Imperial College London

Construction of optimal background fields using semidefinite programming

Giovanni Fantuzzi

supervised by Dr. Andrew Wynn

Imperial College London,
Department of Aeronautics

This thesis is submitted for the degree of Doctor of Philosophy
and the Diploma of Imperial College London

May 2018

Abstract

Quantitative analysis of systems exhibiting turbulence is challenging due to the lack of exact solutions and the cost of accurate simulations, but asymptotic or time-averaged properties can often be bounded rigorously using the background method. This rests on the construction of a background field for the system subject to a spectral constraint, which requires that a background-field-dependent linear operator has non-negative eigenvalues. This thesis develops techniques for the numerical optimisation of background fields and their corresponding bounds.

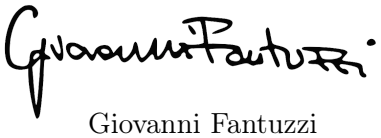
First, bounds on the asymptotic energy of solutions of the Kuramoto–Sivashinsky equation are optimised by solving the Euler–Lagrange (EL) equations for the optimal background field using a time-marching algorithm. It is demonstrated that convergence to incorrect solutions occurs unless the derivation of the EL equations accounts for the multiplicity of eigenvalues in the spectral constraints.

Second, semidefinite programmes (SDPs) are formulated to approximately solve optimisation problems subject to a class of integral inequalities on function spaces, to which spectral constraints can often be reduced. More precisely, inner and outer approximations of the feasible set of an integral inequality with one-dimensional compact integration domain, whose integrand is quadratic in the test functions and affine in the optimisation variables, are constructed using linear matrix inequalities.

These SDP-based techniques, implemented in the MATLAB toolbox QUINOPT, are then utilised to bound the dissipation coefficient C_ε in stress-driven shear flows, and further improved to bound the Nusselt number Nu in Bénard–Marangoni convection at infinite Prandtl number. The results suggest that the existing analytical bounds on C_ε attain the optimal asymptotic scaling, while those on Nu may be lowered by a logarithmic factor upon constructing a non-monotonic background field. It is also concluded that semidefinite programming will offer an efficient, robust, and flexible framework to optimise background fields if the computational challenges presented by large-scale SDPs can be addressed.

Declaration of authorship & copyright

I hereby declare that this thesis is the product of my own research only and no parts of it has been submitted for another degree. References or quotations from other works are fully acknowledged. When parts of this thesis have been published, this has been clearly indicated and references to the published work has been given.



Giovanni Fantuzzi

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Andrew Wynn. If any of the work I have done during my Ph.D. can be considered a success, this is largely thanks to the guidance, encouragement, and advice he has offered throughout the last four years. I am also indebted to him for teaching me how to think, if not like an applied mathematician, at least like a "pure engineer".

I am also extremely grateful to Imperial College, its alumni community, and the Engineering Council for Physical and Engineering Research (EPSRC) for funding this project through the Imperial College President's Ph.D. Scholarship scheme. Without this support, this project would not have been carried out.

Heartfelt thanks also go to all those academics with whom I've had the pleasure to share ideas about this work, and whose suggestions have offered invaluable help. In particular, I am grateful to Prof. Charles Doering, for enthusiastically supporting this work even before it began; to Prof. Sergei Chernyshenko and Dr. David Goluskin, for sharing their boundless insight into the world of bounds; to Prof. Antonis Papachristodoulou, for introducing me to the interesting field of large-scale semidefinite programming; to Prof. Paul Goulart, who helped me discover that writing efficient code can feel extremely rewarding.

Finally, I would also like to thank my parents, my family, and my friends, who have constantly encouraged me to keep going despite any difficulties. Among these wonderful people, special thanks go to Theresa: thank you for sitting through endless conversations (should I say monologues?) about my research, thank you for motivating me when things did not work, thank you for always trying to keep a smile on my face. This journey would not have been the same without you.

Contents

List of figures

List of tables

Chapter 1

Introduction

Systems governed by nonlinear partial differential equations (PDEs), whose dynamics are characterised by complex evolution over time and space, are ubiquitous in physics and engineering. The most notable example is perhaps that of fluid flows, which become turbulent when the force driving the flow is increased. Turbulent flows are a paradigm for dynamical systems whose behaviour is chaotic, in the sense that they may appear to lack any spatio-temporal coherence and that they are extremely sensitive to the initial condition. Such systems are therefore often referred to as *turbulent*.

Given the widespread occurrence of turbulent systems, deriving a quantitative description of their properties is of interest across many scientific disciplines and industries. For instance, a quick but reliable prediction of the aerodynamic drag of a certain wing-body configuration would enable aircraft manufacturers to assess the performance of innovative aircraft concepts, and therefore optimise them in order to minimise fuel consumption. Similarly, knowing how much heat is transported by ocean currents and how much heat, mass, and energy are exchanged by the ocean and the atmosphere is indispensable in climate science to develop accurate climate models for weather forecasting (Cushman-Roisin & Beckers, 2011, chapter 1). As a final example, quantifying the strength of the convective flow of magma in planetary cores can answer questions in astrophysics, such as how mantle convection influences the magnetic field of the Earth (Biggin *et al.*, 2012).

Quantitative analysis of a turbulent system, however, often requires sophisticated methods even if one is interested only in its long-time behaviour or on the time average of a certain quantity, such as aerodynamic losses or the heat transported by a convecting fluid. Asymptotic or average properties are the result of highly-nonlinear and seemingly chaotic processes that take place across a wide range of space and time scales. Consequently, an exact quantitative description of the average or asymptotic behaviour of a turbulent system is normally not available unless one analyses in detail its complex spatio-temporal evolution. To further complicate matters, the state of a system governed by PDEs is a function of space

and time, so the phase space is infinite-dimensional. This makes the mathematical analysis challenging: closed-form solutions of PDEs are unavailable in all but a few exceptional cases, and even proving that a unique solution exists at all times often demands advanced theoretical methods. For example, proving that the Navier–Stokes equations, which ostensibly describe the flow of incompressible fluids, possess a unique smooth solution given smooth initial conditions is a famous open problem (Carlson *et al.*, 2006). In addition, while it is sometimes possible to prove that the asymptotic dynamics are governed only by a finite number of determining modes (Robinson, 2001, chapters 13–16), these are rarely known explicitly and their number often remains large. All these factors not only hinder analytical progress, but also pose limits to the range of turbulent systems that can be studied through direct numerical simulation (DNS) due to the large computational resources required to fully resolve all relevant spatio-temporal scales.

One approach to bypass the difficulties outlined above is to look for rigorous bounds on the long-time or time-averaged properties of a turbulent system, rather than try to compute their value exactly. Bounds are attractive because, as will be explained below, they can be found without numerical integration of the system's turbulent dynamics. Consequently, they give useful quantitative results for systems that are currently beyond the reach of experiments or DNSs. In addition, one can derive bounds that apply to all trajectories in phase space, meaning that they are valid for all possible ways in which the system can evolve, independently of the initial condition. Finally, the mathematical analysis used to derive rigorous bounds proceeds directly from the equations of motion without any additional *ad-hoc* assumptions, such as the closure hypotheses used in many classical theories of hydrodynamic turbulence (Doering & Constantin, 1992). Consequently, one obtains rigorous results against which phenomenological theories and simplified models can be tested in order to validate or reject them.

