optimization with auxiliary polynomials of maximum degree two up to eight. The upper bounds are compared to the maximum Nusselt number among all known solutions (L). The residual r is displayed as defined above, along with δ_1 and δ_2 as defined in (5.6) and (5.7).

	$\mathcal{R} = 10, \quad L = 1.65$				$\mathcal{R} = 100, \quad L = 2.865$			
d	U	$\log_{10}\delta_1$	$\log_{10}\delta_2$	$\log_{10} r$	U	$\log_{10}\delta_1$	$\log_{10}\delta_2$	$\log_{10} r$
2	1.6500000	-12.5	-12	-13.1	2.8684439	-2.9	-14.1	-13.1
4	1.6500000	-14	-14	-12.6	2.8650000	-15.1	-16	-13.2
6	1.6500000	-13.2	-13.4	-10.8	2.8650000	-14.2	-14.9	-12.2
8	1.6500000	-16	-12.9	-9.7	2.8650000	-16	-13	-11.2
	$\mathcal{R} = 500, \quad L = 4.1158818$				$\mathcal{R} = 1000, \quad L = 4.4442202$			
	U	$\log_{10}\delta_1$	$\log_{10}\delta_2$	$\log_{10} r$	U	$\log_{10}\delta_1$	$\log_{10}\delta_2$	$\log_{10} r$
2	4.2815436	-1.4	-14.2	-13.9	4.6304838	-1.4	-13.1	-13
4	4.2099533	-1.6	-9.5	-10.4	4.5377972	-1.7	-11.9	-10.6
6	4.1158840	-6.3	-6.2	-9.1	4.4442203	-8	-8.6	-9.1
8	4.1158827	-6.7	-8.4	-8.7	4.4442202	-8.6	-7.7	-8.5
	$\mathcal{R} = 5000, \quad L = 4.5887705$				$\mathcal{R} = 10000, \quad L = 4.4941571$			
	U	$\log_{10}\delta_1$	$\log_{10}\delta_2$	$\log_{10} r$	U	$\log_{10}\delta_1$	$\log_{10}\delta_2$	$\log_{10} r$
2	4.9247139	-1.1	-13.1	-12.7	4.9622777	-1	-14.6	-13.9
4	4.8410437	-1.3	-10.7	-10.8	4.8940532	-1	-11.6	-10.2
6	4.7433890	-1.5	-7.9	-8.1	4.8386286	-1.1	-9.3	-8.4
8	4.6540188	-1.8	-4.6	-6	4.7949268	-1.2	-4.3	-5.2

5.1.2 Upper bounds at the standard parameters

As a first example we fix $k^2 = 1/2$ and $\sigma = 10$, and consider the dependence of the upper bound on $\mathcal{R}$. In this case, the first bifurcation from the zero state, $\mathcal{R}_{L_1}$, takes its minimal value $\mathcal{R}_c = 27/4$. Figure 5.1(a) shows the upper bounds we computed in this case using SOS methods. Also shown are lower bounds on N^* found by searching among various trajectories of the HK8 system as described in Chapter 3. Close agreement of upper and

lower bounds on N^* implies that the upper bounds are sharp or very nearly so, and that the corresponding trajectories maximize N. The relative gap between upper bounds and lower bounds on N^* established in this chapter are depicted in Figure 5.1(b); for $\mathcal{R} \lesssim 560\,\mathcal{R}_c$, the best upper and lower bounds agree to at least five significant digits.

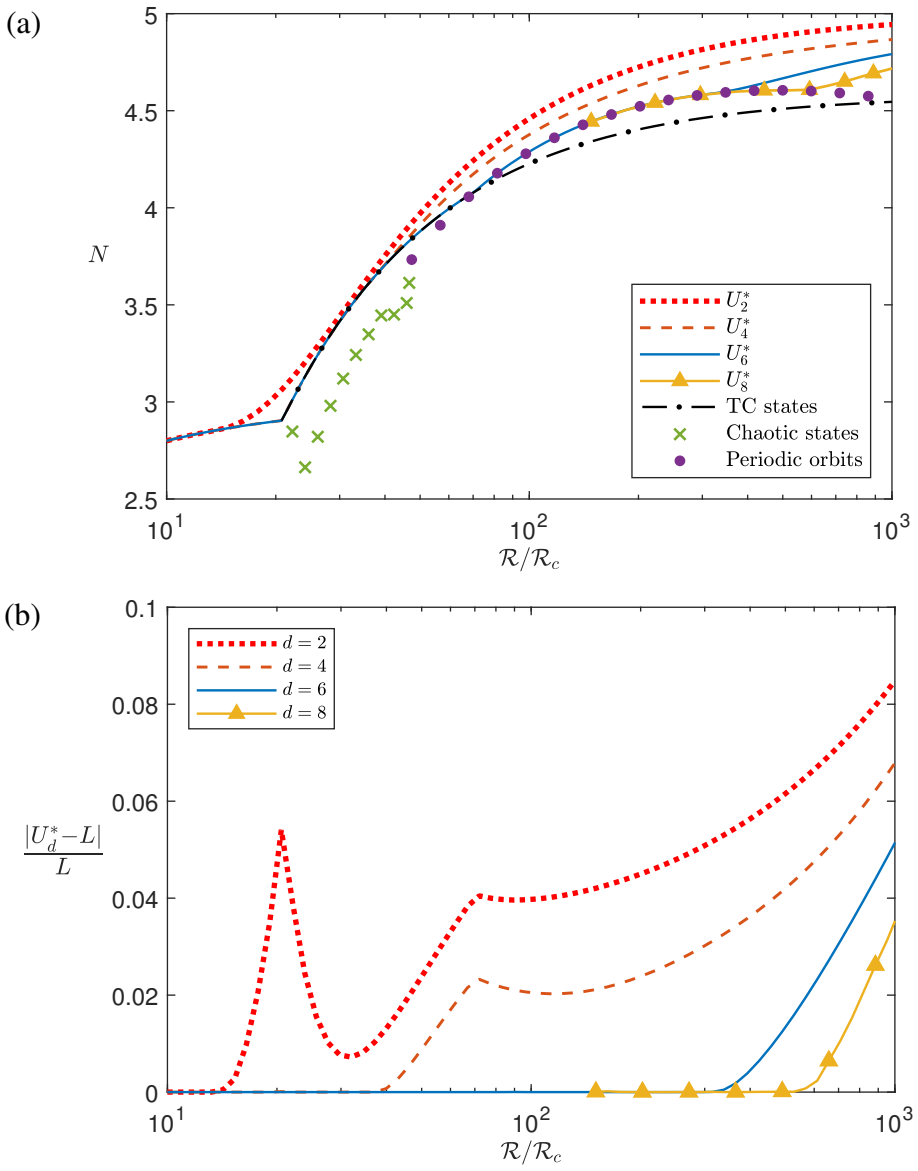


Figure 5.1: (a) Upper bounds (U_d^*) on the truncated Nusselt number (N) computed by solving the SOS optimization (4.13) with degree-d auxiliary polynomials, compared to N on particular solutions of the HK8 model with $k^2 = 1/2$ and $\sigma = 10$. (b) Relative difference between the U_d^* and the lower bound L, determined by finding the maximum N over the known particular solutions obtained in Chapter 3.

Different trajectories saturate the upper bounds (and hence maximize N) over various $\mathcal{R}$

intervals. When $\mathcal{R} \leq \mathcal{R}_{L_1}$, all trajectories satisfy $N = 1$ since they tend to the equilibrium at the origin. On the subsequent interval $\mathcal{R}_{L_1} \leq \mathcal{R} \leq \mathcal{R}_{TC_1} \approx 20.8\,\mathcal{R}_c$, the L_1 equilibria maximize N. At the present k^2 and σ, $N_{TC} > N_{L_1}$ for all $\mathcal{R} > \mathcal{R}_{TC_1}$, and TC equilibria are maximal at least on the interval $\mathcal{R}_{TC_1} \leq \mathcal{R} \lesssim 69\,\mathcal{R}_c$. Time-dependent states appear to maximize N for larger $\mathcal{R}$, with periodic orbits saturating the upper bound on the interval $69\,\mathcal{R}_c \lesssim \mathcal{R} \lesssim 520\,\mathcal{R}_c$. We draw this conclusion because N on the periodic orbits agrees with the best upper bound to within the expected numerical error of our SOS computations. In such cases we say for simplicity that the periodic orbit saturates the bound. Strictly speaking we do not expect periodic orbits to exactly saturate a bound computed with V of any finite degree d, as explained in [23], but we ignore this distinction provided U_d^* is sufficiently converged to the large-d limit. The branch of periodic orbits that saturates the upper bound is the one that emerges, initially unstable, from the Hopf bifurcation at $\mathcal{R}_{H_1} \approx 21.80\,\mathcal{R}_c$. For $\mathcal{R}$ larger than $520\,\mathcal{R}_c$, time-dependent solutions may still maximize N, but our upper bounds on N^* are not sufficiently sharp to confirm it.

5.1.3 Dependence on wavenumber and Prandtl number

We now examine how the upper bounds and the states that saturate them depend on the parameters k^2 and σ. When k^2 is fixed to values other than $1/2$ with σ still fixed at 10, the bounds are qualitatively similar to those depicted in Figure 5.1. We have computed upper bounds on N^* at various wavenumbers and searched among known trajectories for the largest N values. Figure 5.2 shows the upper bounds we computed at five different wavenumbers using V of degree up to eight. The different line styles in Figure 5.2 indicate the type of state that appears to saturate the upper bounds at various $\mathcal{R}$ and k. As in the $k^2 = 1/2$ case, each bound is saturated by L_1 equilibria at small $\mathcal{R}$, by TC equilibria at larger $\mathcal{R}$, and—at least in the smaller-k cases—by periodic orbits at still larger $\mathcal{R}$. The fact that L_1 states maximize N at onset is proved analytically in [66] and in §5.2 below. The Rayleigh number where the TC branch of equilibria bifurcates from the L_1 branch changes with k^2 according to (3.6), but in each case the emerging TC states saturate the bound for some interval of Rayleigh number.

Similarly, we may consider how the situation changes when σ is fixed to various values while $k^2 = 1/2$. The analytical bound proved in §5.2.2 below implies that for $k^2 = 1/2$, the σ-independent L_1 states maximize N for $\mathcal{R}$ at least as large as $\mathcal{R}_T(2^{-1/2}) = 14\,\mathcal{R}_c$, where $\mathcal{R}_T(k)$ is defined by (5.11) below. For $\mathcal{R} > \mathcal{R}_T$, we computed numerical upper bounds on N for various $\sigma \in [0.01, 100]$. In each case, the upper bounds appear to be saturated

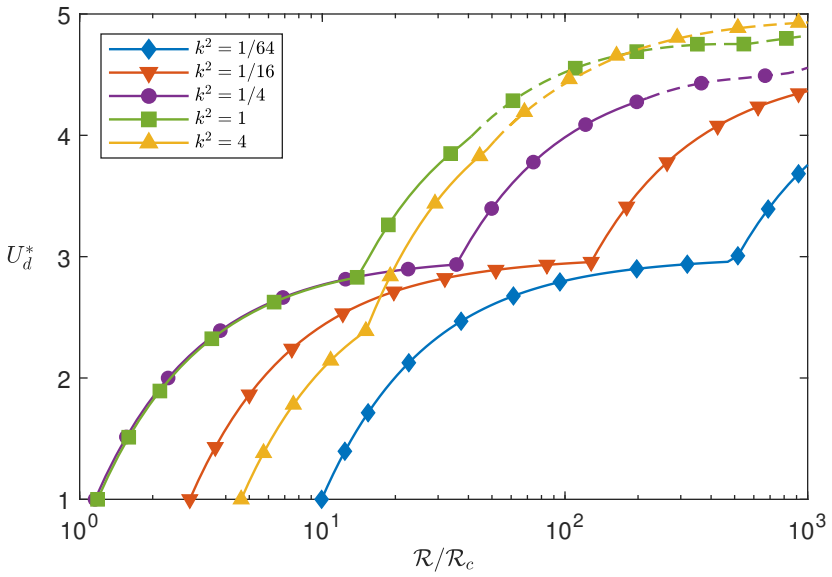


Figure 5.2: Upper bounds on N computed by solving the polynomial optimization problem (4.13) with V of degree up to eight. Upper bounds were computed for various k with $\sigma = 10$ across a range of $\mathcal{R}$. The line style indicates whether the upper bound is saturated by L_1 or TC equilibria (——) or a time-periodic orbit (- - -). For $k^2 = 1/4$ and $k^2 = 1$ sharp upper bounds at very large $\mathcal{R}$ were not confirmed, since the upper bounds of degree eight do not match the maximal N of the time-dependent solutions.

by time-dependent solutions for all $\mathcal{R}$ sufficiently large. The main qualitative distinction between different σ values is whether TC states saturate the upper bounds for some interval of $\mathcal{R}$. We observe in Chapter 3 that N_{TC} exceeds N_{L_1} if and only if (k^2, σ) lies in Region V of Figure 3.2, corresponding to $\sigma \gtrsim 3.523$ when $k^2 = 1/2$. Indeed, for various $\sigma > 3.523$ our upper bounds are saturated by TC states over bounded intervals of $\mathcal{R}$, whereas for smaller σ we found no such intervals.

5.1.4 Optimal wavenumber

In the full PDE model (1.13) of 2D Rayleigh–Bénard convection, steady solutions of each horizontal period exist for sufficiently large $\mathcal{R}$. Thus for the PDE it is natural to search among all horizontal periods for the steady states that maximize heat transport. The analogue in the HK8 model is to maximize N over k—the horizontal wavenumber of modes that are included in the truncated Galerkin expansion (2.41). Thus we consider the quantity

$$N_{k^*}^*(\mathcal{R}, \sigma) := \sup_{k>0} N^*(\mathcal{R}, k, \sigma). \tag{5.8}$$

To find upper bounds on $N^*_{k^*}$, we repeatedly solved the SOS optimization in (4.13) using V of degree 6, sweeping through k and using the MATLAB function `fminsearch` to converge to the global maximum of U^*_6 over k. Lower bounds on $N^*_{k^*}$ were computed by maximizing N over k among each of the three types of states observed to maximize N at various parameters: the L_1 equilibria, the TC equilibria, and stable periodic orbits found via time integration. For the L_1 states, the optimal wavenumber is $(k^*)^2 = 1/2$, corresponding to convection cells of aspect ratio $2\sqrt{2}$. Here N_{L_1} attains its maximum of

$$\max_{k>0} N_{L_1} = 3 - \frac{27}{2\mathcal{R}}, \tag{5.9}$$

in the regime $\mathcal{R} > \mathcal{R}_c$ where these L_1 states exist. The maximum of N_{TC} over k was found with Mathematica by numerically optimizing the analytical expressions for the TC equilibria at various fixed values of $\mathcal{R}$ and σ. At each of these $\mathcal{R}$ and σ values, the maximum value of N over k among stable periodic orbits was determined using a search algorithm similar to the one used to maximize the upper bound, with numerical integration performed to determine N at each iteration. Figure 5.3 shows the upper and lower bounds on $N^*_{k^*}$ in the $\sigma = 10$ case. The upper bounds are sharp or very close to sharp over the full range of $\mathcal{R}$. At $\sigma = 10$, the states that saturate or nearly saturate these bounds are the L_1 equilibria at small $\mathcal{R}$, the TC equilibria at intermediate $\mathcal{R}$, and the stable periodic orbits at larger $\mathcal{R}$. In particular, the maximum value $N^*_{k^*}$ is attained by steady states when $\mathcal{R} \lesssim 7.39\,\mathcal{R}_c$.

To study the effects of changing σ, at various fixed σ we determined the intervals of $\mathcal{R}$ where our k-maximized upper bounds are saturated by the L_1 equilibria, TC equilibria, or time-periodic states. Figure 5.4 summarizes the results. At each σ, we used a bisection search to find the largest $\mathcal{R}$ such that an L_1 equilibrium saturates the k-maximized upper bound. That is, we computed the k-maximized upper bound and increased $\mathcal{R}$ when the upper bound differed from the maximum of N_{L_1} by less than 10^{-5}, and we decreased $\mathcal{R}$ otherwise. Analogous computations for TC equilibria were carried out to find the boundary between regions II and III.

Regions I and II together in Figure 5.4 comprise the parameter regime in the HK8 model where, according to our upper bounds, the maximum of N over all k is attained by steady states. It is an open question whether steady states maximize heat transport in the Boussinesq equations [80]. Results for the PDE suggest that the HK8 model provides heat transport similar to that of the PDE for only a small interval of Rayleigh number. When $\mathcal{R} \gtrsim 2\mathcal{R}_c$, the value of N_{L_1} deviates from the heat transport of the analogous roll state of the PDE (see

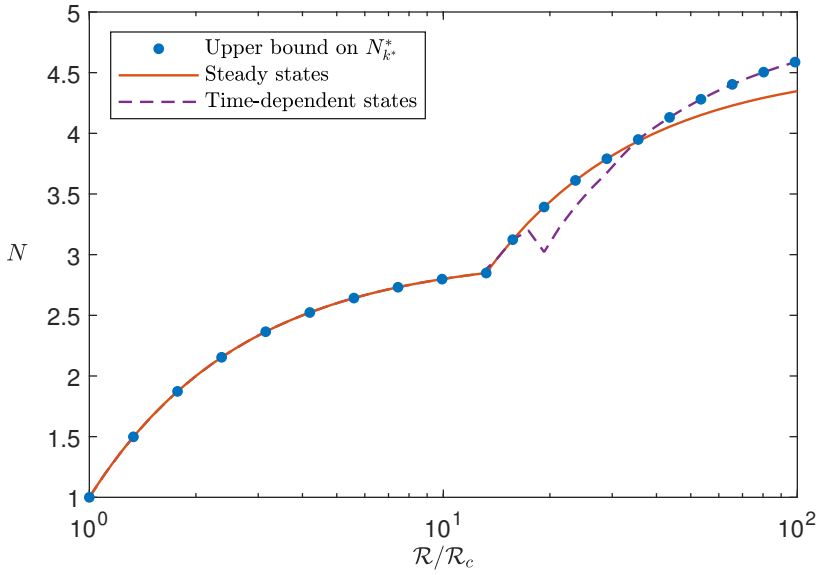


Figure 5.3: Comparison of three k-maximized quantities at $\sigma = 10$: upper bounds on N^*, N values among steady states, and N values among time-dependent states. The maximizer of the state saturating the upper bound, k^*, generally depends on $\mathcal{R}$ and the type of trajectory. Upper bounds were computed using V of degree six.