In the context of turbulent incompressible fluid flows, the derivation of rigorous bounds on time averages was pioneered by Malkus, Howard, and Busse (see for example Malkus, 1954; Howard, 1963, 1972; Busse, 1970, 1979). These authors were inspired by the hypothesis that turbulence organises itself to maximise the transport of a certain quantity, such as heat in convective flows or momentum in shear flows (Malkus, 1954). This suggests the use of variational calculus to find the flow field that maximises the transport of the relevant quantity. Of course, considering flows that satisfy the full equations of motion leads to a variational problem just as difficult as solving the original PDEs. Nonetheless, progress can be made by carrying out the maximisation over a larger set of flow fields, which satisfy only a subset of integral constraints derived from the governing equations. Since any flow observed

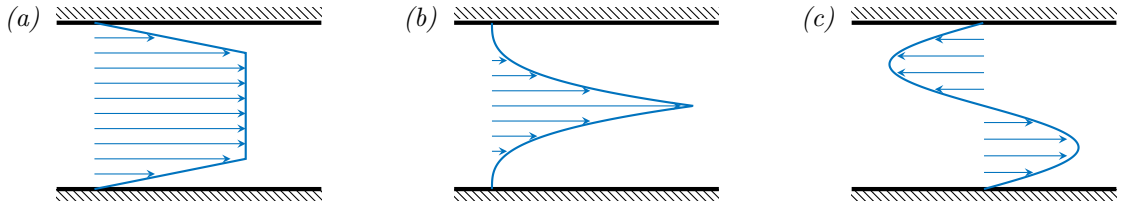


FIGURE 1.1: Possible background fields for a two-dimensional pressure-driven channel flow. *(a)* A piecewise-linear profile. *(b)* A piecewise-cubic profile. *(c)* A sinusoidal profile. In all panels, the background velocity is purely horizontal, varies only in the vertical direction, and vanishes at the walls. None of the profiles satisfies the equations governing the flow.

in reality must also satisfy these constraints, the maximum of this variational problem is a rigorous upper bound on the physically realised turbulent transport. One then hopes that enforcing only a few constraints suffices to obtain a bound that is quantitatively accurate and, most importantly, whose scaling with the relevant non-dimensional parameters (say, the Reynolds number in shear flows or the Rayleigh and Prandtl numbers in convection) is the same as for experimentally or numerically observed flows. However, finding the optimal flow fields typically requires the construction of a sophisticated hierarchy of nested boundary layers (the so-called *multi-α* solution method, see e.g. Busse, 1979) even when one only enforces incompressibility, a global power balance, and a suitably horizontal-and-time averaged version of the equations of motion.

The main limitation of the Malkus–Howard–Busse (MHB) approach to bounding time averages is that a rigorous bound is obtained only upon finding the optimal flow field among all admissible ones. This difficulty was overcome in the 1990s by Doering & Constantin, who introduced the so-called *background method* to bound the time average of flow functionals related to the volume-averaged energy dissipation (see for example Doering & Constantin, 1992, 1994, 1996; Constantin & Doering, 1995*a*,*b*). The background method relies on the decomposition of the flow variables into a steady *background field*, which satisfies the boundary conditions but is otherwise arbitrary, and a time-dependent *fluctuation field*, which satisfies homogeneous boundary conditions. Note that the background field need not solve the governing equations, nor need it correspond to the time average of any physically realised flow: the "unphysical" horizontal velocity fields sketched in figure 1.1, for example, are all valid background velocity fields for a two-dimensional pressure-driven channel flow. Once the decomposition into the background and fluctuation fields is introduced into the equations of motion, an argument similar to the "energy stability" analysis (see Straughan, 2004, for a good introduction to the subject) shows that the functional of interest can be bounded as a function of the background field, provided that a certain quadratic form that depends

on it is positive semidefinite. This condition is equivalent to the non-negativity of the spectrum of a certain linear operator that depends affinely on the background field and, for this reason, it has become widely known as the *spectral constraint*. The key observation is that, while one would like to optimise the bound over all background fields that satisfy the spectral constraint and the prescribed boundary conditions (hereafter referred to as *admissible* background fields), doing so is not required because the construction of any admissible background field yields a rigorous bound. In particular, one can consider a simple background field—typically, a piecewise-linear one like that shown in figure 1.1*(a)*—and use well-known functional estimates to show that the spectral constraint holds.

The fact that the solution of an infinite-dimensional variational problem is not required to produce a rigorous bound represents a great simplification compared to the MHB bounding approach. This was key to the success of the background method, which, since its first formulation, has been applied to bound time-averaged properties of a range of wall-bounded flows. Examples include: plane Couette flow (Doering & Constantin, 1992, 1994; Marchioro, 1994; Nicodemus *et al.*, 1997*a*); pressure-driven channel flow (Constantin & Doering, 1995*b*); stress-driven parallel shear flows (Tang *et al.*, 2004; Hagstrom & Doering, 2014); Rayleigh–Bénard convection with a variety of velocity and thermal boundary conditions (Doering & Constantin, 1996; Constantin & Doering, 1996; Otero *et al.*, 2002; Wittenberg & Gao, 2010; Wittenberg, 2010; Whitehead & Doering, 2011; Goluskin & Doering, 2016); Rayleigh–Bénard convection at infinite Prandtl number (Constantin & Doering, 1999; Doering & Constantin, 2001; Yan, 2004; Doering *et al.*, 2006; Otto & Seis, 2011; Whitehead & Doering, 2012; Whitehead & Wittenberg, 2014); Bénard–Marangoni convection (Hagstrom & Doering, 2010); convection in porous media (Doering & Constantin, 1998; Otero *et al.*, 2004); internally-heated convection (Whitehead & Doering, 2012; Goluskin, 2015, 2016).

It should also be remarked that, before Doering & Constantin's formulation of the background method, similar ideas had been applied to study the long-term dynamics of the Kuramoto–Sivashinky equation (Kuramoto & Tsuzuki, 1975, 1976; Sivashinsky, 1980). This is a nonlinear PDE with a hydrodynamic-type nonlinearity that arises, among others, when studying reaction-diffusion systems near instability (Kuramoto & Tsuzuki, 1975, 1976), the propagation of flame fronts (Sivashinsky, 1977, 1980; Michelson & Sivashinsky, 1977), and the evolution of surface waves in thin liquid films (Sivashinsky & Michelson, 1980). Moreover, it has become a paradigm for systems characterised by long-wavelength instability, *i.e.*, whose Fourier spectrum is unstable at small wavenumbers (Wittenberg, 2002). In this context, the decomposition into a background field and a perturbation was used to prove

asymptotic bounds for the total kinetic energy of the system (Nicolaenko *et al.*, 1985; Collet *et al.*, 1993; Goodman, 1994; Molinet, 2000; Bronski & Gambill, 2006).

The relation between the background method and the MHB approach to bounding time averages was elucidated by Kerswell (1998, 1999, 2001), who showed that the two techniques are dual. This means that the variational problems formulated with the background and MHB methods describe the same saddle point of a certain Lagrangian functional: the former from above, the latter from below (Kerswell, 1998, 1999, 2001; Plasting & Ierley, 2005). Consequently, the optimal bound available to the background method coincides with that obtained from the solution of the MHB variational problem. Duality also offers a useful interpretation of the background field: modulo rescaling, it corresponds to one of the Lagrange multipliers enforcing the dynamical constraints in the MHB method.

Despite the conceptual advantages offered by the background method compared to the MHB approach, the construction of a "good" background field remains a fundamental challenge. While simple profiles and relatively straightforward estimates of the spectral constraint often suffice to obtain non-trivial bounds, their asymptotic scaling with the system parameters may not capture that of the fully optimal result. The best example of this issue comes from the study of Rayleigh–Bénard convection—the buoyancy-driven motion of an incompressible fluid bounded by horizontal plates and heated from below—at infinite Prandtl number. In this case one seeks an upper bound on the Nusselt number Nu, the non-dimensional measure of the vertical heat transfer enhancement by convection, as a function of the non-dimensional thermal forcing described by the Rayleigh number Ra. The traditional background method analysis for this problem requires the construction of a Ra-dependent background temperature field, τ, which varies only in the vertical direction, matches the temperature of the top and bottom plates, and satisfies an appropriate spectral constraint. Both the solution of the dual MHB variational problem using asymptotic methods (Chan, 1971) and a careful numerical optimisation of the background field (Ierley *et al.*, 2006) suggest the optimal bound $Nu \lesssim Ra^{1/3}$ at large Ra. However, restricting attention to piecewise-linear background fields such as those sketched in figures 1.2 yields at best $Nu \lesssim Ra^{2/5}$ when the background field is constant in the bulk of the layer (Doering & Constantin, 2001; Otero, 2002), and $Nu \lesssim Ra^{7/20}$ when it is allowed to vary linearly (Plasting & Ierley, 2005).

This example highlights that optimisation of the background field is fundamental if bounds obtained with the background method are to provide accurate quantitative predictions in practice, or at least as accurate as the method allows. The solution of the variational problem for the optimal bound, however, is difficult to obtain, mostly due to

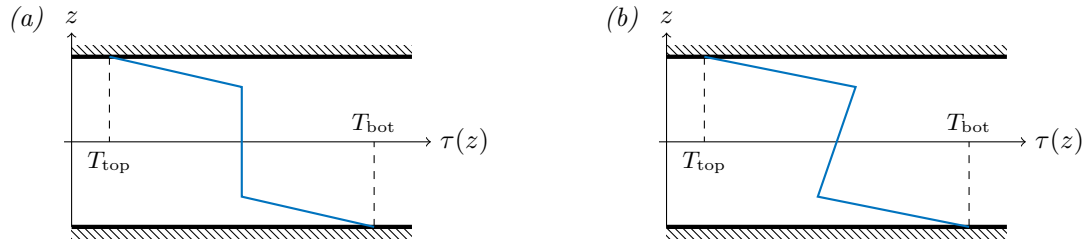


FIGURE 1.2: Piecewise-linear background temperature fields, denoted $\tau(z)$, for the background method analysis of infinite-Prandtl-number Rayleigh–Bénard convection. *(a)* Constant bulk profile, yielding $Nu \lesssim Ra^{2/5}$ (Doering & Constantin, 2001; Otero, 2002). *(b)* Linearly varying bulk profile, yielding $Nu \lesssim Ra^{7/20}$ (Plasting & Ierley, 2005). In both panels, z indicates the vertical direction, T_{top} denotes the temperature of the top plate, and T_{bot} is the (higher) temperature of the bottom plate.

the non-standard nature of the spectral constraint to be imposed on the background field. The classical approach is to consider the Euler–Lagrange (EL) equations for the optimal background and fluctuation fields, derived using the calculus of variations after enforcing the spectral constraint with the usual Lagrange multiplier technique (Doering & Constantin, 1996). An immediate obstacle is that the EL equations are nonlinear, so their solution is often beyond the reach of analytical work (but see Kerswell, 1997, for an example of how the duality between the background method and the MHB approach can be exploited to obtain asymptotic results). The second difficulty is that the EL equations admit multiple solutions, which in geometric terms correspond to different stationary points of the Lagrangian functional of the variational problem. However, all but one of these solutions are "spurious", in the sense that the corresponding background field does not actually satisfy the spectral constraint imposed by the background method, and therefore does not yield a valid bound.

To avoid convergence to such spurious solutions, the numerical optimisation of background fields via the EL equations has traditionally required extremely careful computations. Nicodemus *et al.* (1997*b*, 1998) extended preliminary results by Doering & Hyman (1997) for plane Couette flow using a sophisticated two-stage solution of the boundary-eigenvalue problem associated with the spectral constraint (recall that this is a positivity condition on the spectrum of a linear operator dependent on the background field). However, they considered a restricted form of background field, so their bound can only be consider "semi-optimal". Similar semi-optimal bounds have also been computed in the context of Rayleigh–Bénard convection by considering piecewise-linear profiles and implementing a simpler vanishing determinant technique (Otero *et al.*, 2004; Plasting & Ierley, 2005; Wittenberg & Gao, 2010). The full optimisation of the background field for plane Couette flow was carried out by Plasting & Kerswell (2003), who employed numerical continuation to track the correct solution of

the EL equations for increasing Reynolds numbers. This strategy was also utilised by Tang *et al.* (2004) to optimise bounds on the energy dissipation of stress-driven shear flows, after the imposed stress was approximated by a localised body force.

More recently, Wen *et al.* (2013, 2015) proposed a numerical method that does not rely on numerical continuation, and that appears to be robustly applicable to a wide range of problems. The idea is that, in order to solve the EL equations for the optimal background and fluctuation fields, one can consider all variables to be time-dependent, add suitable time derivatives to the EL equations, and evolve an initial guess for the solution forward in time until convergence to a steady state. This amounts to applying the gradient method to find a saddle point of the Lagrangian functional of the variational problem for the optimal background field, a strategy whose seeds can be found also in an attempt by Gambill (2006, chapter 6) to optimise background fields for the Kuramoto–Sivashinsky equation.[1] Such a "time-marching" approach is attractive because one can leverage established and computationally efficient numerical integration techniques, and its practical efficacy has been demonstrated in the context of two-dimensional porous-media convection at infinite Prandtl-Darcy number. Additionally, and perhaps more importantly, it seems that the time-marching method cannot converge to spurious solutions. Indeed, for three classical problems in fluid mechanics (two-dimensional porous-media convection, plane Couette flow, and two-dimensional Rayleigh–Bénard convection with stress-free isothermal boundaries) Wen *et al.* (2015) have proven that spurious solutions are linearly unstable equilibria of the time-dependent EL equations, and therefore cannot be attracting states.