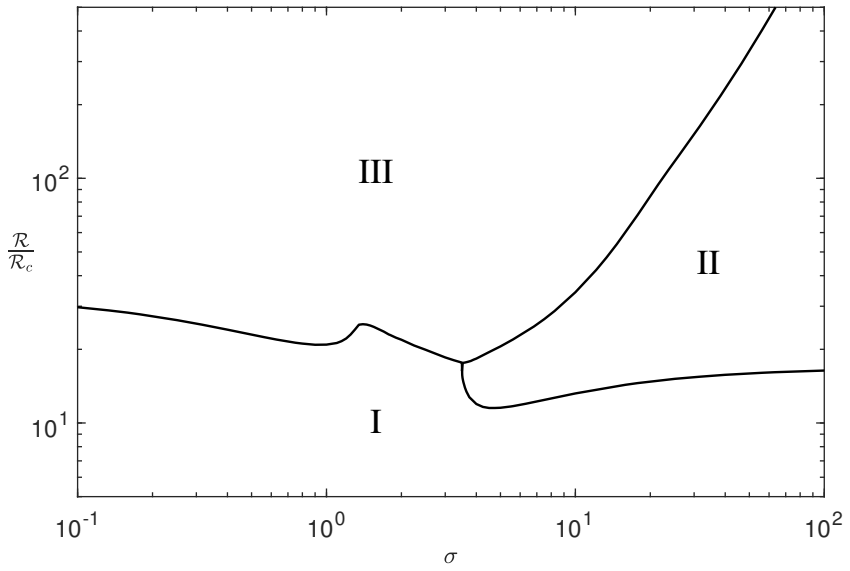


Figure 5.4: Regions in the σ–$\mathcal{R}$ plane where the upper bound on $N^*_{k^*}$—the maximum heat transport over k—is saturated by: (I) L_1 equilibria and (II) TC equilibria. In region (III), time-periodic states appear to saturate the upper bounds, but the upper bounds with V of degree six are not sufficient to prove this at all parameter combinations. The intersection between the three regions occurs near $\sigma = 3.5$ and $\mathcal{R}/\mathcal{R}_c = 17.6$, corresponding to the minimal σ where the Nusselt number of the TC equilibria exceed that of L_1 for sufficiently large $\mathcal{R}$.

[80]). For the HK8 model at all σ and all $\mathcal{R}$ values small enough for the model to capture PDE behavior, Figure 5.4 suggests that steady states maximize N.

5.2 Analytical upper bounds with quadratic auxiliary functions

In principle the bounding framework (4.11) can be applied numerically or analytically, but many of the bounding computations reported in §5.1 would be analytically intractable because the polynomial expression (4.9) for S has hundreds or thousands of terms. Bounds can be derived analytically in the case of quadratic V, however, and we do so in this section. The resulting bounds are rigorous and depend analytically on the parameters $(\mathcal{R}, k, \sigma)$, whereas the numerical bounds in §5.1 were subject to rounding errors in SDP solutions, and they had to be computed anew for each triplet of parameter values.

The best analytical upper bounds on N take different forms in four different regimes of the k–$\mathcal{R}$ parameter plane. These four regimes are shown in Figure 5.5, and the bounds in each are

$$N \leq \begin{cases} 1, & 0 \leq \mathcal{R} \leq \mathcal{R}_{L_1}, \\ N_{L_1}, & \mathcal{R}_{L_1} < \mathcal{R} \leq \mathcal{R}_T, \\ N_{L_1} + \frac{1}{\mathcal{R}} \left[\mathcal{R}_{L_1} - \mathcal{R}_{L_2} + \sqrt{2}\sqrt{(\mathcal{R} - \mathcal{R}_{L_1})^2 + (\mathcal{R} - \mathcal{R}_{L_2})^2} \right], & \\ & \mathcal{R} > \mathcal{R}_T, \ 0 < k \leq k_T, \\ N_{L_1} + \frac{1}{\mathcal{R}} \left[\mathcal{R}_{L_1} - \mathcal{R}_{L_2} + \sqrt{(\mathcal{R}_{L_2} - \mathcal{R}_{L_1})^2 + 4(\mathcal{R} - \sqrt{\mathcal{R}_{L_1}\mathcal{R}_{L_2}})^2} \right], & \\ & \mathcal{R} > \mathcal{R}_T, \ k > k_T, \end{cases} \tag{5.10}$$

Recall that $\mathcal{R}_{L_1}(k)$ and $\mathcal{R}_{L_2}(k)$ are defined by (3.1) and (3.4), and that $N_{L_1} = 3 - 2\,\frac{\mathcal{R}_{L_1}}{\mathcal{R}}$. The character of the bounds changes at $\mathcal{R}_T$, defined by

$$\mathcal{R}_T(k) := \begin{cases} \frac{1}{2}\left(\mathcal{R}_{L_1} + \mathcal{R}_{L_2}\right), & 0 \leq k \leq k_T, \\ \frac{-15}{2(5k^2-4)}\mathcal{R}_{L_1} + \sqrt{\frac{11+5k^2}{5k^2-4}\mathcal{R}_{L_1}\mathcal{R}_{L_2}}, & k > k_T, \end{cases} \tag{5.11}$$

where $k_T \approx 1.00319$ is the positive real root of $(5k^2 + 11)\mathcal{R}_{L_1} = (5k^2 - 4)\mathcal{R}_{L_2}$. Note that the bounds (5.10) are uniform in σ, unlike the bounds reported in §5.1 that were computed numerically using V of degrees 4 and higher.

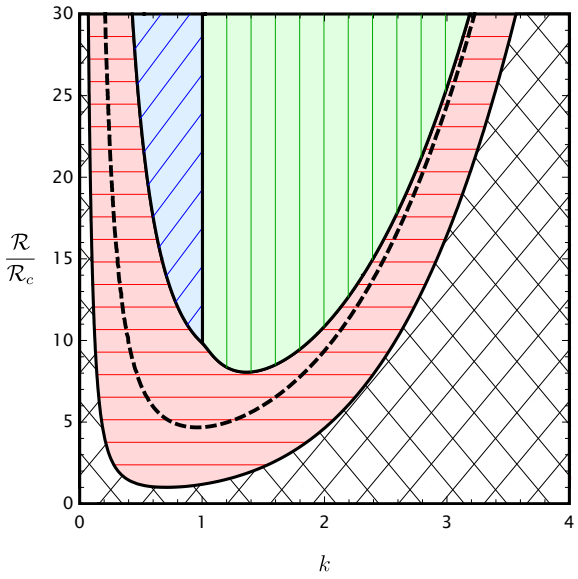


Figure 5.5: Parameter regimes where the four different analytical upper bounds (5.10) on N are proved for the HK8 model. In the cross-hatched region, $N = 1$ for all solutions. The bound $N \leq N_{L_1}$ is proved here for the full horizontally hatched region; it was proved in [66] only for the part of this region below the dashed line (- - -). The diagonally and vertically hatched regions correspond to the third and fourth cases in (5.10), respectively.

The bound in the first regime of (5.10) is sharp and is saturated by the zero equilibrium, since this state is globally attracting below the first instability at $\mathcal{R}_{L_1}$. The bound in the second regime of (5.10) is saturated by the L_1 equilibria. The same bound was proved by Souza & Doering [66] on the strictly smaller parameter regime where $\mathcal{R}_{L_1} < \mathcal{R} \leq \sqrt{\mathcal{R}_{L_1}\mathcal{R}_{L_2}}$. In §5.2.2 we strengthen their result by extending its applicability up to the larger parameter value $\mathcal{R}_T$. The bound in the third regime of (5.10) is new and is proved in §5.2.2. The bound in the fourth regime was proved by Souza & Doering for $\mathcal{R} \leq \sqrt{\mathcal{R}_{L_1}\mathcal{R}_{L_2}}$. Their results are superseded in the second and third regimes by our new bounds. It is possible to improve the bound in the fourth regime using quadratic V, as suggested by the bounds we have computed numerically, but we were unable to derive an analytical expression; partial analytical results are given in Appendix C. Bounds in the third and fourth regimes of (5.10) are not sharp; it is evident from the numerical bounds reported in the previous section that V of higher polynomial degrees provide better bounds.

5.2.1 Sum-of-squares construction in the quadratic case

The quadratic ansatz that we consider for the auxiliary function V need not be the most general possible since some structure can be anticipated, as described in §5.1. First, we

require that V is invariant under the two symmetries (5.3) because this does not change the optimal bound. Second, we require that the cubic terms of $\mathbf{f} \cdot \nabla V$ cancel so that the polynomial S contains only terms of degree two and lower. The most general quadratic V satisfying these two constraints takes the form

$$V = c_1\theta_{02}+c_2\theta_{04}+c_3\psi_{11}^2+c_4\psi_{01}^2+c_5\psi_{12}^2+c_6\theta_{11}^2+c_7\theta_{02}^2+c_8\theta_{12}^2+c_9\psi_{01}\psi_{03}+c_{10}\psi_{03}^2+c_{11}\theta_{04}^2, \tag{5.12}$$

where the coefficients must satisfy:

$$\begin{gathered} c_6 = c_8, \qquad c_7 = 2c_6, \qquad 6c_4 - c_9 - \frac{4(3+k^2)}{k^2+1}c_3 + \frac{4k^2}{k^2+4}c_5 = 0, \\ c_{11} = 2c_8, \qquad \frac{2}{3}c_{10} - c_9 - \frac{4(k^2-5)}{k^2+1}c_3 + \frac{4(k^2-8)}{k^2+4}c_5 = 0. \end{gathered} \tag{5.13}$$

The previous derivation of upper bounds on N for the HK8 model in [66] was presented as an analogue of the "background method" for PDEs [13, 81]. In the PDE setting, the background method can be viewed as a special case of a more general auxiliary functional method—the PDE analogue of our general approach (4.10)—where the auxiliary functional is quadratic [6, 24]. Likewise, the argument in [66] is equivalent to a special case of our present analysis where the quadratic ansatz (5.12) for V has only two free coefficients (the "background values") rather than the six free coefficients in (5.12)–(5.13). Appendix B gives the exact constraints on these coefficients that, if added, would make our analysis equivalent to [66]. We do not impose these unnecessary constraints here, leading to better bounds on N in some parameter regimes.

With the quadratic V ansatz (5.12) and coefficients constrained by (5.13), the expres-

sion (4.9) for the polynomial S that must be SOS becomes

$$\begin{aligned}
S = \; & U - 1 + \left(4c_1 - \frac{2}{\mathcal{R}}\right)\theta_{02} + 16c_8\theta_{02}^2 + \left(16c_2 - \frac{4}{\mathcal{R}}\right)\theta_{04} + 64c_8\theta_{04}^2 \\
& + 2\sigma(k^2+1)c_3\psi_{11}^2 + k\left(-\frac{1}{2}c_1 - 2\mathcal{R}c_8 - 2\frac{\sigma}{k^2+1}c_3\right)\psi_{11}\theta_{11} + 2(k^2+1)c_8\theta_{11}^2 \\
& + 2\sigma(k^2+4)c_5\psi_{12}^2 + k\left(c_2 + 2\mathcal{R}c_8 + 2\frac{\sigma}{k^2+4}c_5\right)\psi_{12}\theta_{12} + 2(k^2+4)c_8\theta_{12}^2 \\
& + \frac{\sigma}{3}\left(c_9 + 4\frac{3+k^2}{k^2+1}c_3 - 4\frac{k^2}{k^2+4}c_5\right)\psi_{01}^2 + 10\sigma c_9\psi_{01}\psi_{03} \\
& + 27\sigma\left(c_9 + 4\frac{k^2-5}{k^2+1}c_3 - 4\frac{k^2-8}{k^2+4}c_5\right)\psi_{03}^2.
\end{aligned} \tag{5.14}$$

For each $(\mathcal{R}, k, \sigma)$, the SOS optimization (4.11) asks for the smallest U such that the c_i can be chosen to make the above expression an SOS polynomial of the state variables. The corresponding value of U provides an upper bound on the time-averaged Nusselt number N over all solutions to the HK8 model. To proceed analytically, we consider the SDP (4.13) that is equivalent to the SOS optimization (4.11). In this formulation, the SOS constraint on expression (5.14) for S is replaced by the equivalent constraint that $S = \mathbf{b}^\mathsf{T}\mathcal{Q}\mathbf{b}$ for some positive semidefinite Gram matrix $\mathcal{Q}$ and vector $\mathbf{b}$ of polynomial basis functions.

We first choose a vector $\mathbf{b}$ such that $S = \mathbf{b}^\mathsf{T}\mathcal{Q}\mathbf{b}$ holds for at least one matrix $\mathcal{Q}$, then we determine when $\mathcal{Q}$ can be positive semidefinite. A Gram matrix representation of S exists if and only if S lies in the span of the scalar polynomial entries of the matrix $\mathbf{b}\mathbf{b}^\mathsf{T}$. Any such $\mathbf{b}$ suffices; the existence of a positive semidefinite $\mathcal{Q}$ does not depend on the choice of $\mathbf{b}$. Here we simply choose the entries of $\mathbf{b}$ to be monomials:

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \mathbf{b}_3 \\ \mathbf{b}_4 \end{bmatrix}, \quad \text{where} \quad \mathbf{b}_1 = \begin{bmatrix} \psi_{11} \\ \theta_{11} \end{bmatrix}, \quad \mathbf{b}_2 = \begin{bmatrix} \psi_{01} \\ \psi_{03} \end{bmatrix}, \quad \mathbf{b}_3 = \begin{bmatrix} \psi_{12} \\ \theta_{12} \end{bmatrix}, \quad \mathbf{b}_4 = \begin{bmatrix} 1 \\ \theta_{02} \\ \theta_{04} \end{bmatrix}. \tag{5.15}$$

We have grouped the entries of $\mathbf{b}$ into the four sub-vectors $\mathbf{b}_i$ to exploit symmetry. In particular, because the expression (5.14) for S is invariant under both transformations in (5.3), we grouped monomials such that $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$, and $\mathbf{b}_4$ are invariant under, respectively, the first transformation only, the second transformation only, neither, and both. We then

restrict $\mathcal{Q}$ to be block diagonal with blocks $\mathcal{Q}_i$ sized according to the $\mathbf{b}_i$. In this case the relation $S = \mathbf{b}^\mathsf{T}\mathcal{Q}\mathbf{b}$ becomes

$$S = \sum_{i=1}^{4} \mathbf{b}_i^\mathsf{T}\mathcal{Q}_i\mathbf{b}_i, \tag{5.16}$$

and this implies that the matrices $\mathcal{Q}_i$ are written as

$$\begin{aligned}
\mathcal{Q}_1 &= \begin{bmatrix} 2\sigma(k^2+1)c_3 & -\frac{k}{2}\left(\frac{1}{2}c_1 + 2\mathcal{R}c_8 + 2\frac{\sigma}{k^2+1}c_3\right) \\ -\frac{k}{2}\left(\frac{1}{2}c_1 + 2\mathcal{R}c_8 + 2\frac{\sigma}{k^2+1}c_3\right) & 2(k^2+1)c_8 \end{bmatrix}, \\
\mathcal{Q}_2 &= \begin{bmatrix} \frac{\sigma}{3}\left(c_9 + 4\frac{3+k^2}{k^2+1}c_3 - 4\frac{k^2}{k^2+4}c_5\right) & 5\sigma c_9 \\ 5\sigma c_9 & 27\sigma\left(c_9 + 4\frac{k^2-5}{1+k^2}c_3 - 4\frac{k^2-8}{k^2+4}c_5\right) \end{bmatrix}, \\
\mathcal{Q}_3 &= \begin{bmatrix} 2\sigma(4+k^2)c_5 & \frac{k}{2}\left(c_2 + 2\mathcal{R}c_8 + 2\frac{\sigma}{k^2+4}c_5\right) \\ \frac{k}{2}\left(c_2 + 2\mathcal{R}c_8 + 2\frac{\sigma}{k^2+4}c_5\right) & 2(k^2+4)c_8 \end{bmatrix}, \\
\mathcal{Q}_4 &= \begin{bmatrix} U-1 & 2c_1 - 1/\mathcal{R} & 8c_2 - 2/\mathcal{R} \\ 2c_1 - 1/\mathcal{R} & 16c_8 & 0 \\ 8c_2 - 2/\mathcal{R} & 0 & 64c_8 \end{bmatrix}.
\end{aligned} \tag{5.17}$$

There is no loss of generality in letting all entries of $\mathcal{Q}$ outside the $\mathcal{Q}_i$ blocks be zero because if there exists any $\mathcal{Q} \succeq 0$ satisfying $S = \mathbf{b}^\mathsf{T}\mathcal{Q}\mathbf{b}$, then there exists such a $\mathcal{Q}$ that is block diagonal [18]. This simplifies matters because the condition $\mathcal{Q} \succeq 0$ is equivalent to $\mathcal{Q}_i \succeq 0$ holding for each block. In other words, S is an SOS polynomial if and only if each $\mathbf{b}_i^\mathsf{T}\mathcal{Q}_i\mathbf{b}_i$ is an SOS polynomial. To prove an upper bound $N \leq U$ in the following analysis, it suffices to find coefficients c_i such that $\mathcal{Q}_i \succeq 0$ for all four matrices in (5.16)–(5.17). A similar analytical procedure was implemented in [23] to exploit symmetry when bounding time averages in the Lorenz equations.

5.2.2 Analytical bounds near the onset of convection

The origin ceases to be globally attracting when the L_1 equilibria emerge as $\mathcal{R}$ increases past $\mathcal{R}_{L_1}$, corresponding to the onset of convection in the PDE model. At each k there exists an interval of Rayleigh number where the L_1 states maximize N. In this subsection we prove that

$$N \leq N_{L_1} = 1 + 2\left(1 - \frac{\mathcal{R}_{L_1}}{\mathcal{R}}\right) \tag{5.18}$$

when $\mathcal{R}_{L_1} \leq \mathcal{R} \leq \mathcal{R}_T$, where $\mathcal{R}_T(k)$ is defined as in (5.10). The regime of the k–$\mathcal{R}$ plane where this bound is proven is represented by horizontal hatching in Figure 5.5. The bound (5.18) may hold at some $\mathcal{R}$ values larger than $\mathcal{R}_T$ for certain k^2 and σ, but this cannot be proved using V that are quadratic.