Such proofs, however, apply only to particular optimal background field problems encountered in fluid dynamics, all of which share a similar structure. A general argument is yet unavailable, and whether the time-marching algorithm can be successfully utilised in other contexts remains an open question. To provide at least a partial answer, in this thesis the time-marching algorithm will be applied to optimise bounds on the asymptotic energy of solutions of the Kuramoto–Sivashinsky equation, derived using the background method. It will be demonstrated that when the EL equations for the optimal background field are formulated following the steps outlined by Wen *et al.* (2015), the time-marching procedure converges to spurious solutions. It will also be shown that a modification of the EL equations seems to resolve the issue in practice, enabling the robust computation of the optimal background field. However, a proof that spurious solutions are linearly unstable remains unavailable because the argument proposed by Wen *et al.* (2015) cannot be extended.

[1] Gambill (2006) did not solve the full EL equations for the Kuramoto–Sivashinsky problem, and only computed "semi-optimal" background fields. However, his numerical method is a time-dependent formulation of the gradient method, very similar to that proposed by Wen *et al.* (2013, 2015).

There are two other issues that deserve further attention. First, even when convergence of the time-marching method to spurious solutions can be rigorously excluded, one cannot guarantee that the desired steady-state solution will actually be reached. As already noted by Wen *et al.* (2015), for certain initial conditions the solution of the time-dependent EL equations may approach a periodic orbit or a chaotic attractor. This was not the case for the problems solved so far, but the lack of convergence to any steady state remains a possible issue in general. Second, there exist optimal background field problems for which the EL equations appear not to be solvable. The difficulty is that, while the convexity of the variational problem guarantees the existence of a unique optimal background field (Doering & Constantin, 1996), an optimal fluctuation field may not always be found. This problem is encountered, for example, when trying to optimise the background field for Bénard–Marangoni convection at infinite Prandtl number: an extra boundary condition arises when deriving the EL equation for the optimal fluctuation, which is therefore over-constrained in general (see chapter 6 for more details). For such optimal background field problems, all existing numerical methods—including the time-marching algorithm by Wen *et al.* (2015)—do not seem applicable, because they rely on a direct solution of the EL equations.

For the reasons just described, this thesis will investigate whether it is possible to optimise background fields using alternative computational techniques, which do not require the solution of the EL equations. Building on work by Fantuzzi & Wynn (2015), a new approach will be developed, based on the fact that a spectral constraint on the background field requires that a linear operator, affinely dependent on the background field, has non-negative eigenvalues. Thus, a spectral constraint can be seen as the infinite-dimensional equivalent of a *linear matrix inequality* (LMI), a positive semidefinitess condition on a matrix whose entries depend affinely on a set of optimisation variables (see chapter 2 for a precise definition). Optimisation problems with LMIs, known as *semidefinite programmes* (SDPs), are well known in the optimisation community and efficient algorithms for their solution (with proven convergence guarantees) exist. In addition, finely tuned implementations are available in open-source software packages such as SeDuMi (Sturm, 1999, 2002), SDPT3 (Toh *et al.*, 1999; Tütüncü *et al.*, 2003), and Mosek (Andersen *et al.*, 2009). This suggests a "discretise-then-optimise" strategy, according to which the infinite-dimensional variational problem for the optimal background field is first recast as a finite-dimensional SDP, and then solved using general-purpose optimisation algorithms. Such a line of work contrasts the traditional approach based on the derivation and numerical solution of the EL equations, which can instead be interpreted as an "optimise-then-discretise" strategy.

One final open problem is how, if at all, numerical optimisation of the background fields can be used to inform rigorous analysis. In fact, one is often interested in bounding the properties of turbulent systems at asymptotically large values of the governing parameters (say, as the Reynolds number of a certain turbulent flow tends to infinity). Numerically optimal bounds at fixed, large parameter values can give a strong indication of the best possible asymptotic behaviour, but one would like to construct *analytically* a background field that achieves it. When piecewise-linear or similarly simple profiles do not suffice, background fields modelled on the numerically optimal ones may be used. However, considerable ingenuity is often required to identify which features of the optimal background fields affect the asymptotic behaviour of the corresponding bounds, and which ones do not. One way to assist the analysis is to optimise the bounds numerically, but only over restricted classes of background fields. For instance, if the fully optimal background fields are non-monotonic, then optimising bounds only over monotonic background fields can reveal if non-monotonicity is important. One benefit of the SDP-based numerical methods developed in this thesis is that extra constraints can be imposed easily, provided that they admit an LMI representation. The extent to which this flexibility can be taken advantage of in order to aid the analytical construction of near-optimal background fields will be examined in chapter 6, in the context of Bénard–Marangoni convection at infinite Prandtl number.

1.1 Outline of the thesis

The general mathematical notation used in this work is defined in section 1.2 below, although additional notation is also introduced as needed throughout the thesis.

Chapter 2 gives a brief overview of Lagrangian duality for finite-dimensional linear optimisation problems, of linear matrix inequalities, and of semidefinite programmes. This chapter is not meant as an extensive review of semidefinite programming, but rather as a short introduction for non-expert readers. For this reason, only notions that are essential for this thesis are covered, while algorithms for the solution of SDPs are not described.

The time-marching algorithm of Wen *et al.* (2015) is applied to bound the asymptotic energy of the solution of the Kuramoto–Sivashinsky equation in chapter 3. Numerical results demonstrate that convergence to spurious solutions can occur, but also that a small modification of the method appears to remove this problem. This modification is suggested by the interpretation of the relevant spectral constraints as an infinite-dimensional LMI, so chapter 3 serves as further motivation for the development of SDP-based methods.

Chapter 4 develops strategies based on semidefinite programming that can be used to solve variational problems arising from the background method analysis. For many classical background method problems, a multi-dimensional spectral constraint can be transformed into a set of one-dimensional spectral constraints upon consideration of Fourier expansions in all but one spatial direction. Each of these Fourier-transformed spectral constraints requires the positivity of an integral quadratic form on a space of functions. Consequently, chapter 4 focusses on optimisation problems subject to a particular class of one-dimensional integral inequality constraints, which encompasses many Fourier-transformed spectral constraints.

The methods developed in chapter 4 are subsequently applied to bound the energy dissipation in stress-driven shear flows (chapter 5) and the convective heat transport in Bénard–Marangoni convection at infinite Prandtl number (chapter 6). Chapter 6 also demonstrates that semidefinite programming allows for the identification of key properties of the optimal background field, to be exploited in rigorous analysis. Other advantages and limitations of optimisation methods based on SDPs are discussed throughout chapters 4–6.

Finally, chapter 7 summarises the main findings of this work, highlights remaining open questions, and outlines challenges to be addressed by future research.

1.2 Notation

As usual, $\mathbb{N}^n$ is the set of non-negative n-dimensional multi-indices (non-negative integer n-tuples), $\mathbb{R}^n$ is the n-dimensional Euclidean space, $\mathbb{C}^n$ is the $2n$-dimensional space of complex-valued n-component vectors, $\mathbb{R}^{m\times n}$ is the space of $m \times n$ matrices, and $\mathbb{S}^n$ is the space of $n \times n$ real-valued symmetric matrices. The real and imaginary parts of a complex-valued quantity q (either a complex number or a complex-valued function) are denoted by $\mathrm{Re}(q)$ and $\mathrm{Im}(q)$, respectively, while q^* is the complex conjugate of q.

Vectors and matrices are denoted, respectively, by lower- and upper-case boldface characters, e.g., a vector $\boldsymbol{v} \in \mathbb{R}^n$ and a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Exceptions to this rule are the zero vector and the zero matrix, both denoted by $\mathbf{0}$. The identity matrix is denoted by $\boldsymbol{I}$, and $\mathbf{1}$ is the vector of ones. For clarity, the size of the zero and identity matrices is sometimes indicated by subscripts, e.g., $\mathbf{0}_{m\times n}$ or $\boldsymbol{I}_n$. All vectors should be understood as column vectors unless otherwise stated. For any vector or matrix, the superscript $\intercal$ denotes transposition.

The usual ℓ^p ($1 \le p < \infty$) and ℓ^∞ norms of a vector $\boldsymbol{v} \in \mathbb{K}^n$ ($\mathbb{K} \equiv \mathbb{R}$ or $\mathbb{K} \equiv \mathbb{C}$) are

$$\|\boldsymbol{v}\|_p := \left(\sum_{i=1}^{n} |v_i|^p\right)^{\frac{1}{p}}, \qquad \|\boldsymbol{v}\|_\infty := \max_{1\le i\le n} |v_i|\,,$$

and the shorthand notation $\|\boldsymbol{v}\| \equiv \|\boldsymbol{v}\|_2$ is also used for the standard Euclidean norm.

The range, null space, trace, and rank of a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ are denoted by $\mathrm{img}(\boldsymbol{A})$, $\ker(\boldsymbol{A})$, $\mathrm{tr}(\boldsymbol{A})$, and $\mathrm{rank}(\boldsymbol{A})$, respectively. The Frobenius inner product of matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$, denoted by $\langle \boldsymbol{A}, \boldsymbol{B}\rangle$, is defined as $\langle \boldsymbol{A}, \boldsymbol{B}\rangle := \mathrm{tr}(\boldsymbol{A}^\mathsf{T}\boldsymbol{B})$. Accordingly, the Frobenius norm of $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is

$$\|\boldsymbol{A}\|_\mathrm{F} := \langle \boldsymbol{A}, \boldsymbol{A}\rangle = \left(\sum_{i=1}^{n}\sum_{j=1}^{m} |A_{i,j}|^2\right)^{\frac{1}{2}}.$$

Given vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$, the inequalities $\boldsymbol{u} \geq \boldsymbol{v}$ and $\boldsymbol{u} > \boldsymbol{v}$ should be interpreted element-wise, *i.e.*, $u_i \geq v_i$ and $u_i > v_i$ for $i = 1, \ldots, n$. For a matrix $\boldsymbol{A} \in \mathbb{S}^n$, the notation $\boldsymbol{A} \succeq 0$ (resp. $\boldsymbol{A} \succ 0$) means that $\boldsymbol{A}$ is positive semidefinite (resp. positive definite), *i.e.*, all its eigenvalues are non-negative (resp. positive) or, equivalently, $\boldsymbol{v}^\mathsf{T}\boldsymbol{A}\boldsymbol{v} \geq 0$ (resp. > 0) for all $\boldsymbol{v} \in \mathbb{R}^n$. In addition, the inequalities $\boldsymbol{A} \succeq \boldsymbol{B}$ and $\boldsymbol{A} \succ \boldsymbol{B}$ should be interpreted, respectively, as $\boldsymbol{A} - \boldsymbol{B} \succeq 0$ and $\boldsymbol{A} - \boldsymbol{B} \succ 0$.

Given an open bounded set $\Omega \subset \mathbb{R}^n$, let the compact set $\bar{\Omega}$ be its closure and, with the shorthand notation $\mathrm{d}^n\boldsymbol{x} := \mathrm{d}x_1 \cdots \mathrm{d}x_n$, denote its n-dimensional volume by

$$|\Omega| := \int_\Omega \mathrm{d}^n\boldsymbol{x}.$$

For a positive integer q, $C^m(\bar{\Omega}, \mathbb{R}^q)$ is the space of m-times continuously differentiable functions mapping $\bar{\Omega}$ to $\mathbb{R}^q$; the shorthand notation $C^m(\bar{\Omega})$ is also used instead of $C^m(\bar{\Omega}, \mathbb{R})$. Moreover, given $1 \leq p < \infty$, $L^p(\Omega, \mathbb{R}^q)$ is the usual Lebesgue space of p-integrable vector-valued functions, while $L^\infty(\Omega, \mathbb{R}^q)$ is the space of essentially bounded functions. The standard Lebesgue norms of $\boldsymbol{f} \in L^p(\Omega, \mathbb{R}^q)$ and $\boldsymbol{g} \in L^\infty(\Omega, \mathbb{R}^q)$, denoted by $\|\boldsymbol{f}\|_p$ and $\|\boldsymbol{g}\|_\infty$, are

$$\|\boldsymbol{f}\|_p := \left[\int_\Omega \sum_{i=1}^{n} |f_i(\boldsymbol{x})|^p \, \mathrm{d}^n\boldsymbol{x}\right]^{\frac{1}{p}}, \qquad \|\boldsymbol{g}\|_\infty := \max_{i=1,\ldots,n}\left[\operatorname*{ess\,sup}_{\boldsymbol{x}\in\Omega} g_i(\boldsymbol{x})\right].$$

Recall that the essential supremum, ess sup, is the smallest supremum over all sets $\Omega \setminus Z$ with Z a zero-measure set (Robinson, 2001, section 1.4.3).

For $\alpha \geq 1$, the α-th derivative of a function $f(x)$ is denoted by $\partial^\alpha f$. When $\alpha = 1$ and $\alpha = 2$, however, primes (e.g., f') are used where convenient to lighten the notation. When differentiating a multi-variable functions α times with respect to one of its variables, the differentiation variable is specified explicitly, for instance $\partial_x^\alpha f := \frac{\partial^\alpha f}{\partial x^\alpha}$.