To prove (5.18), we let $U = N_{L_1}$ in the expression (5.14) for S. When this bound holds it is saturated by the L_1 equilibria. The auxiliary function method can give a sharp bound on time averages only if S vanishes pointwise on all trajectories that saturate the bound [75], so in the present case S must vanish on the L_1 equilibria, whose nonzero coordinates are given by (3.2). This is possible only if all four terms in the sum $S = \mathbf{b}_i^{\mathsf{T}} \mathcal{Q}_i \mathbf{b}_i$ are SOS polynomials that vanish at the L_1 equilibria. The second and third terms vanish there for any $\mathcal{Q}_i$ because $\mathbf{b}_2$ and $\mathbf{b}_3$ vanish. The first and fourth terms, on the other hand, vanish at the L_1 equilibria if and only if they take the form

$$\mathbf{b}_1^{\mathsf{T}} \mathcal{Q}_1 \mathbf{b}_1 = q_1 \left[\psi_{11} - \tfrac{k}{(k^2+1)^2} \theta_{11} \right]^2, \quad \mathbf{b}_4^{\mathsf{T}} \mathcal{Q}_4 \mathbf{b}_4 = q_4 \left[\theta_{02} - (\mathcal{R} - \mathcal{R}_{L_1}) \right]^2 + q_5 \theta_{04}^2, \quad (5.19)$$

where the SOS constraints require $q_1, q_4, q_5 \geq 0$. Applying the above identities on the right-hand side of $S = \mathbf{b}_i^{\mathsf{T}} \mathcal{Q}_i \mathbf{b}_i$ and equating coefficients on each side of this equality determines four of the coefficients of V:

$$c_1 = -\frac{1}{2\mathcal{R}}, \quad c_2 = \frac{1}{4\mathcal{R}}, \quad c_3 = \frac{(k^2+1)^4}{8\sigma k^2 \mathcal{R}(\mathcal{R} - \mathcal{R}_{L_1})}, \quad c_8 = \frac{1}{8\mathcal{R}(\mathcal{R} - \mathcal{R}_{L_1})}. \quad (5.20)$$

To establish the bound (5.18), the eleven coefficients of the V ansatz (5.12) must satisfy not only the four expressions above but also the five constraints in (5.13). This ensures $\mathcal{Q}_1, \mathcal{Q}_4 \succeq 0$ when $\mathcal{R} \geq \mathcal{R}_{L_1}$, so it remains only to choose coefficients c_5 and c_9 such that $\mathcal{Q}_2, \mathcal{Q}_3 \succeq 0$. These matrix inequalities can be satisfied if and only if $\mathcal{R}_{L_1} \leq \mathcal{R} \leq \mathcal{R}_T$. To see this, we observe that a 2-by-2 matrix is positive semidefinite if and only if its upper left entry and determinant are both nonnegative. Applying this criterion gives four inequalities that are equivalent to $\mathcal{Q}_2$ and $\mathcal{Q}_3$ being positive semidefinite. We performed quantifier elimination using the `Reduce` and `Exists` commands in Mathematica to determine that these inequalities can be satisfied if and only if $\mathcal{R}_{L_1} \leq \mathcal{R} \leq \mathcal{R}_T$. Thus, quadratic V yield the sharp bound $N^* = N_{L_1}$ on this parameter regime.

5.2.3 Analytical bounds at larger Rayleigh number

In the regime where $\mathcal{R} \geq \mathcal{R}_T$, we have proved a new analytical bound when $k \leq k_T \approx 1.00319$ but not when $k > k_T$. The sum-of-squares bound in the former case is

$$N \leq 3 - \frac{2\mathcal{R}_{L_1}}{\mathcal{R}} + \frac{1}{\mathcal{R}}\left[\mathcal{R}_{L_1} - \mathcal{R}_{L_2} + \sqrt{2}\sqrt{(\mathcal{R}-\mathcal{R}_{L_1})^2 + (\mathcal{R}-\mathcal{R}_{L_2})^2}\right]. \tag{5.21}$$

To derive this bound we consider the expression (5.14) for S where U is equal to the right-hand side of (5.21). As in the previous subsection, we must show that the resulting S can be written in the form $S = \sum_{i=1}^{4} \mathbf{b}_i^{\mathsf{T}} \mathcal{Q}_i \mathbf{b}_i$, where each term in the sum is an SOS polynomial—or, equivalently, where each $\mathcal{Q}_i \succeq 0$.

Allowing for fully general SOS constraints leads to analytical difficulties, even with the simplifying block diagonal structure of $\mathcal{Q}$. Unlike in §5.2.2, we cannot anticipate where the polynomial S must vanish. Instead we simplify the analysis by making assumptions on the forms of the SOS representations. In particular we observe that, in the $\sigma \to \infty$ limit, the Lorenz triplets $\{\psi_{11}, \theta_{11}, \theta_{02}\}$ and $\{\psi_{12}, \theta_{12}, \theta_{04}\}$ decouple, and the maximal N is obtained when each triplet is in the L_1 and L_2 states, respectively (see Appendix A.1 for details). This motivates us to assume that the SOS representations of the first and third SOS polynomials take the form

$$\mathbf{b}_1^{\mathsf{T}} \mathcal{Q}_1 \mathbf{b}_1 = q_1 \left[\psi_{11} - \frac{k}{(k^2+1)^2}\theta_{11}\right]^2, \quad \mathbf{b}_3^{\mathsf{T}} \mathcal{Q}_3 \mathbf{b}_3 = q_3 \left[\psi_{12} + \frac{k}{(k^2+4)^2}\theta_{12}\right]^2. \tag{5.22}$$

The above constraints are stronger than the general SOS conditions and could potentially lead to suboptimal bounds, but this appears to not occur. At various fixed parameter values in this regime, upper bounds computed numerically with the optimal choice of quadratic V agree precisely with the analytical bound (5.21).

With the assumption (5.22) on SOS representations, the coefficients of V must satisfy

$$\begin{aligned} c_3 &= \frac{(k^2+1)}{\sigma}\mathcal{R}_{L_1} c_8, & c_5 &= \frac{(k^2+4)}{\sigma}\mathcal{R}_{L_2} c_8, \\ c_1 &= 4(\mathcal{R}_{L_1} - \mathcal{R})c_8, & c_2 &= 2(\mathcal{R}_{L_2} - \mathcal{R})c_8. \end{aligned} \tag{5.23}$$

As a result of (5.22), the matrices $\mathcal{Q}_1$ and $\mathcal{Q}_3$ defined in (5.17) each have determinant zero, and so $\mathcal{Q}_1, \mathcal{Q}_3 \succeq 0$ as long as $c_8 \geq 0$. The condition $\mathcal{Q}_2 \succeq 0$ requires $c_8 > 0$ because $c_8 = 0$ would imply that $\det \mathcal{Q} = -16\sigma^2 c_9^2$, which is negative for all c_9. Furthermore, it can be

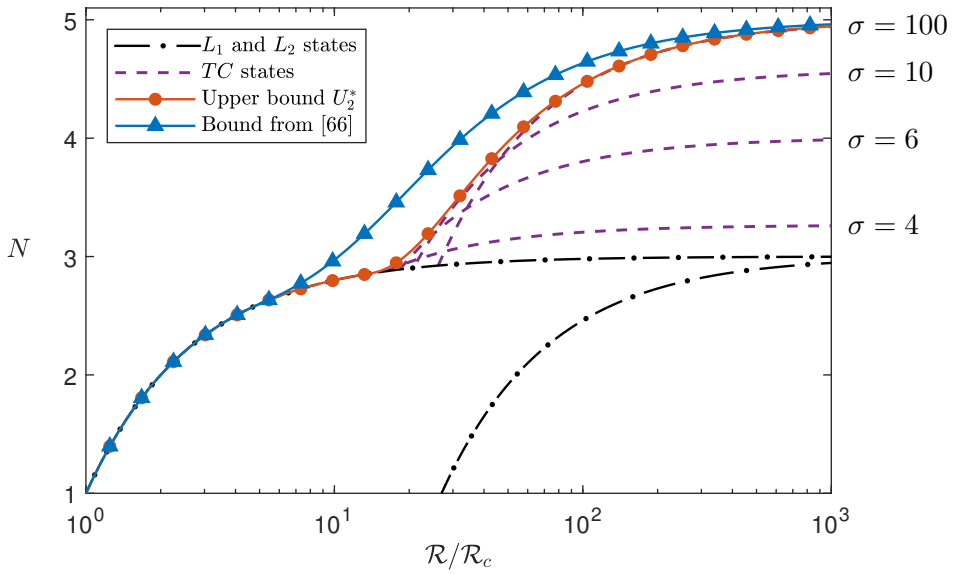


Figure 5.6: Analytical upper bounds on N^* in the $k^2 = 1/2$ case, proved with optimal quadratic auxiliary functions (U_2^*) and with the suboptimal choice of [66]. The bounds are uniform in σ. Values of N_{TC} are shown at several selected values of σ to demonstrate near-convergence of U_2^* to the envelope of steady state Nusselt numbers over σ.

shown that if c_8 is positive, $\mathcal{Q}_4 \succeq 0$ whenever $\det \mathcal{Q}_4 \geq 0$. We also observe that $\det \mathcal{Q}_4 = 0$ must hold at the minimal U; otherwise, there would exist smaller U such that $\mathcal{Q}_4 \succeq 0$. After the relations (5.23) are applied to $\mathcal{Q}_4$, the $\det \mathcal{Q}_4 = 0$ condition becomes

$$U - 1 = \frac{(8(\mathcal{R}_{L_1} - \mathcal{R})c_8 - 1/\mathcal{R})^2 + (8(\mathcal{R}_{L_2} - \mathcal{R})c_8 - 1/\mathcal{R})^2}{16c_8}. \tag{5.24}$$

Minimizing U over positive c_8 yields the bound (5.21), where the minimizer is

$$c_8 = \left[32\mathcal{R}^2\left((\mathcal{R} - \mathcal{R}_{L_1})^2 + (\mathcal{R} - \mathcal{R}_{L_2})^2\right)\right]^{-1/2}. \tag{5.25}$$

It remains to find c_9 such that $\mathcal{Q}_2 \succeq 0$. Again performing quantifier elimination with Mathematica's `Reduce` and `Exists` commands, we find that such c_9 exist if and only if $k \leq k_T$. This condition and the $\mathcal{R} \geq \mathcal{R}_T$ condition define the regime in the $\mathcal{R}$–k plane where we proved the bound (5.21). For the standard wavenumber $k^2 = 1/2$, Figure 5.6 compares the optimal analytical bounds—(5.18) and (5.21)—that can be proved using quadratic V, to the upper bound from [66], as well as to the N values of various steady states.

5.2.4 Quadratic bounds compared to steady states at maximal Prandtl number

The analytical bounds (5.10) proved using quadratic V are uniform in σ, so they are also upper bounds on the maximum of N over σ. That is,

$$N^*_{\sigma^*}(k,\mathcal{R}) := \max_{\sigma>0} N^*(k,\sigma,\mathcal{R}) \le U^*_2(k,\mathcal{R}). \tag{5.26}$$

When $\mathcal{R}_{L_1} \le \mathcal{R} \le \mathcal{R}_T(k)$, the quadratic-$V$ upper bound is saturated by the σ-independent L_1 states, and $N^*_{\sigma^*} = U^*_2$. For $\mathcal{R} > \mathcal{R}_T$, the uniform-in-$\sigma$ bounds cannot always be sharp at particular σ because there are cases where N is maximized by σ-dependent TC states. However, this does not rule out the possibility that the uniform-in-σ bounds may be sharp upper bounds on $N^*_{\sigma^*}$. Investigating this possibility, we find that the analytical bounds (5.10) are nearly equal to $N^*_{\sigma^*}$ but slightly larger in general when $\mathcal{R} > \mathcal{R}_T$.

The relationship between the quadratic-V upper bounds and the quantity $N^*_{\sigma^*}$ may be visualized by constructing an envelope of N_{TC} curves at multiple values of σ. Figure 5.6 shows a few such curves in the $k^2 = 1/2$ case. The SOS bound follows the contour of this envelope, but a small separation occurs after the SOS bound diverges from N_{L_1}. As shown in Figure 5.7 for various fixed k, the bounds provided by quadratic auxiliary functions are almost but not quite saturated by $N^*_{\sigma^*}$ when $\mathcal{R} > \mathcal{R}_T$. The $N^*_{\sigma^*}$ values used in Figure 5.7 we obtained by finding exact expressions for N_{TC} with computer algebra, then maximizing the result over σ for various fixed values of $\mathcal{R}$ and k. In each case, the maximizing σ^* lies in region V of Figure 3.2—the parameter regime in the k^2–σ plane where the branch of TC equilibria connects only to the L_1 branch. As $\mathcal{R} \to \infty$, the quadratic-V bounds and the infinite-σ limit of N_{TC} both asymptote to $N = 5$; see Appendix A.1 for details on the infinite-σ limit.

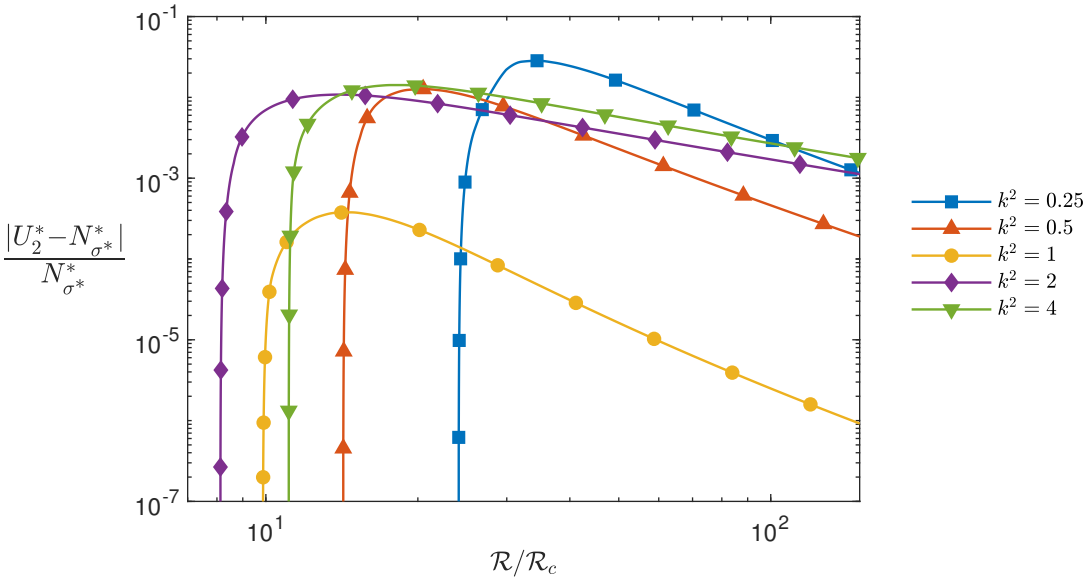


Figure 5.7: Difference between the best upper bound provable using quadratic auxiliary functions (U_2^*) and the value $N_{\sigma^*}^*$ defined by (5.26), for various fixed values of k^2. The difference vanishes as $\mathcal{R}$ decreases towards $\mathcal{R}_T(k)$ and in the limit $\mathcal{R} \to \infty$.

CHAPTER 6

Heat Transport in Other Reduced Models

In this chapter, we examine a hierarchy of reduced models that build on the HK8 system studied in Chapters 3 and 5. We apply the auxiliary function method to construct upper bounds on maximal heat transport, N^* (2.19), for several models in the hierarchy. Various solutions of each model are computed through numerical bifurcation analysis and numerical integration of the ODEs. Each particular solution provides a lower bound on maximal heat transport for the chosen truncated model and allows comparison with the upper bounds. Whenever the Rayleigh number is sufficiently small so that the reduced system can be considered a reliable approximation to the PDE, the analytical upper bounds are saturated by the branch of equilibria that arise as the first instability of the zero state.

6.1 Hierarchy of truncated models

The truncated models studied here are constructed according to the procedure outlined in Chapter 2. First, ψ and θ are written as Fourier series:

$$\psi(x,z,t) = \sum_{(m,n)\in M_\psi} \psi_{mn}(t)\,\mathrm{sc}(mkx)\sin(nz), \tag{6.1}$$

$$\theta(x,z,t) = \sum_{(m,n)\in M_\theta} \theta_{mn}(t)\,\mathrm{sc}(mkx)\sin(nz), \tag{6.2}$$

where sc represents either sine or cosine, based on the phase convention established in §2.2: we select sine for ψ modes and cosine for θ modes when $(m+n)$ is even, and make the opposite choices when $(m+n)$ is odd. Here $k = 2/A$ is the fundamental horizontal wavenumber for a domain of aspect ratio A. A reduced order model is formed by truncating the series expansions (6.1)–(6.2) to a finite number of terms, substituting the resulting expressions into the Boussinesq equations (1.13), and projecting the result onto the selected

Fourier modes. This procedure yields a system of ODEs governing the time evolution of the coefficients ψ_{mn} and θ_{mn}.

In the following discussion, let the sets of ordered pairs (m, n) of all selected modes for ψ and θ in the reduced model be given by M_ψ and M_θ, respectively. We consider models that are distinguished in the sense that each model obeys the truncated energy, temperature, and vorticity balance laws described in §2.4. All distinguished models obey the following rules:

Criterion 1 (Energy balance)**.** If $(m, n) \in M_\psi \cap M_\theta$, then $(0, 2n) \in M_\theta$ [73].

Criterion 2 (Vorticity balance)**.** If $(p, q) \in M_\psi$ and $(p, s) \in M_\psi$, then $(0, |q - s|) \in M_\psi$ if and only if $(0, q + s) \in M_\psi$.

The benefits of considering such models include boundedness of all trajectories of the ODEs and equivalence of the two definitions of the time-averaged Nusselt number, (2.17) and (2.18).

We direct our focus to the subset of truncated models including one or more "shear modes"—i.e., those of the form ψ_{0n}. The smallest model obeying each of the above criteria is the HK4 model, a modified version of the Lorenz equations that includes the Fourier modes $\psi_{11}, \theta_{02}, \theta_{11}$, and ψ_{01}. In this model, the shear mode decays exponentially along all orbits since it satisfies the simple uncoupled equation $\dot{\psi}_{01} = -\sigma\psi_{01}$, so that the dynamics are indistinguishable from those of the Lorenz equations. The next smallest model—the first to exhibit nontrivial shear flow—is the HK8 model studied in Chapters 3 and 5. We construct a hierarchy of distinguished models with shear that build on these two initial cases. First, define an ordering on the mode pairs as $(m_1, n_1) > (m_2, n_2)$ if and only if

$$m_1 + n_1 > m_2 + n_2 \text{ or } (m_1 + n_1 = m_2 + n_2 \text{ and } n_1 > n_2). \tag{6.3}$$

Let HKM_i be the i^{th} model in the hierarchy, containing M_i modes. To construct the next model:

1. Find the smallest pair with $m, n > 0$ not included in HKM_i, according to (6.3).
2. Add the corresponding modes ψ_{mn} and θ_{mn} to the system.
3. Add $\theta_{0,2n}$ and $\psi_{0,2n-1}$, if not already included in HKM_i.

The final condition ensures that all models in the hierarchy are distinguished according to the above energy and vorticity rules. In the limit as $M_i \to \infty$, the procedure enumerates all pairs (m, n) with strictly positive indices, all shear modes with odd n, and all temperature modes with $m = 0$ and even n. Therefore all Fourier modes that satisfy the phase convention are included for sufficiently large M_i. The modes in the truncated Fourier series for several models in the hierarchy are listed in Table 6.1, and a schematic of the selection procedure is shown in Figure 6.1.

Table 6.1: Additional modes required to construct each HK model from the previous one in the hierarchy up to $M_i = 44$.

Model	Additional modes	Model	Additional Modes
HK4	$\psi_{01}, \psi_{11}, \theta_{02}, \theta_{11}$	HK26	ψ_{32}, θ_{32}
HK8	$\psi_{03}, \psi_{12}, \theta_{04}, \theta_{12}$	HK28	ψ_{41}, θ_{41}
HK10	ψ_{21}, θ_{21}	HK32	$\psi_{09}, \psi_{15}, \theta_{0,10}, \theta_{15}$
HK14	$\psi_{05}, \psi_{13}, \theta_{06}, \theta_{13}$	HK34	ψ_{24}, θ_{24}
HK16	ψ_{22}, θ_{22}	HK36	ψ_{33}, θ_{33}
HK18	ψ_{31}, θ_{31}	HK38	ψ_{42}, θ_{42}
HK22	$\psi_{07}, \psi_{14}, \theta_{08}, \theta_{14}$	HK40	ψ_{51}, θ_{51}
HK24	ψ_{23}, θ_{23}	HK44	$\psi_{0,11}, \psi_{16}, \theta_{0,12}, \theta_{16}$

6.2 Particular solutions of models in the HK hierarchy

6.2.1 Equilibria of truncated models

We first turn our attention to the equilibria of models in the HK hierarchy. These solutions provide candidates for the optimal heat transport and therefore bound N^* from below. Such solutions are of interest because it is conjectured that steady solutions maximize heat transport in the Boussinesq equations [80]. For all HK models, the zero equilibrium is globally attracting for sufficiently small $\mathcal{R}$. As the Rayleigh number increases, the zero state undergoes a series of pitchfork bifurcations, giving rise to pairs of equilibria via an instability in the variables ψ_{mn} and θ_{mn}. These equilibria emerge at the Rayleigh number $\mathcal{R}_{mn}$, given by

$$\mathcal{R}_{mn} := \left((mk)^2 + n^2\right)^3 / (mk)^2. \tag{6.4}$$

Depending on the model, these equilibria may simply be three dimensional analogues of the nonzero equilibria of the Lorenz model, or may take a more complicated form where

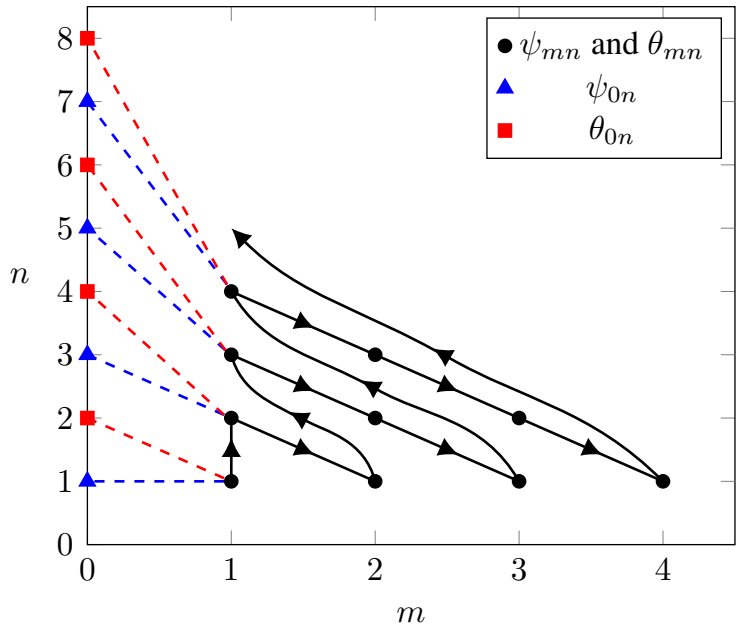


Figure 6.1: Schematic of the mode selection procedure for the first 10 models in the HK hierarchy, where arrows indicate the order of selection and dashed lines connect modes added simultaneously. Modes with $m = 0$ are added when the next $(m + n)$ shell is reached. Each model contains the modes of all previous models, beginning with the HK4 system.

additional variables must be nonzero. In the simpler case, we denote these equilibria as L_{mn}. The nonzero variables in the L_{mn} equilibria are given by

$$\psi_{mn} = \pm\tfrac{\sqrt{8}}{(mk)^2+n^2}\sqrt{\mathcal{R}-\mathcal{R}_{mn}}, \qquad \theta_{mn} = \pm(-1)^{m+n}\sqrt{8}\tfrac{(mk)^2+n^2}{mk}\sqrt{\mathcal{R}-\mathcal{R}_{mn}}, \qquad \theta_{0,2n} = \tfrac{1}{n}\left(\mathcal{R}-\mathcal{R}_{mn}\right). \tag{6.5}$$

Each of the L_{mn} equilibria can be related to the equilibria of the Lorenz equations by a linear change of variables. The value of N at the L_{mn} equilibria, computed from either (2.17) or (2.18), is

$$N_{L_{mn}} = 3 - 2\mathcal{R}_{mn}/\mathcal{R} \tag{6.6}$$

In higher-order reduced systems, the equilibria arising at $\mathcal{R}_{mn}$ often take a more complicated form. This occurs due to additional terms in the reduced system that result from nonlinear interactions between Fourier modes.

The equilibria that bifurcate at $\mathcal{R}_{11}$ are of particular interest because they correspond to a pair of steady convection rolls that are globally attracting at onset. In the simplest case, these equilibria are called L_{11} and are equivalent to the equilibria that emerge at the first

instability of the HK8 model (3.2). More generally, we call this branch of equilibria the *primary branch*. The smallest model where the primary branch deviates from the expressions for the L_{11} states occurs in the HK14 system. In this case, the primary branch has seven nonzero components: $\psi_{11}, \psi_{13}, \theta_{02}, \theta_{04}, \theta_{06}, \theta_{11}$ and θ_{13}. The equilibria no longer satisfy (6.5) because the ODE for θ_{13} includes a term proportional to $\psi_{11}\,\theta_{02}$. As a result, additional variables must be nonzero to satisfy the algebraic equation corresponding to $\dot{\theta}_{13} = 0$.

The primary branches of equilibria for several reduced order models are depicted in Figure 6.2. For each of the models depicted, the temperature profiles display slightly unphysical behavior at $\mathcal{R} = 5\mathcal{R}_c$, including internal temperature maxima. These features are not present at the onset of convection and develop at some larger $\mathcal{R}$, indicating that the reduced model is not capturing the full physics of the Boussinesq equations.

In general, the expressions for the equilibria that emerge at $\mathcal{R}_{mn}$ deviate from (6.5) for models large enough such that $\psi_{m,3n}$ and $\theta_{m,3n}$ are included. Similar to the case of HK14, nonlinear pairing between ψ_{mn} and $\theta_{0,2n}$ occurs in the $\theta_{m,3n}$ equation, so that additional variables must be nonzero. We observe that the inclusion of additional modes generally enhances heat transport from (6.6).

We studied the bifurcation structure of the HKM_i models in more detail with the numerical continuation package MATCONT [10]. For these and subsequent computations within this chapter, the model parameters k^2 and σ were fixed at $1/2$ and 10 respectively. To improve numerical stability for all computations performed on the HK models, we scale the ψ modes by $\mathcal{R}^{-1/2}$, θ modes by $\mathcal{R}^{-1}$, and time by $\mathcal{R}^{1/2}$. As a result, the Nusselt number for the scaled HK models is

$$N = 1 + \sum_{(0,2n)\in M_\theta} (2n)\,\overline{\theta_{0,2n}}. \tag{6.7}$$

With $M_i \leq 44$, we locate all branches of equilibria detectable for $\mathcal{R}/\mathcal{R}_c < 1000$, with $\mathcal{R}$ as the bifurcation parameter. We began by continuing all branches that bifurcate from the zero state at $\mathcal{R}_{mn}$. For these and each additional branch located, we continued the equilibria until $\mathcal{R}/\mathcal{R}_c > 1000$, or until the curve terminates. This process was repeated for each equilibria stemming from any of the pitchfork bifurcations detected, until no additional branches of equilibria were found. Results for a few selected models are displayed in Figure 6.3. The primary equilibria of the HK14 model display appreciably greater heat transport than those of the HK10 model at all values of $\mathcal{R}$ past onset. This is due in large part to the mechanism described above where the L_{11} equilibria are augmented with additional modes.

We maximized N among all computed equilibria for each model. In all cases, the

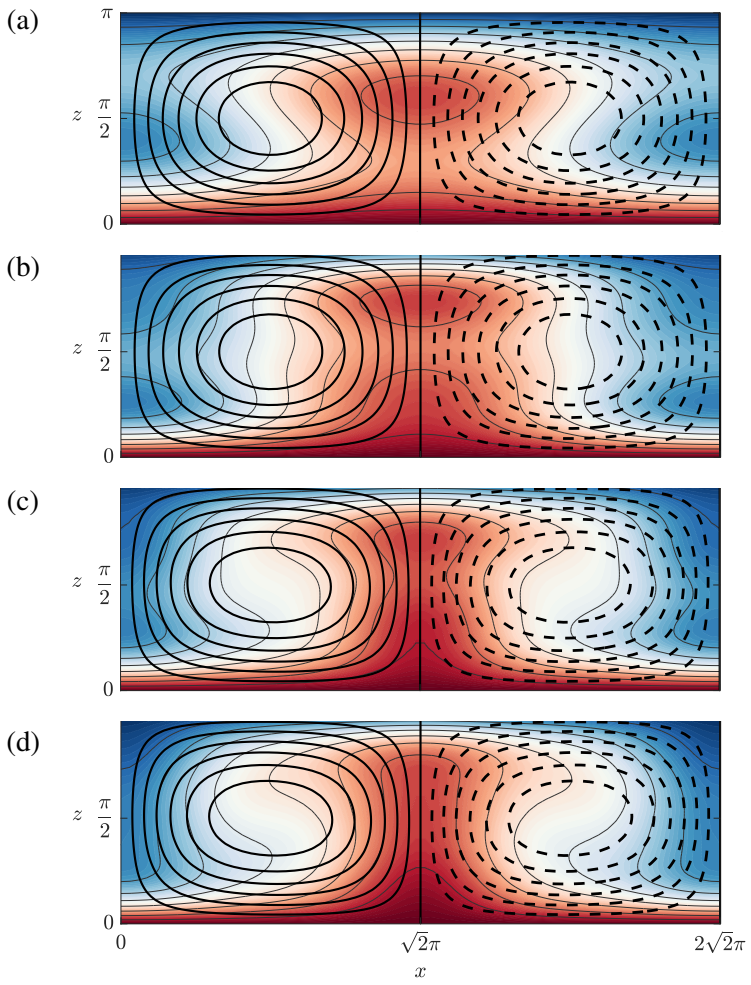


Figure 6.2: Streamlines overlaid on contours of the temperature (T) for approximations of steady convection whose mode amplitudes in the Galerkin expansion (2.4) are given by the primary branches of equilibria. The plots above depict several reduced models at $k^2 = 1/2$, $\sigma = 10$ and $\mathcal{R} = 5\mathcal{R}_c$: (a) the HK8 model (b) the HK14 model (c) the HK22 model and (d) the HK32 model. The T scale ranges from 0 (dark) to 1 (light). Positive and negative vorticity are indicated by solid and dashed streamlines, respectively. In each model, additional modes pair with the nonzero variables of the L_{11} states due to nonlinear interactions between modes, enhancing heat transport across the domain.

primary equilibria are maximal from onset until some larger value of $\mathcal{R}$. Eventually, the heat transport of the primary states may be surpassed by some other state, but it appears that this only happens when $\mathcal{R}$ is well beyond the point where unphysical behavior is first observed. The primary equilibria were compared to the analogous states of the Boussinesq equations—those arising from the first instability from the static state. We observe that the HK14 model predicts slightly larger heat transport than the PDE for some $\mathcal{R}$. This may occur as a result of the partially filled shell in the hierarchy, where only one Lorenz triple

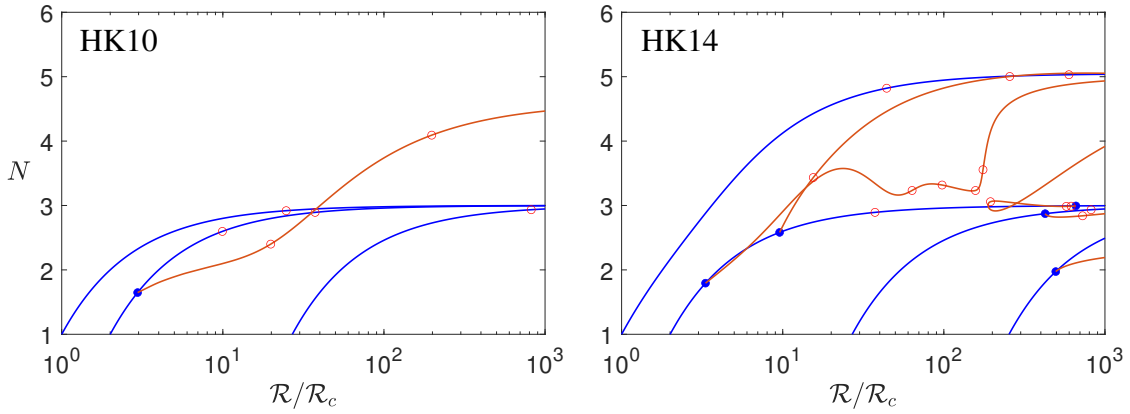


Figure 6.3: Bifurcation diagrams for the HK10 and HK14 models, at $k^2 = 1/2$ and $\sigma = 10$. Filled circles indicate pitchfork bifurcations and open circles denote Hopf bifurcations. All curves were computed by numerical continuation with a resolution of approximately 0.1 in units of $\mathcal{R}/\mathcal{R}_c$.

with $m + n = 4$ is included in the truncation.

The bifurcation structure of the HK models becomes more complex with larger M_i, and the number of equilibria and Hopf bifurcations rapidly increases with the dimension of the ODE. Therefore it is not practical to attempt to locate every equilibrium branch when the dimension becomes sufficiently large. For larger models, we consider only the primary branch of equilibria, and conjecture that these equilibria transport heat optimally at all physically relevant values of $\mathcal{R}$. The primary equilibria for several models are shown in Figure 6.4.

6.2.2 Time integration of the HK ODEs

For models in the HK hierarchy with $M_i \leq 44$, additional candidates for the maximal N were obtained by directly computing the time average integral (1.19) with $\Phi = N$. Numerical solutions were obtained by directly integrating the ODEs for 10^4 to 10^5 time units with the MATLAB solver `ode45` starting from randomly generated initial conditions within $[1, 1]^{M_i}$. The absolute and relative tolerances of the solver were set to 10^{-12} and 10^{-9}, respectively. In cases where more than one stable solution was detected at a particular Rayleigh number, we report the maximum N among such solutions. The results are displayed in Figure 6.5. When the Rayleigh number is sufficiently small, integrating the ODEs yields stable equilibria, while at larger $\mathcal{R}$ we identify trajectories that are time-periodic or even chaotic.

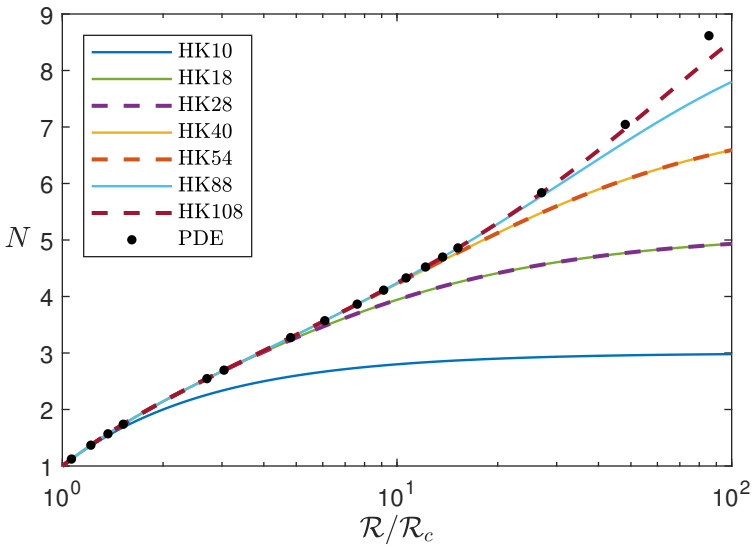


Figure 6.4: Heat transport among the primary branch of equilibria found via numerical continuation for several selected truncated models. Values of Nu at the equilibria arising from the first instability of the Boussinesq equations—analogues to the primary equilibria—are plotted for comparison. For each model with $M_i \leq 40$, the primary branches were determined to provide maximal heat transport among all equilibria until $\mathcal{R}/\mathcal{R}_c$ was greater than 20. Data for the PDE was computed by Baole Wen [80].

6.3 Upper bounds on N

We apply the auxiliary function method with sum-of-squares optimization introduced in Chapter 4 to determine upper bounds on N^* for models in the HK hierarchy. In the nomenclature introduced in Chapter 4, we seek U_d^*: the minimum upper bound that can be proved with degree d auxiliary functions. In general U_d^* depends on both the chosen model and the parameters $\mathcal{R}, \sigma$, and k.

The number of terms in the general ansatz for the auxiliary function $V \in \mathbb{P}_{n,d}$ grows rapidly in both the dimension n and degree d. As a result, if the SDP corresponding to (4.11) is solved without simplification, the computational cost becomes prohibitive in all but the simplest examples. To make these computations tractable, we simplify the structure of the V ansatz by taking advantage of the structure of the HK ODEs. Following the procedure outlined in §4.5, the monomial basis for the auxiliary function V is reduced by imposing symmetry and highest degree cancellation conditions. Table 6.2 shows the number of monomials before and after reduction for several models with degree four V. Variables are scaled as described in §6.2 so that the relevant dynamics lie within $[-1, 1]^{M_i}$.

Bounds are constructed for HK models with degree four V when $M_i \leq 40$, and with degree two V when $M_i \leq 60$. For each model in the HK hierarchy with $M_i \leq 40$, the

Table 6.2: Number of monomials in the ansatz for the auxiliary function V of degree 4 for several models in the HK hiearchy before and after reducing the ansatz using the structure of the ODEs. The number of monomials before reduction is $\binom{M_i+4}{4}$, where M_i is the dimension of the model. The time required to solve the SDP is reported for both the reduced and unreduced problems (the unreduced problem was not solved for $M_i > 18$ due to memory constraints). As a rule of thumb, the memory and time requirements scale roughly as $O(n^3)$ when the corresponding Gram matrix is of dimension n. Computation time for the HK26 model was comparatively slow because it has only one sign symmetry.

	Unreduced		Reduced	
Model	Monomials	Time (s)	Monomials	Time (s)
HK8	495	2	88	0.5
HK10	1000	5	158	0.7
HK14	3059	60	381	3
HK16	4844	230	447	7
HK18	7314	450	574	9
HK22	14949	-	977	50
HK24	20474	-	1189	100
HK26	27404	-	1433	815
HK28	35959	-	1697	250

primary branch of equilibria saturate the upper bounds when the Rayleigh number is slightly larger than $\mathcal{R}_{11}$. The auxiliary function method provides the sharpest bounds when $M_i = 14$. This model is the first one whose primary equilibria deviate from the form (6.5), with seven nonzero modes when $\mathcal{R} > \mathcal{R}_{11}$, resulting in larger heat transport than the Lorenz equilibria. As such, the primary equilibria for HK14 are maximal for a larger range of Rayleigh number than smaller HK models, but are still simple enough to admit sharp bounds with V of low degree.