Let $u = u(t, \boldsymbol{x}, z)$ be a real-valued function of a time variable $t \in [0, +\infty)$, a horizontal coordinate vector $\boldsymbol{x} \in \Omega \subset \mathbb{R}^d$, and a vertical coordinate $z \in (a, b)$ with (a, b) bounded. The

overline notation $\overline{u}$ denotes the horizontal and infinite-time average of u, while $\langle u \rangle$ denotes its volume and infinite-time average. More precisely,

$$\overline{u}(z) := \limsup_{T \to +\infty} \frac{1}{|\Omega|\, T} \int_0^T \int_\Omega u(t, \boldsymbol{x}, z)\, \mathrm{d}^d \boldsymbol{x}\, \mathrm{d}t,$$
$$\langle u \rangle := \limsup_{T \to +\infty} \frac{1}{(b-a)\, |\Omega|\, T} \int_0^T \int_a^b \int_\Omega u(t, \boldsymbol{x}, z)\, \mathrm{d}^d \boldsymbol{x}\, \mathrm{d}z\, \mathrm{d}t.$$

Given functions $f(x)$ and $g(x)$ with $x > 0$, the notation $f \sim g$ indicates that f and g are asymptotically equivalent up to a positive constant, meaning that there exists a constant $C > 0$ such that

$$\lim_{x \to +\infty} \frac{f(x)}{g(x)} = C.$$

Similarly, the notation $f \lesssim g$ indicates that f is asymptotically bounded by g up to a positive constant, meaning that there exists a constant $C > 0$ such that

$$\lim_{x \to +\infty} \frac{f(x)}{g(x)} \leq C.$$

Finally, functionals are denoted by calligraphic letters and their arguments are specified inside curly braces. Thus, $\mathcal{F}\{u\}$ denotes the functional $\mathcal{F}$ acting on u. The functionals encountered in this work are defined on affine spaces of functions $u : \Omega \to \mathbb{R}^n$ that satisfy linear boundary conditions (homogeneous, inhomogeneous, or periodic) on the boundary of Ω. Let such an affine function space be denoted by $U = \{u = u_0 + v, v \in V\}$ where u_0 is any function satisfying the prescribed boundary conditions and V is the linear space defined by the homogeneous version of the boundary condition. Then, $u + hv \in U$ for all $u \in U$, $h \in \mathbb{R}$, and $v \in V$, so $\mathcal{F}\{u + hv\}$ is well defined. Given $u \in U$, the variation (or functional derivative) of $\mathcal{F}$ at u is the functional $\frac{\delta \mathcal{F}}{\delta u} : V \to \mathbb{R}$ such that

$$\frac{\delta \mathcal{F}}{\delta u}\{v\} := \lim_{h \to 0} \frac{\mathcal{F}\{u + hv\} - \mathcal{F}\{u\}}{h},$$

provided that the limit exists and is finite. If V is a Hilbert space and $\frac{\delta \mathcal{F}}{\delta u}$ is a bounded linear functional, then $\frac{\delta \mathcal{F}}{\delta u}$ can be identified with an element of V by virtue of the Riesz representation theorem (see for instance Zeidler, 1995, section 2.10). In such cases, the notation $\frac{\delta \mathcal{F}}{\delta u}$ is used also to indicate this representing element.

Chapter 2

Lagrangian duality, linear matrix inequalities and semidefinite programming: a review

Linear matrix inequalities and linear optimisation problems with linear matrix inequalities, known as semidefinite programmes, are well known in the optimisation and control communities. These concepts, together with the idea of Lagrangian duality, play a major role in this thesis, so they are briefly reviewed here. The purpose of this chapter is to give a self-contained introduction to material that will be utilised throughout chapters 3–6. For an in-depth treatment of Lagrangian duality, linear matrix inequalities, and semidefinite programming the reader is referred to the excellent works by Boyd *et al.* (1994), Vandenberghe & Boyd (1996), Boyd & Vandenberghe (2004), and Parrilo (2013).

2.1 Lagrangian duality

This section offers a brief introduction to Lagrangian duality for inequality-constrained optimisation problems. The material is adapted from Boyd & Vandenberghe (2004, chapter 5), and we refer the interested reader to their work for a detailed treatment of the subject.

Let $\boldsymbol{x} = (x_i)_{i=1}^n \in \mathbb{R}^n$ be an optimisation variable, let $f_0, \ldots, f_m : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ be given functions, and consider the optimisation problem (referred to as the *primal problem*)

$$\begin{aligned} \min_{\boldsymbol{x}\in\Omega} \quad & f_0(\boldsymbol{x}), \\ \text{s.t.} \quad & f_i(\boldsymbol{x}) \geq 0, \quad i = 1, \ldots, m. \end{aligned} \tag{2.1}$$

Assume that the problem is feasible, meaning that there exists a *feasible point* $\boldsymbol{x} \in \Omega$ that satisfies the inequality constraints, that there exists an *optimal point* $\boldsymbol{x}^\star$, which minimises $f_0(\boldsymbol{x})$ subject to the constraints, and that the optimal value $p^\star := f_0(\boldsymbol{x}^\star)$ is finite.

Lagrangian duality provides a way to study (2.1) by augmenting the objective function using the constraints. Given $\boldsymbol{y} = (y_i)_{i=1}^m \in \mathbb{R}^m$, the *Lagrangian function* (or simply *Lagrangian*) for (2.1) is

$$L(\boldsymbol{x}, \boldsymbol{y}) := f_0(\boldsymbol{x}) - \sum_{i=1}^{m} y_i \, f_i(\boldsymbol{x}), \tag{2.2}$$

and the *Lagrange dual function*, or simply *dual function*, is defined as

$$g(\boldsymbol{y}) := \inf_{\boldsymbol{x} \in \Omega} L(\boldsymbol{x}, \boldsymbol{y}). \tag{2.3}$$

The vector $\boldsymbol{y}$ is known as the *dual variable*, and each entry y_i is known as the *Lagrange multiplier* for the inequality constraint $f_i(\boldsymbol{x}) \geq 0$.

The dual function is useful to derive lower bounds on the optimal value $p^\star$ of (2.1). In fact, for any $\boldsymbol{y} \geq \boldsymbol{0}$ one has

$$g(\boldsymbol{y}) = \inf_{\boldsymbol{x} \in \Omega} L(\boldsymbol{x}, \boldsymbol{y}) \leq L(\boldsymbol{x}^\star, \boldsymbol{y}) = f_0(\boldsymbol{x}^\star) - \sum_{i=1}^{m} y_i \, f_i(\boldsymbol{x}^\star) \leq f_0(\boldsymbol{x}^\star) = p^\star, \tag{2.4}$$

where the last inequality holds because $\boldsymbol{y} \geq \boldsymbol{0}$ and $f_i(\boldsymbol{x}^\star) \geq 0$. In particular, the best lower bound on $p^\star$ is found upon solving the so-called *dual problem* associated with (2.1),

$$\begin{aligned} \sup_{\boldsymbol{y}} \quad & g(\boldsymbol{y}) \\ \text{s.t.} \quad & \boldsymbol{y} \geq \boldsymbol{0}. \end{aligned} \tag{2.5}$$

Let $d^\star$ denote the optimal value of the dual problem (2.5). The property that $d^\star \leq p^\star$ is known as *weak duality*, and it holds for a general problem. If the equality $d^\star = p^\star$ is satisfied, then the dual problem gives a sharp bound on $p^\star$ and one says that *strong duality* holds. Strong duality does not always hold, but if the primal problem (2.1) is convex (meaning that f_0, $-f_1$, …, $-f_m$ are convex functions) then a sufficient condition for strong duality, known as *Slater's condition*, can be established (Boyd & Vandenberghe, 2004, section 5.2.3). The following definitions are required.