When M_i is increased for fixed $\mathcal{R}$, the bound U_4^* increases noticeably upon progression to the HK14 and HK32 models. The upper bounds vary by smaller amounts within each shell, and in some cases U_4^* decreases from one model to the next at the same parameter set. This appears to be caused primarily by the enhancement of heat transport that occurs when progression to the next model results in the pairing of modes with larger wavenumber with the nonzero modes of the primary branch of equilibria. The occurs only when the total wavenumber $(m + n)$ of the next shell is even, since modes of odd total wavenumber do not pair with the L_{11} equilibria. A consequence of this pairing mechanism is that the shear

modes are identically zero at the primary equilibria. Similarly, in numerical simulations of the full PDE, zonal flow has been observed to decrease the time-averaged heat transport [25].

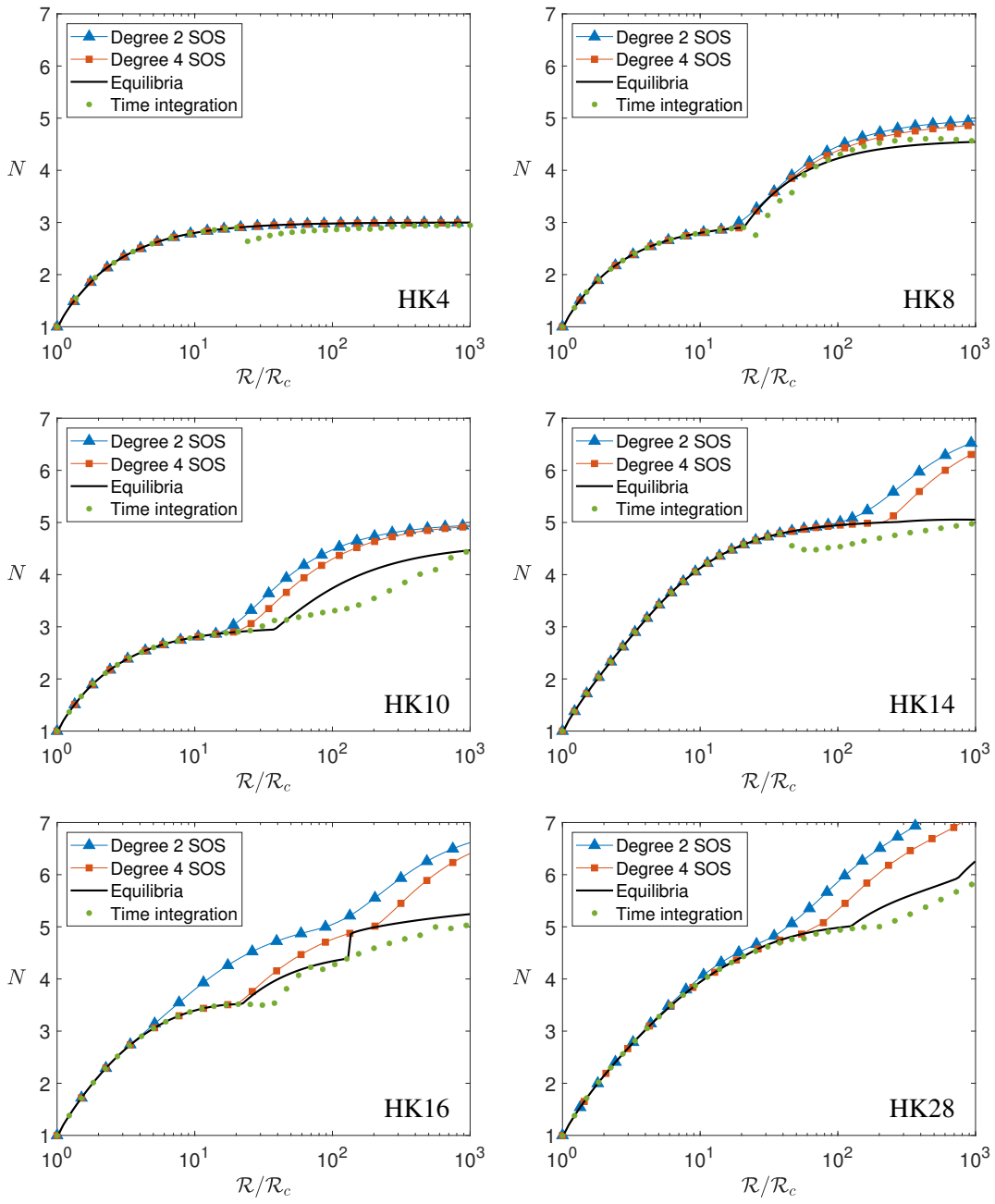


Figure 6.5: Upper bounds on N obtained with the auxiliary function method for a few selected models in the HK hierarchy. The maximum value of the Nusselt number obtained from numerical continuation and direct integration of the ODEs are shown for comparison.

SDPs with degree two V provide more conservative upper bounds, but are less computationally taxing to compute, allowing bounds to be constructed for much larger systems. Such bounds are displayed in Figure 6.6 for the HK40, HK54 and HK70 models. These are the models that complete their respective shell in the HK hierarchy. When progressing to a shell with even modes, the primary states are augmented, often resulting in a jump in heat transport. At each $\mathcal{R}$, the SDP bounds and the primary branch of equilibria provide an enclosure in N around the maximal Nusselt number, N^*.

The bound U_2^* appears to be constant in σ for each model in the hierarchy, just as observed for the HK8 model in §5.2. For each model with $M_i \leq 58$, U_2^* varied by less than 1% when $\sigma \in [0.01, 100]$ and other parameters are held fixed. As a result, the degree two upper bounds are generally not sharp among solutions with fixed σ. When $\sigma = 10$ and $k^2 = 1/2$, the degree two bounds are sharp for only a small interval of Rayleigh number in all models with $M_i > 4$.

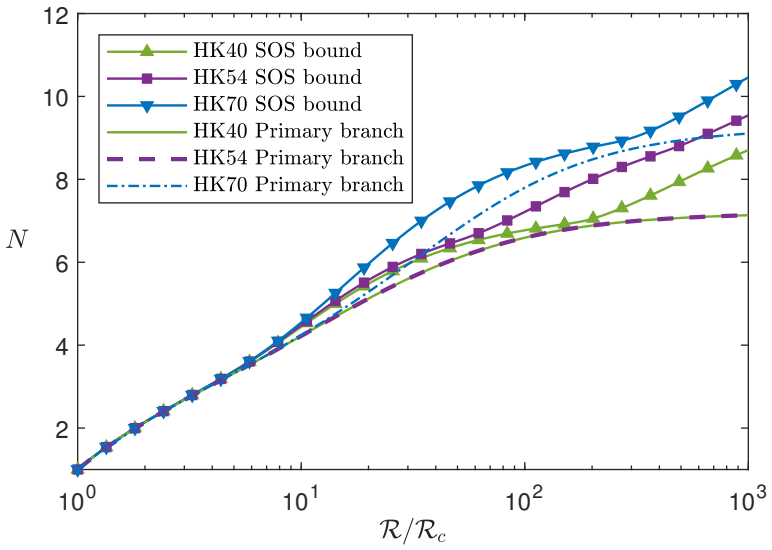


Figure 6.6: Upper bounds on N computed with degree two auxiliary functions compared to the L_{11} equilibria for the HK40, HK54 and HK70 models. Values of $N_{L_{11}}$ coincide for the equilibria of the HK40 and HK70 models over this range of $\mathcal{R}$.

CHAPTER 7

Discussion and Future Work

In this thesis, reduced order models are constructed to approximate the physics of Rayleigh–Bénard convection. Chapter 2 includes a derivation of the general form of Galerkin-truncated models subject to stress-free isothermal boundary conditions in a horizontally periodic domain. Criteria are presented for truncating the Fourier series such that the resulting models satisfy energy, temperature and vorticity balance laws derived from the Boussinesq equations. It is theorized that imposing these restrictions results in models that more faithfully represent the physics of the PDE [76]. The energy balance, in particular, implies boundedness of all solutions to the reduced system and equivalence of the two definitions of the truncated version of the time-averaged heat transport, N [73]. The other two balance laws used for model selection in this thesis—the temperature and vorticity laws presented in §2.4—are of physical relevance but their importance is not fully understood. Future studies may determine the implications of imposing these constraints on the ODE dynamics, or include additional constraints based on the physics of convection.

The auxiliary function method is presented in Chapter 4, including a general discussion of its application to reduced order convection models. A crucial step in the method is the selection of an auxiliary function, V, allowing global upper bounds on the heat transport to be established. For polynomial dynamical systems, the auxiliary functions are typically taken to be polynomials of fixed degree d. The number of monomials required to characterize the ansatz for V of fixed degree increases rapidly in both the degree and the number of variables, placing a significant limitation on the feasibility of generalizing the method to larger models. One way to deal with this issue is to reduce the number of monomials in the ansatz. In this thesis, we impose sign symmetry and highest-degree cancellation conditions that offer a drastic reduction in the size of the monomial basis without sacrificing the quality of the upper bounds. The computational complexity involved in solving the corresponding SDPs with an $n \times n$ Gram matrix scales roughly as n^3 [39], so constraints on the monomial

basis that lead to a simpler Gram matrix would have a significant impact on computation time and memory cost.

A second possible improvement can be achieved by using algorithms that scale better with the size of the problem. First order methods, including the alternating direction method of multipliers (ADMM) [83], improve computation time at the expense of accuracy by exploiting sparsity within the SDP formulation. DSOS and SDSOS optimization (diagonally dominant and scaled diagonally dominant SOS) [1] introduce alternatives to the SOS constraint that allow more efficient computation at the risk of obtaining more conservative bounds. This is because these methods consider auxiliary functions in a strictly smaller class of polynomials. Such techniques could allow computation of SDPs of larger systems, or with higher degree V.

Even with the above improvements, the truncated Fourier series may provide an inadequate description of the physics at larger Rayleigh numbers. Strongly forced convective flows develop a wide range of length scales that require a large number of Fourier modes to fully resolve. To develop bounds that are less sensitive to these length scales, sum-of-squares optimization can be applied with auxiliary functionals acting on the function space of the PDE. This approach has been implemented for other systems directly [24], and is an extension of the background method applied to produce bounds on time-averaged quantities in fluid dynamics [6].

In Chapter 5, the auxiliary function method was applied using sum-of-squares optimization to establish upper bounds on the mean heat transport, N, among all solutions of the HK8 system, a truncated version of Rayleigh's PDE model [58]. Values of N were also calculated along various particular solutions to the HK8 system. The upper bounds were sharp in many cases, as confirmed by their coincidence with N on known steady or time-periodic solutions. For purposes of numerical computation, SOS optimization was performed via semidefinite programming with auxiliary functions of degrees 2, 4, 6, and 8 at various choices of the model parameters $\mathcal{R}$, σ, and k. Moreover, upper bounds were derived analytically using quadratic auxiliary functions yielding estimates that depend explicitly on the parameters $\mathcal{R}$ and k, and improve upon a previous result.

For all values of k and σ where we computed bounds on the HK8 model by SOS optimization, steady states maximize N for a range of $\mathcal{R}$ above onset. Specifically, for all k and σ there exists an interval of $\mathcal{R}$ where the L_1 equilibria maximize N. This interval contains the σ-independent interval $\mathcal{R}_{L_1} \leq \mathcal{R} \leq \mathcal{R}_T$, as follows from our analytical bounds. When σ is sufficiently large, there exists a bounded interval of $\mathcal{R}$ within $(\mathcal{R}_T, \infty)$ where the

TC equilibria saturate the upper bound; for smaller σ the L_1 equilibria transport optimally among the steady states of the HK8 model. Numerical results suggest that time-dependent trajectories often maximize N for sufficiently large $\mathcal{R}$. The emergence of time-periodic solutions that transport more heat than any steady state contrasts with the Lorenz equations where the L_1 equilibrium maximizes N for all $\mathcal{R}$ beyond onset [23, 41, 65]. Motivated by the physics of the full PDE model, where the aspect ratio of convection cells need not be fixed, we also maximized our bounds over k. This maximization procedure is often applied in studies of maximal heat transport for the full PDE [27, 67, 79]. This naturally divides the σ–$\mathcal{R}$ parameter space into regions (Figure 5.4) where the so-maximized bounds are maximized by each of the three types of solutions described above. For a range of $\mathcal{R}$ after the onset of convection, bounds are always saturated by steady states. This means that maximal heat transport is achieved by steady flows, at least for the small range of $\mathcal{R}$ where the HK8 model faithfully reflects the full PDE.

For models in the HK hierarchy studied in Chapter 6, we observe that N is most often maximized by steady states, especially when the Rayleigh number is sufficiently small to allow quantitative comparison to the PDE. The primary branch equilibria—the steady states that emerge as the first instability of the static state when $\mathcal{R}$ is raised—maximize heat transport for a range of $\mathcal{R}$ at the onset of convection. The heat transport of the primary states is enhanced when additional modes are included in the reduced order model that pair with the nonzero variables of the primary branch. The first such enhancement occurs in the HK14 model, and additional jumps in the total heat transport typically occur when the HK hierarchy begins a "shell" in the hierarchy that has even total wavenumber $(m + n)$. Other distinguished modal hierarchies could potentially be constructed that take advantage of this phenomenon to achieve greater heat transport with a similar number of modes.

We also observe that the states providing optimal heat transport do not include shear modes except when the Rayleigh number is well beyond the point where the reduced models closely approximate the heat transport of the full Boussinesq equations. In fact, we observe that the shear modes are identically zero along all equilibria that bifurcate from the zero state (the L_{mn} equilibria and higher-dimensional analogues). Steady states exhibiting shear flow—analogues of the TC equilibria studied in the HK8 model—were only observed to emerge as bifurcations from equilibria other than the zero state. The study of reduced order models without shear is another possible direction for future research. Such models arise naturally if one considers a fully enclosed fluid domain, rather than imposing horizontal periodicity. Another possible direction of future research is to consider other types of

boundary conditions, such as no-slip or fixed-flux conditions along the walls.

BIBLIOGRAPHY

[1] A. A. Ahmadi and A. Majumdar. DSOS and SDSOS optimization: More tractable alternatives to sum of squares and semidefinite optimization. *SIAM Journal on Applied Algebra and Geometry*, 3(2):193–230, 2019.

[2] D. A. Ayala. MATLAB source code, 2017.

[3] J. T. Beale, T. Kato, and A. Majda. Remarks on the breakdown of smooth solutions for the 3-D Euler equations. *Communications in Mathematical Physics*, 94(1):61–66, 1984.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] F. H. Busse. Convection driven zonal flows and vortices in the major planets. *Chaos*, 4(2):123–134, 1994.

[6] S. Chernyshenko. Relationship between the methods of bounding time averages. *arXiv:1704.02475v2*, 2017.

[7] S. I. Chernyshenko, P. Goulart, D. Huang, and A. Papachristodoulou. Polynomial sum of squares in fluid dynamics: A review with a look ahead. *Philosophical Transactions of the Royal Society A*, 372:20130350, 2014.

[8] M. D. Choi, T. Y. Lam, and B. Reznick. Sums of squares of real polynomials. In *Proceedings of Symposia in Pure Mathematics*, volume 58, pages 103–126, 1994.

[9] J. Deng, T. Y. Hou, and X. Yu. Geometric properties and nonblowup of 3D incompressible Euler flow. *Communications in Partial Differential Equations*, 30(1-3):225–243, 2005.

[10] A. Dhooge, W. Govaerts, and Y. A. Kuznetsov. MATCONT: A MATLAB package for numerical bifurcation analysis of ODEs. *ACM Transactions on Mathematical Software*, 29(2):141–164, 2003.

[11] P. H. Diamond, S.-I. Itoh, K. Itoh, and T. S. Hahm. Zonal flows in plasma—a review. *Plasma Physics and Controlled Fusion*, 47(5):R35–R161, 2005.

[12] C. R. Doering. Turning up the heat in turbulent thermal convection. *Proceedings of the National Academy of Sciences of the United States of America*, 117(18):9671–9673, 2020.

[13] C. R. Doering and P. Constantin. Variational bounds on energy dissipation in incompressible flows: Shear flow. *Physical Review E*, 53(6):5957–5981, 1996.

[14] C. R. Doering and J. D. Gibbon. *Applied Analysis of the Navier–Stokes Equations*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 1995.

[15] C. R. Doering, J. D. Gibbon, and D. D. Holm. A new diagnostic for the relative accuracy of Euler codes. *arXiv:1002.2961*, 2010.

[16] M. Duponcheel, P. Orlandi, and G. Winckelmans. Time-reversibility of the Euler equations as a benchmark for energy conserving schemes. *Journal of Computational Physics*, 227(19):8736–8752, 2008.

[17] G. Fantuzzi, D. Goluskin, D. Huang, and S. I. Chernyshenko. Bounds for deterministic and stochastic dynamical systems using sum-of-squares optimization. *SIAM Journal on Applied Dynamical Systems*, 15(4):1962–1988, 2016.

[18] K. Gatermann and P. A. Parrilo. Symmetry groups, semidefinite programs, and sums of squares. *Journal of Pure and Applied Algebra*, 192(1-3):95–128, 2004.

[19] J. D. Gibbon. The three-dimensional Euler equations: where do we stand? *Physica D*, 237(14-17):1894–1904, 2008.

[20] J. D. Gibbon and D. D. Holm. The dynamics of the gradient of potential vorticity. *Journal of Physics A: Mathematical and Theoretical*, 43(17):172001, 2010.

[21] A. Gluhovsky, C. Tong, and E. Agee. Selection of modes in convective low-order models. *Journal of the Atmospheric Sciences*, 59:1383–1393, 2002.

[22] D. Goluskin. *Zonal flow driven by convection and convection driven by internal heating*. PhD thesis, Columbia University, 2013.

[23] D. Goluskin. Bounding averages rigorously using semidefinite programming: Mean moments of the Lorenz system. *Journal of Nonlinear Science*, 28(2):621–651, 2018.

[24] D. Goluskin and G. Fantuzzi. Bounds on mean energy in the Kuramoto–Sivashinsky equation computed using semidefinite programming. *Nonlinearity*, 32(5):1705–1730, 2019.

[25] D. Goluskin, H. Johnston, G. R. Flierl, and E. A. Spiegel. Convectively driven shear and decreased heat flux. *Journal of Fluid Mechanics*, 759:360–385, 2014.

[26] R. Grauer and T. C. Sideris. Numerical computation of 3D incompressible ideal fluids with swirl. *Physical Review Letters*, 67:3511–3514, 1991.

[27] P. Hassanzadeh, G. P. Chini, and C. R. Doering. Wall to wall optimal transport. *Journal of Fluid Mechanics*, 751:627–662, 2014.

[28] D. Henrion and M. Korda. Convex computation of the region of attraction of polynomial control systems. *IEEE Transactions on Automatic Control*, 59(2):297–312, 2014.

[29] K. B. Hermiz, P. N. Guzdar, and J. M. Finn. Improved low-order model for shear flow driven by Rayleigh–Bénard convection. *Physical Review E*, 51(1):325–331, 1995.

[30] B. J. Hoskins. On the use and significance of isentropic potential vorticity maps. *Quarterly Journal of the Royal Meteorological Society*, 111(470):877–946, 1985.

[31] T. Y. Hou and R. Li. Dynamic depletion of vortex stretching and non-blowup of the 3-D incompressible Euler equations. *Journal of Nonlinear Science*, 16:639–664, 2006.

[32] L. N. Howard and R. Krishnamurti. Large-scale flow in turbulent convection: A mathematical model. *Journal of Fluid Mechanics*, 170:385–410, 1986.

[33] K. P. Iyer, J. D. Scheel, J. Schumacher, and K. R. Sreenivasan. Classical 1/3 scaling of convection holds up to Ra = 1015. *Proceedings of the National Academy of Sciences of the United States of America*, 117(14):7594–7598, 2020.

[34] H. Johnston and C. R. Doering. A comparison of turbulent thermal convection between conditions of constant temperature and constant flux. *Physical Review Letters*, 102(6), 2009.

[35] R. Krishnamurti and L. N. Howard. Large-scale flow generation in turbulent convection. *Proceedings of the National Academy of Sciences*, 78(4):1981–1985, 1981.

[36] M. V. Lakshmi, G. Fantuzzi, J. D. Fernández-Caballero, Y. Hwang, and S. I. Chernyshenko. Finding extremal periodic orbits with polynomial optimization, with application to a nine-mode model of shear flow. *SIAM Journal on Applied Dynamical Systems*, 19(2):763–787, 2020.

[37] J. B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11:796–817, 2001.

[38] J. Löfberg. YALMIP: A toolbox for modeling and optimization in MATLAB. In *IEEE International Conference on Robotics and Automation*, pages 284–289, Taipei, Taiwan, 2004.

[39] J. Löfberg. Pre- and post-processing sum-of-squares programs in practice. *IEEE Transactions on Automatic Control*, 2009.

[40] M. D. Long and T. W. Becker. Mantle dynamics and seismic anisotropy. *Earth and Planetary Science Letters*, 297(3-4):341–354, 2010.

[41] E. N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(3):130–141, 1963.

[42] G. Luo and T. Y. Hou. Potentially singular solutions of the 3D axisymmetric Euler equations. *Proceedings of the National Academy of Sciences of the United States of America*, 111(36):12968–12973, 2014.

[43] W. V. R. Malkus. The heat transport and spectrum of thermal turbulence. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 225(1161):196–212, 1954.

[44] W. V. R. Malkus. Non-periodic convection at high and low Prandtl number. *Mémoires la Société R. des Sci. Liège, Collect. IV*, 6:125–128, 1972.

[45] W. V. R. Malkus and G. Veronis. Finite amplitude cellular convection. *Journal Fluid Mechanics*, 59(3):225–260, 1958.

[46] S. Mishra and L. V. Spinolo. Accurate numerical schemes for approximating initial-boundary value problems for systems of conservation laws. *Journal of Hyperbolic Differential Equations*, 12(1):61–86, 2015.

[47] MOSEK ApS. *MOSEK optimization toolbox for MATLAB*, 2018. Release 9.0.98.

[48] K. G. Murty and S. N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39(2):117–129, 1987.

[49] Y. Nesterov. Squared functional systems and optimization problems. In H. Frenk, K. Roos, T. Terlaky, and S. Zhang, editors, *High performance optimization*, pages 405–440. Springer, 2000.

[50] J. J. Niemela, L. Skrbek, K. R. Sreenivasan, and R. J. Donnelly. Turbulent convection at very high Rayleigh numbers. *Nature*, 406(6794):439–439, 2000.

[51] J. J. Niemela and K. R. Sreenivasan. Confined turbulent convection. *Journal of Fluid Mechanics*, 481(481):355–384, 2003.

[52] M. L. Olson, D. Goluskin, W. W. Schultz, and C. R. Doering. Heat transport bounds for a truncated model of Rayleigh–Bénard convection via polynomial optimization. *arXiv:2004.07204v1*, 2020.

[53] P. A. Parrilo. *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*. PhD thesis, California Institute of Technology, 2000.

[54] P. A. Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming*, 96(2):293–320, 2003.

[55] V. Powers and T. Wörmann. An algorithm for sums of squares of real polynomials. *Journal of Pure and Applied Algebra*, 127:99–104, 1998.

[56] C. H. B. Priestley. Convection from a large horizontal surface. *Australian Journal of Physics*, 7(1):176, 1954.

[57] A. Pumir and E. D. Siggia. Development of singular solutions to the axisymmetric Euler equations. *Physics of Fluids A. Fluid Dynamics*, 4(7):1472–1491, 1992.

[58] L. Rayleigh. LIX. On convection currents in a horizontal layer of fluid, when the higher temperature is on the under side. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 32(192):529–546, 1916.

[59] B. Reznick. Some concrete aspects of Hilbert's 17th Problem. In *Real algebraic geometry and ordered structures (Baton Rouge, LA, 1996)*, volume 253, pages 251–272. American Mathematical Society, Providence, RI, 2000.

[60] B. Saltzman. Finite amplitude free convection as an initial value problem—I. *Journal of the Atmospheric Sciences*, 19(4):329–341, 1962.

[61] W. H. Schubert, E. Ruprecht, R. F. A. Hertenstein, R. N. Ferreira, R. K. Taft, C. M. Rozoff, P. A. E. Ciesielski, and H. chi Kuo. English translations of twenty-one of Ertel's papers on geophysical fluid dynamics. *Meteorologische Zeitschrift*, 13:527–576, 2004.

[62] N. Z. Shor. Class of global minimum bounds of polynomial functions. *Cybernetics*, 23(6):731–734, 1988.

[63] N. Z. Shor. *Nondifferentiable Optimization and Polynomial Problems*, volume 24 of *Nonconvex Optimization and Its Applications*. Springer US, Boston, MA, 1998.

[64] E. Siggia. High Rayleigh number convection. *Annual Review of Fluid Mechanics*, 26(1):137–168, 1994.

[65] A. N. Souza and C. R. Doering. Maximal transport in the Lorenz equations. *Physics Letters A*, 379:518–523, 2015.

[66] A. N. Souza and C. R. Doering. Transport bounds for a truncated model of Rayleigh–Bénard convection. *Physica D: Nonlinear Phenomena*, 2015.

[67] A. N. Souza, I. Tobasco, and C. R. Doering. Wall-to-wall optimal transport in two dimensions. *Journal of Fluid Mechanics*, 889, 2020.

[68] E. A. Spiegel. A generalization of the mixing-length theory of turbulent convection. *The Astrophysical Journal*, 138(8):216, 1963.

[69] E. A. Spiegel. Convection in stars I. Basic Boussinesq convection. *Annual Review of Astronomy and Astrophysics*, 9:323–352, 1971.

[70] E. A. Spiegel and G. Veronis. On the Boussinesq approximation for a compressible fluid. *The Astrophysical Journal*, 131(1):442, 1960.

[71] R. J. A. M. Stevens, D. Lohse, and R. Verzicco. Prandtl and Rayleigh number dependence of heat transport in high Rayleigh number thermal convection. *Journal of Fluid Mechanics*, 688:31–43, 2011.

[72] C. A. Stewart and D. L. Turcotte. The route to chaos in thermal convection at infinite Prandtl number: 1. Some trajectories and bifurcations. *Journal of Geophysical Research*, 94(B10), 1989.

[73] J.-L. Thiffeault. Modeling shear flow in Rayleigh–Bénard convection. Master's thesis, University of Texas at Austin, 1995.

[74] J.-L. Thiffeault and W. Horton. Energy-conserving truncations for convection with shear flow. *Physics of Fluids A*, 8(7):1715–19, 1996.

[75] I. Tobasco, D. Goluskin, and C. R. Doering. Optimal bounds and extremal trajectories for time averages in nonlinear dynamical systems. *Physics Letters A*, 382(6):382–386, 2017.

[76] Y. M. Treve and O. P. Manley. Energy conserving Galerkin approximations for 2-D hydrodynamic and MHD Bénard convection. *Physica D: Nonlinear Phenomena*, 4(3):319–342, 1982.

[77] B. Turkington. Vortex rings with swirl: Axisymmetric solutions of the Euler equations with nonzero helicity. *SIAM Journal on Mathematical Analysis*, 20(1):57–73, 1989.

[78] P. Urban, V. Musilová, and L. Skrbek. Efficiency of heat transfer in turbulent Rayleigh-Bénard convection. *Physical Review Letters*, 107(1):014302, 2011.

[79] B. Wen, G. P. Chini, R. R. Kerswell, and C. R. Doering. Time-stepping approach for solving upper-bound problems: Application to two-dimensional Rayleigh–Bénard convection. *Physical Review E*, 92(4):043012, 2015.

[80] B. Wen, D. Goluskin, M. LeDuc, G. P. Chini, and C. R. Doering. Steady coherent convection between stress-free boundaries. *arXiv:2007.02530v1*, 2020.

[81] J. P. Whitehead and C. R. Doering. Ultimate state of two-dimensional Rayleigh–Bénard convection between free-slip fixed-temperature boundaries. *Physical Review Letters*, 106(24), 2011.

[82] A. A. Wing, K. Emanuel, C. E. Holloway, and C. Muller. Convective self-aggregation in numerical simulations: A review. *Surveys in Geophysics*, 38(6):1173–1197, 2017.

[83] Y. Zheng, G. Fantuzzi, A. Papachristodoulou, P. Goulart, and A. Wynn. Chordal decomposition in operator-splitting methods for sparse semidefinite programs. *Mathematical Programming*, 180(1):489–532, 2020.

APPENDIX A

Limiting Cases of the HK8 Model

A.1 The HK8 system in the infinite Prandtl number limit

In this appendix, we examine the HK8 model in the limit of large Prandtl number, and determine upper bounds on N using SOS optimization. In this limit, the HK8 dynamics are reduced to that of two 2-dimensional ODEs that are rescaled versions of the Lorenz equations in the same limit. Proper balancing of terms in the HK8 system suggests that as $\sigma \to \infty$, the shear modes ψ_{01} and ψ_{03} are $O(\sigma^{-1})$, and all other variables are $O(1)$ as $\sigma \to \infty$. Scaling the state variables of the HK8 system according to these assumptions allows the dynamics to be reduced to two 2-dimensional systems:

$$\begin{aligned} \dot{\theta}_{11} &= \frac{k^2+1}{\mathcal{R}_{L_1}}\big(\mathcal{R} - \mathcal{R}_{L_1} - \theta_{02}\big)\theta_{11}, \\ \dot{\theta}_{02} &= -4\theta_{02} + \frac{k^2+1}{2\mathcal{R}_{L_1}}\theta_{11}^2, \end{aligned} \tag{A.1}$$

and

$$\begin{aligned} \dot{\theta}_{12} &= \frac{k^2+4}{\mathcal{R}_{L_2}}\big(\mathcal{R} - \mathcal{R}_{L_2} - 2\theta_{04}\big)\theta_{12}, \\ \dot{\theta}_{04} &= -16\theta_{04} + \frac{k^2+4}{\mathcal{R}_{L_2}}\theta_{12}^2, \end{aligned} \tag{A.2}$$

where

$$\begin{aligned} \psi_{11} &= \frac{k}{(k^2+1)^2}\theta_{11}, \qquad \psi_{01} = -\frac{3k}{4\sigma}\psi_{11}\psi_{12}, \\ \psi_{12} &= -\frac{k}{(k^2+4)^2}\theta_{12}, \quad \psi_{03} = \frac{k}{36\sigma}\psi_{11}\psi_{12}. \end{aligned} \tag{A.3}$$

Under a suitable change of variables, (A.1) and (A.2) are each equivalent to the large–σ limit of the Lorenz equations studied previously in [72]:

$$\begin{aligned} \dot{y} &= (\rho - 1 - z)\, y, \\ \dot{z} &= -\beta z + y^2, \end{aligned} \tag{A.4}$$

corresponding to the restriction of Lorenz equations to the plane $x = y$. Under this restriction, the nonzero equilibria $(y, z) = (\pm\sqrt{\beta(\rho-1)}, \rho-1)$ are globally stable within their respective half-plane ($y > 0$ or $y < 0$) for all $\rho > 1$, and trajectories cannot become chaotic [72]. To obtain (A.4) from (A.1) we change variables according to

$$\begin{gathered}\theta_{11} = \sqrt{2}\mathcal{R}_{L_1}\, y, \qquad \theta_{02} = \mathcal{R}_{L_1}\, z, \qquad (k^2+1)t \mapsto t, \\ \beta = 4/(k^2+1), \qquad \rho = \mathcal{R}/\mathcal{R}_{L_1}.\end{gathered} \tag{A.5}$$

and a similar scaling may be applied to θ_{12} and θ_{04} to obtain (A.4) from (A.2). Therefore the dynamics of (A.1) and (A.2) can each be understood by studying the ODE (A.4).

When $\mathcal{R} > \mathcal{R}_{L_2}$, (A.1) and (A.2) each have three equilibria: a pair of Lorenz-like equilibria corresponding to the L_1 or L_2 states, and the zero equilibrium. Any combination of these provides an equilibrium for the full HK8 system, and therefore there are nine in total. The four such equilibria where both Lorenz-like systems are nonzero are stable when they exist, and correspond to the large-σ limit of the TC states. The maximum N over all equilibria in the large-σ limit is $N_{TC} - 1 = (N_{L_1} - 1) + (N_{L_2} - 1)$. It can be shown using SOS optimization with degree two auxiliary functions that for sufficiently large $\mathcal{R}$, the TC equilibria saturate the upper bound U_d^*, and thus:

$$\max_{\mathbf{x}(t)} N = \begin{cases} 1, & 0 < \mathcal{R} \le R_{L_1}, \\ 3 - 2\mathcal{R}_{L_1}/\mathcal{R}, & 2\mathcal{R}_{L_1} < \mathcal{R} \le \mathcal{R}_{L_2}, \\ 5 - 2\mathcal{R}_{L_1}/\mathcal{R} - 2\mathcal{R}_{L_2}/\mathcal{R}, & \mathcal{R} > \mathcal{R}_{L_2}. \end{cases} \tag{A.6}$$

These upper bounds are a limiting case of the bounds constructed at finite σ in §5.2. In the infinite-Prandtl number limit of the HK8 system, equilibria saturate the upper bound on N for all $\mathcal{R}$ and k. This provides a contrast to the bounds at finite Prandtl number determined in §5.1, where time-dependent states were observed to maximize N for $\mathcal{R}$ sufficiently large.

A.2 TC equilibria in the infinite-$\mathcal{R}$ limit

Expressions for the TC equilibria may be obtained using symbolic manipulation, yet their exact formulae are too complicated to analyze directly. In order to better understand the behavior of the TC equilibria, we compute asymptotic formulae for these equilibria in the large-$\mathcal{R}$ limit. Solutions to the truncated model in this limit have almost no correspondence to solutions of the full PDE; the purpose of this analysis is purely to gain a better understanding of the HK8 model.

The qualitative behavior of the TC states at large $\mathcal{R}$ may be categorized by dividing the k^2–σ plane into three distinct parameter regimes, much like the analysis performed in §3.1. Let S_1 and S_2 be the regions where the TC branch connects to only L_1 or L_2, respectively,

defined by

$$S_1 := \{(k^2,\sigma) : (10\sigma - 3\sigma^2)(k^2+4)^2 + 2(k^2+1)(5k^2+11) \le 0\}, \tag{A.7}$$
$$S_2 := \{(k^2,\sigma) : (10\sigma + 3\sigma^2)(k^2+1)^2 + 2(k^2+4)(5k^2-4) \le 0\}. \tag{A.8}$$

Below, we prove that the large-$\mathcal{R}$ limit of N is

$$N_0^1 = \frac{20(k^2+1)(5k^2+11) + 2(65k^4+313k^2+698)\sigma + 45(k^2+4)^2\sigma^3}{20(k^2+1)(5k^2+11) + 2(65k^4+403k^2+788)\sigma + 9(k^2+4)^2\sigma^3}, \quad (k^2,\sigma) \in S_1,$$
$$N_0^2 = \frac{20(k^2+4)(5k^2-4) + 2(35k^4-83k^2-442)\sigma + 45(k^2+1)^2\sigma^3}{20(k^2+4)(5k^2-4) + 2(35k^4+7k^2-82)\sigma + 9(k^2+1)^2\sigma^3}, \quad (k^2,\sigma) \in S_2. \tag{A.9}$$

The regions S_1 and S_2 correspond with regions V and I–III, respectively, in Figure 3.2. In the part of parameter space that separates S_1 and S_2 (called regions III–IV in Figure 3.2), the TC branch connects to both L_1 and L_2 via pitchfork bifurcations, and as a result TC equilibria only exist for a finite range of $\mathcal{R}$. Hence, the large-$\mathcal{R}$ limit of the TC states need only be considered in S_1 and S_2.

We first derive algebraic conditions on the TC equilibria, using the fact that all eight variables are nonzero at the TC states. Define the variables $X = \psi_{11}^2$ and $Y = \psi_{12}^2$. Then, assuming X and Y are nonzero, the algebraic system whose solutions are equilibria of the HK8 model reduces to

$$\begin{aligned} 0 &= k\left(\mathcal{R}_{L_1} - \mathcal{R}\right) + \frac{k}{8}(k^2+1)^2 X + \left((k^2+1)\alpha + \frac{5k}{12\sigma}(k^2+4)^2\right) Y \\ &\quad + \left(\frac{k^2}{8}\alpha - \frac{5k^2}{12\sigma}\beta\right) XY, \\ 0 &= k\left(\mathcal{R}_{L_2} - \mathcal{R}\right) - \left((k^2+4)\beta + \frac{5k}{12\sigma}(k^2+1)^2\right) X + \frac{k}{8}(k^2+4)^2 Y \\ &\quad - \left(\frac{k^2}{8}\beta + \frac{5k^2}{12\sigma}\alpha\right) XY, \end{aligned} \tag{A.10}$$

where

$$\alpha = \frac{k(5k^2+11)}{12\sigma^2}, \qquad \beta = \frac{k(5k^2-4)}{12\sigma^2}. \tag{A.11}$$

After further simplification, (A.10) takes the form

$$X = \frac{k\mathcal{R} + C_0 + C_1 Y}{D_0 + D_1 Y}, \tag{A.12}$$

where Y solves the quadratic equation

$$Y^2 + (A_0\mathcal{R} + A_1)Y + (B_0\mathcal{R} + B_1) = 0. \tag{A.13}$$

The constants A_i, B_i, C_i, D_i are independent of $\mathcal{R}$ and may be determined from (A.10). To determine asymptotic expansions for X and Y as $\mathcal{R} \to \infty$, we let $\varepsilon = 1/\mathcal{R}$ and multiply (A.13) by ε, resulting in the singular perturbation problem

$$\varepsilon Y^2 + (A_0 + A_1\varepsilon)Y + (B_0 + B_1\varepsilon) = 0. \tag{A.14}$$

Solutions to (A.13) could also be determined using the quadratic formula, but simple analytical expressions for N are simpler to derive with perturbation methods.