Definition 2.1. The *affine hull* of a set $\Omega \in \mathbb{R}^n$, denoted $\operatorname{aff}(\Omega)$, is the affine space spanned by Ω, *i.e.*,

$$\operatorname{aff}(\Omega) := \left\{ \boldsymbol{x} \in \mathbb{R}^n \,:\, \exists k \in \mathbb{N},\ \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \Omega,\ \theta_1, \ldots, \theta_k \in \mathbb{R} \right.$$
$$\left. \text{such that } \sum_{i=1}^{k} \theta_i = 1 \text{ and } \boldsymbol{x} = \sum_{i=1}^{k} \theta_i \boldsymbol{x}_i \right\}.$$

Definition 2.2. For $\boldsymbol{x} \in \mathbb{R}^n$, let $B(\boldsymbol{x}, r) := \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| < r\}$ be the open ball centred at $\boldsymbol{x}$ with radius r. The *relative interior* of a set $\Omega \in \mathbb{R}^n$, denoted $\operatorname{relint}(\Omega)$, is the set

$$\operatorname{relint}(\Omega) := \{\boldsymbol{x} \in \Omega : B(\boldsymbol{x}, r) \cap \operatorname{aff}(\Omega) \subset \Omega \text{ for some } r > 0\}.$$

In other words, the relative interior of a set $\Omega \in \mathbb{R}^n$ is the interior of Ω when viewed as a subset of the affine hull $\operatorname{aff}(\Omega)$.

Proposition 2.1 (Slater's condition, Boyd & Vandenberghe, 2004, section 5.3.2)**.** *Assume that the optimisation problem* (2.1) *is convex, meaning that* $f_0, -f_1, \ldots, -f_m$ *are convex functions. If there exists* $\boldsymbol{x}_0 \in \operatorname{relint}(\Omega)$ *such that* $f_i(\boldsymbol{x}_0) > 0$ *for all* $i = 1, \ldots, m$*, then strong duality holds. Furthermore, there exists a dual variable* $\boldsymbol{y}^\star$ *such that* $\boldsymbol{y}^\star \geq \boldsymbol{0}$ *and* $g(\boldsymbol{y}^\star) = d^\star$*, i.e., the dual problem* (2.5) *attains its optimal value.*

To conclude this section, suppose that strong duality holds for (2.1). Let $\boldsymbol{x}^\star$ and $\boldsymbol{y}^\star$ be the optimal solutions of the primal problem (2.1) and the dual problem (2.5), respectively. Then, the pair $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ is a saddle point for the Lagrangian $L(\boldsymbol{x}, \boldsymbol{y})$ (Boyd & Vandenberghe, 2004, section 5.4), in the sense that for all $\boldsymbol{x} \in \Omega$ and all $\boldsymbol{y} \geq 0$

$$L(\boldsymbol{x}^\star, \boldsymbol{y}) \leq L(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \leq L(\boldsymbol{x}, \boldsymbol{y}^\star). \tag{2.6}$$

Consequently, an optimal solution of (2.1) can be computed by finding a saddle point of $L(\boldsymbol{x}, \boldsymbol{y})$, and in particular by solving the max-min or min-max problems

$$p^\star = \sup_{\boldsymbol{y} \geq \boldsymbol{0}} \inf_{\boldsymbol{x} \in \Omega} L(\boldsymbol{x}, \boldsymbol{y}), \qquad\qquad p^\star = \inf_{\boldsymbol{x} \in \Omega} \sup_{\boldsymbol{y} \geq \boldsymbol{0}} L(\boldsymbol{x}, \boldsymbol{y}). \tag{2.7}$$

2.2 Linear matrix inequalities

Let $\boldsymbol{y} = (y_i)_{i=1}^m \in \mathbb{R}^m$ be an optimisation variable and $\boldsymbol{F}_i \in \mathbb{S}^n$, $i = 0, \ldots, m$ be given symmetric matrices. A linear matrix inequality (LMI) is a constraint of the form

$$\boldsymbol{F}(\boldsymbol{y}) := \boldsymbol{F}_0 - \sum_{i=1}^m y_i \boldsymbol{F}_i \succeq 0. \tag{2.8}$$

In other words, an LMI is a positive semidefiniteness condition on a symmetric matrix $\boldsymbol{F}(\boldsymbol{y})$, whose entries depend affinely on $\boldsymbol{y}$ (any such matrix can be rewritten in the above form). It is not difficult to check that (2.8) is a convex constraint on $\boldsymbol{y}$, meaning that its *feasible set* $S := \{\boldsymbol{y} \in \mathbb{R}^m : \boldsymbol{F}(\boldsymbol{y}) \succeq 0\}$ is convex. Indeed, if two vectors $\boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^m$ satisfy (2.8), then for any $\theta \in [0, 1]$ so does their convex combination $\theta\boldsymbol{y} + (1 - \theta)\boldsymbol{z}$.

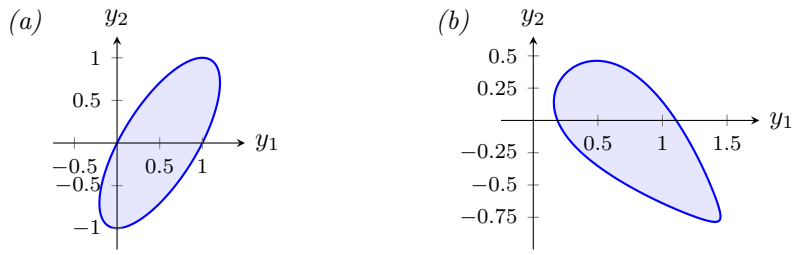


FIGURE 2.1: *(a)* Feasible set of the LMI in example 2.1, enclosed within the ellipse $2y_1^2 - 2y_1y_2 + y_2^2 - 2y_1 + y_2 = 0$. *(b)* Feasible set of the LMI in example 2.2, defined by the polynomial inequalities $p_0(\boldsymbol{y}) \geq 0$, $p_1(\boldsymbol{y}) \geq 0$, $p_2(\boldsymbol{y}) \geq 0$, $p_3(\boldsymbol{y}) \geq 0$, and $p_4(\boldsymbol{y}) \geq 0$.