A.2.1 Outer approximation

The equilibria corresponding to the infinite-$\mathcal{R}$ limit of the TC equilibria in S_2 can be found by substituting the expansion $Y \sim \sum_n \varepsilon^n Y_n$ into (A.14). This yields a hierarchy of equations for Y_n, with the leading term Y_0 given by

$$Y_0 = -\frac{B_0}{A_0}. \tag{A.15}$$

Substituting the series for Y into (A.12) yields a geometric series with leading order

$$x_0 = \frac{k}{D_0 + D_1 Y_0}. \tag{A.16}$$

Because all eight variables must be real and nonzero, the above expansions provide limiting behavior for the TC equilibria as long as $x_0, Y_0 > 0$, corresponding exactly to the set S_2. Within this region, (A.15) and (A.16) determine the limiting behavior for all eight variables on the TC branch. In this case it is more useful to express N directly in terms of X and Y, yielding an asymptotic series for N. After simplification, the volume-averaged expression (2.44) for N becomes

$$N = 1 + \frac{k}{4\mathcal{R}}\left(\frac{(k^2+1)^2}{k}X + (\alpha - \beta)XY + \frac{(k^2+4)^2}{k}Y\right). \tag{A.17}$$

The expressions (A.15) and (A.16) imply that the leading order behavior of N in S_2 is given by

$$\begin{aligned} N_0^2 &= \frac{k}{4}\left(\frac{(k^2+1)^2}{k}x_0 + (\alpha - \beta)(x_0 Y_0)\right) \\ &= \frac{20(k^2+4)(5k^2-4) + 2(35k^4 - 83k^2 - 442)\sigma + 45(k^2+1)^2\sigma^3}{20(k^2+4)(5k^2-4) + 2(35k^4 + 7k^2 - 82)\sigma + 9(k^2+1)^2\sigma^3}. \end{aligned} \tag{A.18}$$

Further terms in the expansion may be computed using the asymptotic series computed above. It can be shown that this expansion is valid in the set S_2, where the TC branch connects only to L_2, as elsewhere the leading-order terms of X or Y will be negative. Within the interior of S_2, N_0^2 is strictly increasing in both k^2 and σ, reaching its maximum value of 3 along the entirety of the interior boundary of S_2, and approaching its minimum value of 1 as σ vanishes. The leading-order term has a jump discontinuity at the point $(k^2, \sigma) = (4/5, 0)$.

Convergence of the asymptotic series may be demonstrated by comparing its first few terms against computed values of N at chosen parameter values. As expected, the leading-order expansion converges at a rate of $O(\varepsilon^2)$ as $\varepsilon \to 0$, while including more terms improves the order of accuracy.

A.2.2 Inner approximation

The large-$\mathcal{R}$ limit of the TC equilibria in S_1 can be found after rescaling (A.14) by $y = \varepsilon Y$, to obtain

$$y^2 + (A_0 + A_1\varepsilon)y + (B_0\varepsilon + B_1\varepsilon^2) = 0. \tag{A.19}$$

This scaling emerges when seeking a dominant balance between the first two terms of (A.14). We then proceed as in A.2.1 to find that the leading-order expansion for N is given by

$$\begin{aligned} N_0^1 &= 1 + \frac{k}{4}\left((\alpha - \beta)X_0 y_0 + \frac{(k^2+4)^2}{k} y_0\right) \\ &= \frac{20(k^2+1)(5k^2+11) + 2(65k^4+313k^2+698)\sigma + 45(k^2+4)^2\sigma^3}{20(k^2+1)(5k^2+11) + 2(65k^4+403k^2+788)\sigma + 9(k^2+4)^2\sigma^3}. \end{aligned} \tag{A.20}$$

where

$$y_0 = -A_0, \qquad X_0 = \frac{k + C_1 y_0}{D_1 y_0}. \tag{A.21}$$

This solution corresponds to the asymptotic state of the TC branch in S_1, which is of particular interest since the TC equilibria maximize N among the steady states of the HK8 model at sufficiently large $\mathcal{R}$ in this parameter regime. The leading-order term N_0^1 is nearly constant in k^2 and is strictly increasing in σ, rapidly approaching 5 as $\sigma \to \infty$. As (k^2, σ) approaches the interior boundary of S_1, the TC equilibria approximate L_1 states and therefore N_0^1 tends to 3.

APPENDIX B

Special Case of the Sum-of-Squares Method: Results of Souza and Doering

The analysis of Souza & Doering in [66] amounts to a special case of the quadratic SOS approach implemented for the HK8 model in §5.2, with the coefficients of the general quadratic auxiliary function (5.12)–(5.13) restricted more than necessary. The bounds they prove are identical to those proven in §5.2 when $\mathcal{R} \leq \sqrt{\mathcal{R}_{L_1}\mathcal{R}_{L_2}}$, but for larger $\mathcal{R}$ they prove instead that [66]

$$N \leq N_{L_1} + \frac{1}{\mathcal{R}}\left[\mathcal{R}_{L_1} - \mathcal{R}_{L_2} + \sqrt{2}\sqrt{(\mathcal{R}-\mathcal{R}_{L_1})^2 + (\mathcal{R}-\mathcal{R}_{L_2})^2}\right]. \tag{B.1}$$

On this interval, more general quadratic V give sharper bounds, as shown by the results of §5.1 and §5.2.

To show how the analysis of [66] fits into our present framework, let us derive the large-$\mathcal{R}$ bound of (B.1) in the language of §5.2. Let z_1 and z_2 be constants to be chosen below—the "background variables" in the language of [66]—and define the constant

$$\alpha = \frac{z_1 + z_2}{z_1^2 + z_2^2}. \tag{B.2}$$

Our approach in §5.2 reduces to the special case of [66] if the coefficients c_i of V in (5.12) are restricted such that

$$\begin{aligned} c_1 &= \frac{1-2\alpha z_1}{2\mathcal{R}}, & c_2 &= \frac{1-2\alpha z_2}{4\mathcal{R}}, & c_3 &= \frac{(k^2+1)(\alpha-1)}{8\sigma\mathcal{R}}, \\ c_5 &= \frac{(k^2+4)(\alpha-1)}{8\sigma\mathcal{R}}, & c_8 &= \frac{\alpha}{8\mathcal{R}^2}, & c_9 &= 0. \end{aligned} \tag{B.3}$$

With the coefficients constrained by (5.13) and (B.3), there are only two free parameters in the expression for V, and therefore the auxiliary function is determined by specifying z_1 and z_2. Under these restrictions on V, the minimal upper bound such that the polynomial S is sum-of-squares is

$$N \leq 1 + 2(z_1 + z_2), \tag{B.4}$$

provided z_1 and z_2 are chosen so that $\alpha \geq 0$ and

$$\frac{\mathcal{R}_{L_1}}{\mathcal{R}}(\alpha - 1) - \alpha(z_1 - 1)^2 \geq 0, \tag{B.5}$$

$$\frac{\mathcal{R}_{L_2}}{\mathcal{R}}(\alpha - 1) - \alpha(z_2 - 1)^2 \geq 0. \tag{B.6}$$

The bound (B.1) in the $\mathcal{R} > \sqrt{\mathcal{R}_{L_1}\mathcal{R}_{L_2}}$ case is then constructed by taking [66]

$$z_1 = \frac{\sqrt{\mathcal{R}_{L_1}}\left[-\mathcal{R}_{L_1} + 2\mathcal{R}\sqrt{\frac{\mathcal{R}_{L_2}}{\mathcal{R}_{L_1}}} - \mathcal{R}_{L_2} + \sqrt{(\mathcal{R}_{L_1} + \mathcal{R}_{L_2})^2 + 4\mathcal{R}\left(\mathcal{R} - 2\sqrt{\mathcal{R}_{L_1}\mathcal{R}_{L_2}}\right)}\right]}{2\mathcal{R}(\mathcal{R}_{L_1} + \mathcal{R}_{L_2})},$$

$$z_2 = \sqrt{\frac{\mathcal{R}_{L_2}}{\mathcal{R}_{L_1}}}(z_1 - 1) + 1. \tag{B.7}$$

This bound is valid whenever $\mathcal{R} > \sqrt{\mathcal{R}_{L_1}\mathcal{R}_{L_2}}$ but is not as tight as the result obtained in §5.1 and §5.2 using the most general quadratic ansatz for V.

APPENDIX C

Analytical Bounds in the Larger-Wavenumber Regime

This appendix gives partial results towards the analytical optimization of N with quadratic V in the regime where $\mathcal{R} > \mathcal{R}_T$ and $k > k_T$ (see Figure 5.5). In this regime the bound (5.21) does not hold, so the assumptions (5.22) are not valid. Our best analytical bound coincides with that of [66]:

$$N \leq N_{L_1} + \frac{1}{\mathcal{R}} \left[\mathcal{R}_{L_1} - \mathcal{R}_{L_2} + \sqrt{(\mathcal{R}_{L_2} - \mathcal{R}_{L_1})^2 + 4(\mathcal{R} - \sqrt{\mathcal{R}_{L_1}\mathcal{R}_{L_2}})^2} \right]. \tag{C.1}$$

The exact auxiliary function required to prove this bound is given in Appendix B. However, this bound is not optimal among quadratic V; numerical solution to (4.13) with quadratic V gives sharper bounds at many parameter values.

In order for the construction of optimal V to become analytically tractable, we want to further restrict V in a way that will be justified *a posteriori* by the sharpness of the resulting bounds, much as was done for the smaller-k regime in §5.2.3. If the bounds are saturated by the TC equilibria, then S must vanish there. Since $\psi_{01} = -27\psi_{03}$ on the TC equilibria,

$$\mathbf{b}_2^T \mathcal{Q}_2 \mathbf{b}_2 = q_2 \left(\psi_{01} + 27\psi_{03}\right)^2. \tag{C.2}$$

We further observe in SOS computations with quadratic V that the determinants of $\mathcal{Q}_1$ and $\mathcal{Q}_3$ are zero up to the tolerance of the solver. This implies that for some q_1, q_2, A_1, and A_2,

$$\mathbf{b}_1^\mathsf{T} \mathcal{Q}_1 \mathbf{b}_1 = q_1 \left(\psi_{11} - A_1\theta_{11}\right)^2, \qquad \mathbf{b}_3^\mathsf{T} \mathcal{Q}_3 \mathbf{b}_3 = q_3 \left(\psi_{12} - A_2\theta_{12}\right)^2. \tag{C.3}$$

These restrictions impose the coefficient relationships

$$\begin{aligned} c_1 &= 8\frac{k^2+1}{k}\sqrt{\sigma c_3 c_8} - 4\mathcal{R}c_8 - \frac{4\sigma}{k^2+1}c_3, \quad c_3 = \frac{(5k^2-4)(k^2+1)}{(5k^2+11)(k^2+4)}c_5, \\ c_2 &= 4\frac{k^2+4}{k}\sqrt{\sigma c_5 c_8} - 2\mathcal{R}c_8 - \frac{2\sigma}{k^2+4}c_5, \quad c_9 = \frac{108}{(k^2+4)(5k^2+11)}c_5. \end{aligned} \tag{C.4}$$

The semidefinite constraints will be satisfied if c_5 and c_8 are each nonnegative, so it remains to determine c_5 and c_8 that minimize U. By the same argument used in §5.2.3, the optimal

$\mathcal{Q}_4$ must have a determinant of zero, in which case the SOS optimization is equivalent to

$$\min_{c_5, c_8 \geq 0} U, \tag{C.5}$$

where

$$\begin{aligned} U - 1 = \frac{1}{16 c_8} \Bigg[& \left(8\mathcal{R} c_8 + \frac{1}{\mathcal{R}} + \frac{8\sigma\mu}{k^2+1} c_5 - 16 \frac{k^2+1}{k} (\sigma \mu c_5 c_8)^{1/2} \right)^2 \\ & + \left(8\mathcal{R} c_8 + \frac{1}{\mathcal{R}} + \frac{8\sigma}{k^2+4} c_5 - 16 \frac{k^2+4}{k} (\sigma c_5 c_8)^{1/2} \right)^2 \Bigg], \end{aligned} \tag{C.6}$$

with

$$\mu = \frac{(5k^2-4)(k^2+1)}{(5k^2+11)(k^2+4)}. \tag{C.7}$$

Numerical solutions of (C.5)–(C.7) agree with numerical solutions to the full SOS optimization problem in this parameter regime, suggesting the assumptions (C.2)–(C.3) leading to this simpler minimization problem are not overly restrictive. However, we have not been able to derive an analytical solution to (C.5)–(C.7) that is simple enough to be useful.

APPENDIX D

Application of a Diagnostic Procedure to Analyze Euler Codes

D.1 Background

Numerical solvers for the 3D incompressible Euler equations are essential tools in the endeavor to understand the many intricacies of fluid mechanics. One noteworthy application of Euler codes has been their use in the search for evidence of finite time singularity formation in incompressible, inviscid flows from smooth initial data. This problem, analogous to the well-known Millennium Prize Problem for the Navier-Stokes equations, has resulted in a wealth of research conducted towards its resolution. Numerical studies by Grauer and Sideris [26], Pumir and Siggia [57], Luo and Hou [42], and many others have addressed the problem, yet uncertanties have persisted, and in some cases even convincing studies were contradicted by later evidence [31]. A more comprehensive list of studies of singularity formation for the incompressible Euler equations is provided in [19].

The Euler equations arise as the inviscid limit of the Navier Stokes equations, expressed as

$$\begin{aligned} \frac{D\mathbf{u}}{Dt} &= -\nabla p, \\ \nabla \cdot \mathbf{u} &= 0, \end{aligned} \tag{D.1}$$

for $\mathbf{x} \in \Omega \subseteq \mathbb{R}^n$, where $\mathbf{u}(\mathbf{x}, t)$ is the divergence-free velocity field, $p(\mathbf{x}, t)$ is the pressure field, and D/Dt is the material derivative, given by

$$\frac{D}{Dt} = \partial_t + \mathbf{u} \cdot \nabla. \tag{D.2}$$

It is convenient to express this equation in terms of the vorticity field, where the vorticity is defined by $\boldsymbol{\omega} = \nabla \times \mathbf{u}$. The corresponding evolution equations, derived by taking the curl of (D.1), are

$$\frac{D\boldsymbol{\omega}}{Dt} = \boldsymbol{\omega} \cdot \nabla \mathbf{u}, \tag{D.3}$$

Several criteria have been established to provide evidence for or against the existence of a finite-time singularity of the incompressible Euler equations, including the Beale-Kato-

Majda (BKM) blowup criterion [3]. This condition states that a solution to the 3D Euler equations with smooth initial data forms a singularity at time T if and only if

$$\int_0^T \|\boldsymbol{\omega}(\cdot,t)\|_\infty \,\mathrm{d}t = \infty, \tag{D.4}$$

where $\|\cdot\|_\infty$ is the maximum norm over the domain. Alternatively, the Deng-Hou-Yu (DHY) nonblowup criterion [9] provides a pair of conditions on the length and curvature of vortex lines that cannot simultaneously be satisfied in the presence of a singularity.

For example, Luo and Hou [42] produced evidence for the potential finite-time blowup of solutions in the 3D axisymmetric case. They solved the axisymmetric Euler equations numerically in a vertically periodic cylinder with period H, and solid outer boundary. They observed evidence of a singularity starting from the following initial conditions on the azimuthal components of the velocity and vorticity fields:

$$u^\theta(r,z) = 100r \cdot e^{-30(1-r^2)^4} \sin\left(\frac{2\pi}{H} z\right) \tag{D.5}$$

$$\omega^\theta(r,z) = 0. \tag{D.6}$$

Careful evidence was given to support the existence of a singularity, including checking the BKM and Deng-Hou-Yu criteria, verifying conservation of energy, and providing evidence of self-similar solutions.

While the BKM and DHY criteria are often cited as evidence of singularity formation, including in the study above, applying them assumes the flow is well-resolved. It is conceivable that a computed solution may appear to become singular while the true solution exhibits only rapid, non-singular vorticity growth. Any candidate for singular behavior will pose a significant numerical challenge to even the most highly accurate solvers, leading to uncertainty in this delicate classification. Even if one can ensure that the numerical method converges, it is still possible that it does not converge to the correct solution, as demonstrated by Mishra and Spinolo [46]. In light of these issues, devising additional tests demonstrating the validity of the computed solutions may create a more complete picture for assessing the potential singularity.

The time reversibility of the Euler equations for energy conserving schemes allows one to estimate the numerical errors by assessing the program's ability to recover the initial condition from a future state, as proposed by Duponcheel et al. [16]. This benchmark allows analysis of the accuracy in temporal discretization. By contrast, the diagnostic implemented in this Appendix is an attempt to address the errors in spatial discretization. The combination of spatial and temporal diagnostics would provide a more complete analysis of the success of the solver, and so it is to this end that the diagnostic is presented.

D.2 The diagnostic procedure

We next detail the proposed test, first introduced by Doering, Gibbon, and Holm [15]. First, let $A(\mathbf{x}, t)$ be a passive tracer field whose evolution under the dynamics of (D.3) is governed by the advection equation:

$$\frac{DA}{Dt} = 0. \tag{D.7}$$

This quantity could represent the concentration of dye or the potential temperature[1] in geophysical flows [20], and its distribution does not influence the flow. As the test will be constructed from gradients of the dye concentration, its initial distribution is crucial to the application of the diagnostic. It may be scattered throughout the domain, or concentrated to illuminate some particular feature of the flow.

Next we define the potential vorticity q as

$$q(\mathbf{x}, t) = \boldsymbol{\omega}(\mathbf{x}, t) \cdot \nabla A(\mathbf{x}, t). \tag{D.8}$$

Ertel's Theorem [61] states that the potential vorticity is also conserved by the flow, i.e.

$$\frac{Dq}{Dt} = 0. \tag{D.9}$$

The potential vorticity is well known in the field of geophysical fluid mechanics, where it represents a modified vorticity field, independent of the effects of pressure and latitude. Since it is conserved by the Euler equations, the potential vorticity is often a quantity of interest in atmospheric and ocean science [30], where it can be used as a Lagrangian tracer either for predictive models or to reconstruct the flow pattern from present information. In addition, the potential vorticity has been under recent scrutiny from Gibbon and Holm [20] for more general applications.

Now define a divergence free vector field $\mathbf{B}$ as

$$\mathbf{B}(\mathbf{x}, t) = \nabla q(\mathbf{x}, t) \times \nabla A(\mathbf{x}, t) \tag{D.10}$$

It is straightforward to verify that B obeys the same evolution equation as the vorticity:

$$\frac{D\mathbf{B}}{Dt} = \mathbf{B} \cdot \nabla \mathbf{u}. \tag{D.11}$$

Any algorithm designed to compute the evolution of the vorticity field governed by (D.3) can be directly applied to evolve $\mathbf{B}$, and a small modification allows the quantities A and q to be computed simultaneously.

Once these quantities have been defined, the diagnostic procedure is as follows [15]:

[1] The potential temperature is the temperature a fluid parcel would have if it were transported adiabatically to a given reference pressure.

1. Specify initial conditions for $\mathbf{u}$ and A, and compute the emergent initial data for q and $\mathbf{B}$ from (D.8) and (D.10)

2. Simultaneously evolve each quantity forward in time to obtain $\boldsymbol{\omega}(\mathbf{x}, t)$, $A(\mathbf{x}, t)$, $q(\mathbf{x}, t)$, and $\mathbf{B}(\mathbf{x}, t)$.

3. Compute $q_1(\mathbf{x}, t) = \boldsymbol{\omega}(\mathbf{x}, t) \cdot \nabla A(\mathbf{x}, t)$ and $\mathbf{B}_1(\mathbf{x}, t) = \nabla q_1(\mathbf{x}, t) \times \nabla A(\mathbf{x}, t)$, and compare to their corresponding evolved quantities $q(\mathbf{x}, t)$ and $\mathbf{B}(\mathbf{x}, t)$.

4. Compute $\mathbf{B}_2(\mathbf{x}, t) = \nabla q(\mathbf{x}, t) \times \nabla A(\mathbf{x}, t)$ from the evolved vorticity q, and compare this to the evolved $\mathbf{B}(\mathbf{x}, t)$.

Because the diagnostic fields contain information about the velocity and vorticity fields and their derivatives, the diagnostics can be used to illustrate the numerical errors introduced during evolution of the Euler variables. Clearly, if no errors were introduced during computation, the evolved diagnostics would not deviate from the corresponding quantities computed prior to computation. Furthermore, since $\mathbf{B}$ and $\boldsymbol{\omega}$ are subject to the same equations of motion, they may each be evolved in exactly the same manner, enabling the test to be implemented without significant difficulty.

D.3 Application of the diagnostic procedure

The proposed test was adapted into an existing Euler code written by Diego Ayala [2]. Spatial derivatives are computed with a pseudo-spectral method, and a Runge-Kutta method of order 2 was used for time discretization to generate the data for all of the plots shown below. Axisymmetry allows the Euler equations to be reduced to two dimensions, and thus all computations are performed in the r–z plane. The computational domain Ω is periodic in the axial direction. At the outer boundary, $\mathbf{u}$ satisfies the Dirichlet boundary conditions

$$\hat{\mathbf{v}} \cdot \mathbf{u} = 0, \quad \mathbf{x} \in \partial\Omega, \tag{D.12}$$

where $\hat{\mathbf{v}}$ is the outward normal vector on $\partial\Omega$.

For such axisymmetric flows, it is convenient to define the following variables (see Turkington [77] for a more detailed discussion):

$$\zeta(r, z, t) = \frac{1}{r}\,\omega^{\theta}(r, z, t), \qquad \gamma(r, z, t) = r u^{\theta}(r, z, t),$$

where ω^{θ}, u^{θ} are the azimuthal components of vorticity and velocity, respectively. Next, transform the coordinate system by defining:

$$\rho = \frac{1}{2} r^2. \tag{D.13}$$

We then define a stream function ψ such that $u^z = (\partial_r \psi)/r$, and $u^r = -(\partial_z \psi)/r$. In ρ–z coordinates, ψ solves the Poisson problem

$$L\psi = \zeta, \qquad (\rho, z) \in \Omega, \tag{D.14}$$

where L is defined by

$$L = -\frac{1}{2\rho}\frac{\partial^2}{\partial z^2} - \frac{\partial^2}{\partial \rho^2}. \tag{D.15}$$

Let $\{f, g\} := \partial_\rho f \partial_z g - \partial_\rho f \partial_z f$ denote the Jacobian of the functions f and g. Then (D.3) implies that ζ and γ evolve according to:

$$\frac{d\zeta}{dt} + \{\zeta, \psi\} + \{\gamma, \gamma/2\rho\} = 0, \tag{D.16}$$

$$\frac{d\gamma}{dt} + \{\gamma, \psi\} = 0, \tag{D.17}$$

for all $(z, \rho) \in \Omega$.

In the axisymmetric case, the vectors ∇q and ∇A lie in the r-z plane, so the diagnostic vector is given by $\mathbf{B} = B^\theta \, \hat{e}_\theta$. For simplicity, we take $B := B^\theta$. The equations of motion for B take the form:

$$\frac{dB}{dt} = \{\psi, B\} - \frac{1}{2\rho}\frac{\partial \psi}{\partial z} B \tag{D.18}$$

In the following, let initial conditions on ζ and γ be given by

$$\zeta(\rho, z, 0) = \gamma(\rho, z, 0) = C\rho \exp\left[-\left(z^2 + (\rho - 0.1)^2\right)^2\right], \tag{D.19}$$

where C is a normalizing constant, here defined such that the initial enstrophy of the initial data takes a given value E_0, that is

$$E_0 = \frac{1}{2}\int_\Omega |\boldsymbol{\omega}(\rho, z, 0)|^2 \, \mathrm{d}\rho \, \mathrm{d}z. \tag{D.20}$$

These initial data describe a single vortex ring with swirl, where the azimuthal components of velocity and vorticity are localized within a toroidal region centered about $r = 0$ (See Figure D.1). Fluid particles experience both toroidal and poloidal rotation.

Applying the diagnostic also requires a choice of initial condition for the tracer A. In this study, the initial condition for A is given by

$$A(\rho, z, 0) = -C_A z\rho(\rho - 8)^8 \exp\left[-\left(z^2 + (\rho - 0.01)^2\right)\right], \tag{D.21}$$

where the constant C_A is chosen so that $\|A\|_2 = 1$. The resulting distribution of the tracer is a pair of Gaussian-like impulses in the r–z plane, as shown in Figure D.1. The initial tracer concentration was designed so that its largest gradients are localized near the peaks in

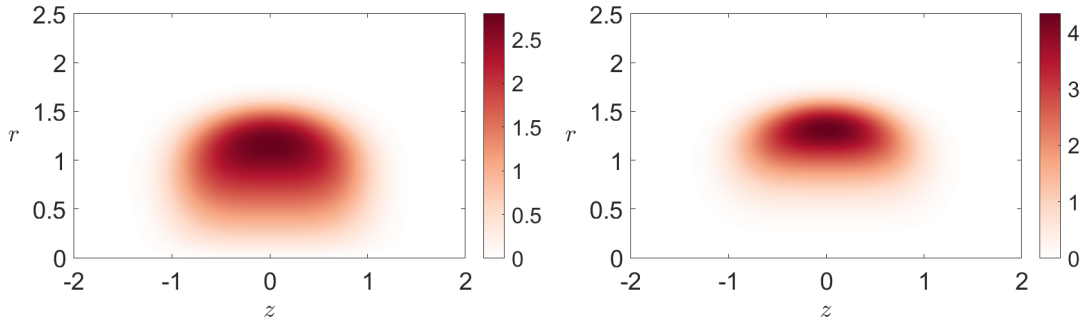
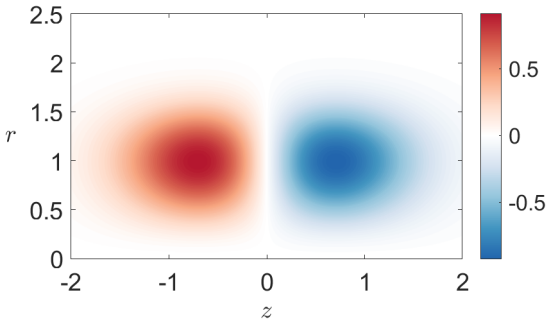


Figure D.1: Initial conditions for u_θ (left), ω_θ (right), and A (below).



the initial data for ω^θ and u^θ, so that the evolution of the q and $\mathbf{B}$ fields capture the salient features of the dynamics.

We implement the diagnostic procedure outlined in the previous section in numerical simulations with resolutions of 256^2, 512^2, and 1024^2 in Fourier space. A snapshot of the flow after evolution forward to $t = 1$ at a resolution of 1024^2 is shown in Figure D.2. As a consequence of the initial vorticity, flow through the center of the ring quickly destroys the symmetry about $z = 0$, and the ring begins to roll into a crescent shape. As the vorticity stretches, a region of negative vorticity begins to form on the front edge of the ring.

The diagnostic fields A, q and B were evolved alongside the vorticity and velocity fields, updating each quantity at each time step. The difference between the evolved and computed diagnostics are displayed in Figure D.3. The largest discrepancy in the diagnostic is localized near the edges of the vortex. This is not unexpected, due to the large gradients near the edges and rapid localization of vorticity experienced by the flow.

This procedure naturally admits two measures of the global computational error

$$\frac{\|q - q_1\|_2}{\|q\|_2} \qquad \text{and} \qquad \frac{\|\mathbf{B} - \mathbf{B}_2\|_2}{\|\mathbf{B}\|_2}, \tag{D.22}$$

in some relevant spatial norm. Figure D.4 depicts the time evolution these quantities in the L^2 norm in simulations with spatial resolutions of 256^2, 512^2, and 1024^2. Here the diagnostic indicates that, as one may expect, increasing the resolution results in reduced computation error. Additionally, the diagnostic test indicates that simulating the flow at higher resolution maintains accuracy for a longer period of time, although at all resolutions

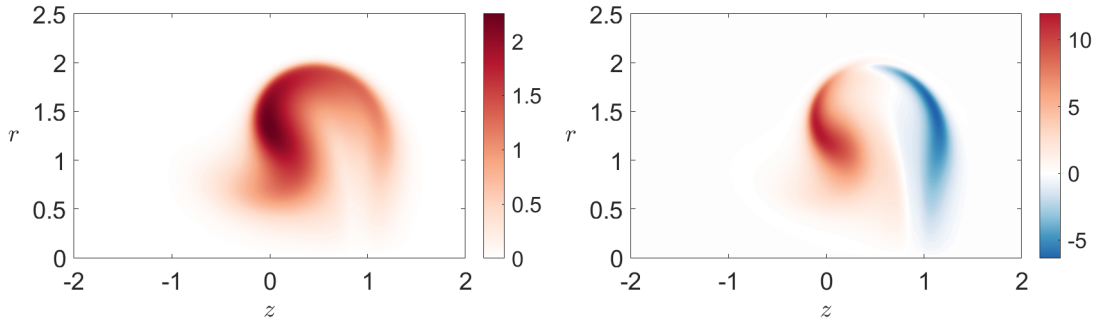
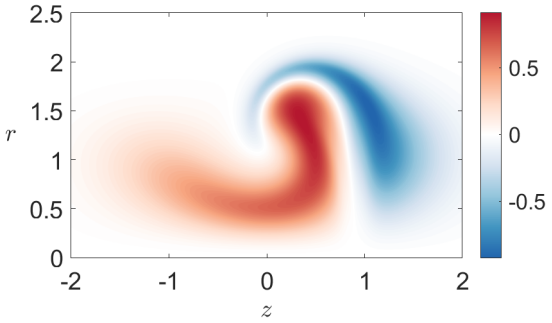


Figure D.2: Azimuthal velocity (left), vorticity (right), and A (below) at time $t = 1$, computed at a resolution of 1024^2 from the initial conditions (D.19)–(D.21)



the simulation loses validity well before $t = 2$. This suggests that localized mesh refinement may be necessary to perform accurate simulations over an extended period of time. At each resolution, we observe a sharp increase in error when the second measure of the diagnostic error nears 10^{-4}. After this point, errors quickly begin to accumulate, indicating that the simulation may be starting to break down.

D.4 Conclusions and Future Research

The diagnostic procedure introduced in [15] was implemented for a simulation of a vortex ring with swirl under the dynamics of the Euler equations. By evolving higher order derivatives of the flow, this test detects errors in the spatial discretization. It could be applied to measure numerical errors in other simulations of the Euler equations, and may be of particular use in analyzing candidates for finite time singularities. An obvious next step in testing the procedure would be to apply the test to initial conditions that lead to more complex behavior. For instance, initial conditions that lead to collision of vortex rings are challenging computationally due to the rapid enstrophy growth and small length scales. The vortex stretching mechanism in $\mathbf{B}$ may enhance the small-scale features of the flow and illuminate problematic locations in the computational domain.

Distribution of the passive tracer A is important to the procedure outlined above. Careful placement of the initial concentration helps illuminate the features of the flow which are most taxing on the numerical scheme. Future studies could introduce methods for allocating

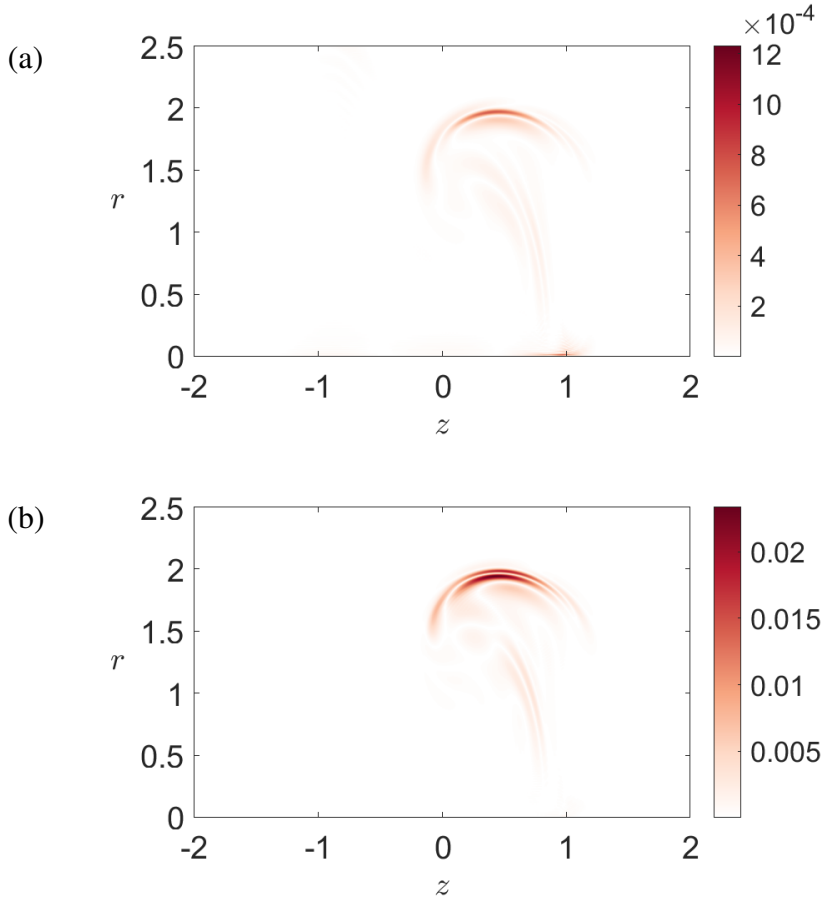


Figure D.3: Relative error diagnostics at $t = 1$, computed at a resolution of 1024^2. In (a), the field $|q - q_1|$ is plotted, while (b) depicts the field $|\mathbf{B} - \mathbf{B}_2|$.

the initial tracer distribution to study the relevant dynamics. For example, data from previous simulations could inform the choice of the initial condition for A in the next trial.

Another idea is to implement the diagnostic procedure for mesh refinement, since the error in the quantities q and $\mathbf{B}$ indicate areas where higher resolution may be required. This requires a carefully designed process for distributing the dye, as mentioned above. If applied properly, this could be especially useful in identifying candidates for a singularity, as choosing an adequate mesh can be crucial to the success of the solver.

Further research is needed to investigate the mathematical properties of both the potential vorticity and the $\mathbf{B}$ vector. At present, the research into the quantities has been limited, and a deeper investigation may open up other possibilities. Establishing a relationship between measures of the diagnostic error and more universal measures of error relating to the flow would be significant in justifying the applicability of the diagnostic.

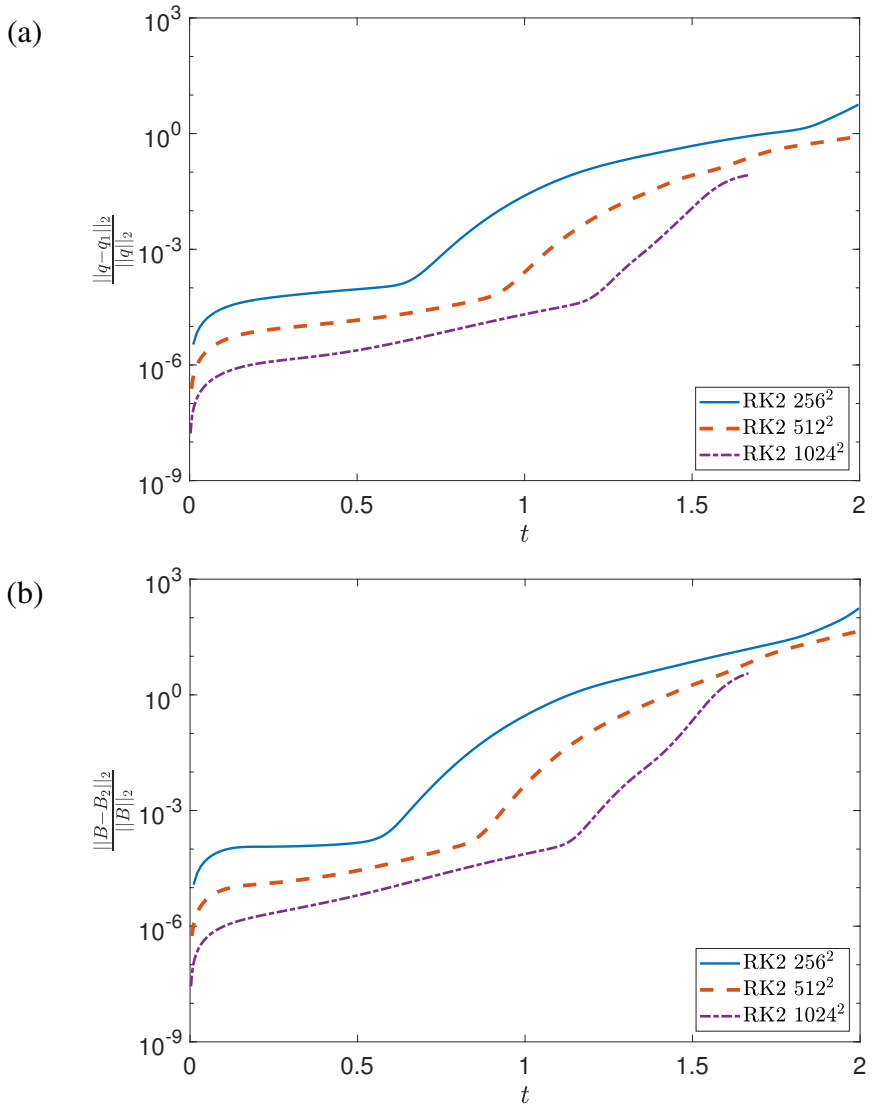


Figure D.4: Propagation of two measures of the diagnostic error in the first two time units, at spatial resolutions of 256^2, 512^2 and 1024^2. Time discretization was performed by the Runge-Kutta method of order two. The potential vorticity error is shown in (a) and the $\mathbf{B}$ diagnostic error in (b).