Example 2.1. Let $\boldsymbol{y} \in \mathbb{R}^2$ and consider the 2×2 LMI

$$\boldsymbol{F}(\boldsymbol{y}) := \begin{bmatrix} 2y_1 - y_2 & y_2 \\ y_2 & 2 - 2y_1 + y_2 \end{bmatrix} \succeq 0.$$

This LMI can be rewritten in the form (2.8) with

$$\boldsymbol{F}_0 = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}, \qquad \boldsymbol{F}_1 = \begin{bmatrix} -2 & 0 \\ 0 & 2 \end{bmatrix}, \qquad \boldsymbol{F}_2 = \begin{bmatrix} 1 & -1 \\ -1 & -1 \end{bmatrix}.$$

The feasible set of this LMI can be found analytically by requiring that the eigenvalues of $\boldsymbol{F}(\boldsymbol{y})$ are non-negative, which is true when $2y_1^2 - 2y_1y_2 + y_2^2 - 2y_1 + y_2 \leq 0$. This polynomial inequality defines the ellipse plotted in figure 2.1*(a)*.

Example 2.2. Let $\boldsymbol{y} \in \mathbb{R}^2$ and consider the 5×5 LMI

$$\boldsymbol{F}(\boldsymbol{y}) = \begin{bmatrix} 3y_1 & y_2 - y_1 & 2y_2 & y_2 - 1 & 0 \\ y_2 - y_1 & 5 - y_2 & -y_2 & y_1 & 0 \\ 2y_2 & -y_2 & 2 - y_1 & y_1 + y_2 & 0 \\ y_2 - 1 & y_1 & y_1 + y_2 & 2 + y_2 & -1 \\ 0 & 0 & 0 & -1 & 5 \end{bmatrix} \succeq 0.$$

This LMI can be rewritten in the form (2.8) with

$$\boldsymbol{F}_0 = \begin{bmatrix} 0 & 0 & 0 & -1 & 0 \\ 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ -1 & 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & -1 & 5 \end{bmatrix}, \boldsymbol{F}_1 = \begin{bmatrix} -3 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \boldsymbol{F}_2 = \begin{bmatrix} 0 & -1 & -2 & -1 & 0 \\ -1 & 1 & 1 & 0 & 0 \\ -2 & 1 & 0 & -1 & 0 \\ -1 & 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

As in the previous example, the feasible set of the LMI can be found by requiring that all eigenvalues of $\boldsymbol{F}(\boldsymbol{y})$ are non-negative. The eigenvalues are the roots of the characteristic polynomial $p(t) = \det[\boldsymbol{F}(\boldsymbol{y}) - t\boldsymbol{I}] = t^5 - p_4(\boldsymbol{y})t^4 + p_3(\boldsymbol{y})t^3 - p_2(\boldsymbol{y})t^2 + p_1(\boldsymbol{y})t - p_0(\boldsymbol{y})$, where

$$\begin{aligned}
p_0(\boldsymbol{y}) &:= 20y_1^4 + 20y_1^3y_2 - 106y_1^3 - 5y_1^2y_2^2 - 226y_1^2y_2 - 133y_1^2 + 88y_1y_2^2 - 43y_1y_2 \\
&\quad + 295y_1 - 363y_2^2 + 110y_2 - 50, \\
p_1(\boldsymbol{y}) &:= 4y_1^4 + 4y_1^3y_2 - 41y_1^3 - y_1^2y_2^2 - 105y_1^2y_2 - 176y_1^2 + 48y_1y_2^2 - 29y_1y_2 + 364y_1 \\
&\quad - 341y_2^2 + 129y_2 + 45, \\
p_2(\boldsymbol{y}) &:= -4y_1^3 - 12y_1^2y_2 - 60y_1^2 + 6y_1y_2^2 - 4y_1y_2 + 161y_1 - 99y_2^2 + 47y_2 + 121, \\
p_3(\boldsymbol{y}) &:= -6y_1^2 + 30y_1 - 9y_2^2 + 5y_2 + 67, \\
p_4(\boldsymbol{y}) &:= 2y_1 + 14.
\end{aligned}$$

Consequently, by Descartes' rule of signs, the 5×5 LMI above is feasible if $\boldsymbol{y}$ belongs to the region of the $\mathbb{R}^2$ plane defined by the polynomial inequalities $p_i(\boldsymbol{y}) \geq 0$, $i = 1, \ldots, 4$. As illustrated in figure 2.1*(b)*, this region is convex.

2.3 LMI-representable constraints

Linear matrix inequalities can be used to represent many types of constraints typically encountered in optimisation. This section reviews some LMI-representable constraints that will be encountered in chapters 4–6. Throughout, $\boldsymbol{y} \in \mathbb{R}^m$ denotes the optimisation variable.

Linear inequalities. Let $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{b} \in \mathbb{R}^n$ be given. The n linear inequalities $\boldsymbol{Ay} \geq \boldsymbol{b}$ are clearly equivalent to the diagonal LMI

$$\begin{bmatrix}
\sum_{j=1}^m y_j A_{1,j} - b_1 & 0 & \cdots & 0 \\
0 & \sum_{j=1}^m y_j A_{2,j} - b_2 & & \vdots \\
\vdots & & \ddots & 0 \\
0 & \cdots & 0 & \sum_{j=1}^m y_j A_{n,j} - b_n
\end{bmatrix} \succeq 0. \tag{2.9}$$

Convex quadratic constraints. Let $\boldsymbol{A} \in \mathbb{S}^m$ be a given positive definite matrix. Let $\boldsymbol{b} \in \mathbb{R}^m$ and $c \in \mathbb{R}$ also be given. Since $\boldsymbol{A} \succ 0$ by assumption, Schur's complement condition (Boyd & Vandenberghe, 2004, appendix A.5.5) implies that the convex quadratic

constraint $\boldsymbol{y}^\mathsf{T}\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b}^\mathsf{T}\boldsymbol{y} - c \leq 0$ is equivalent to the LMI

$$\begin{bmatrix} \boldsymbol{A}^{-1} & \boldsymbol{y} \\ \boldsymbol{y}^\mathsf{T} & \boldsymbol{b}^\mathsf{T}\boldsymbol{y} + c \end{bmatrix} \succeq 0. \tag{2.10}$$

Second-order cone constraints. Let $\boldsymbol{A} \in \mathbb{R}^{n\times m}$, $\boldsymbol{b} \in \mathbb{R}^n$, $\boldsymbol{c} \in \mathbb{R}^m$ and $d \in \mathbb{R}$ be given. The convex constraint $\|\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}\| \leq \boldsymbol{c}^\mathsf{T}\boldsymbol{y} + d$ is known as a second-order cone constraints (SOCC). Upon squaring both sides of the inequality, the SOCC can be posed as an LMI because, by Schur's complement condition (Boyd & Vandenberghe, 2004, appendix A.5.5), one has

$$\|\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}\| \leq \boldsymbol{c}^\mathsf{T}\boldsymbol{y} + d \quad \Leftrightarrow \quad \begin{bmatrix} (\boldsymbol{c}^\mathsf{T}\boldsymbol{y} + d)\boldsymbol{I} & \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b} \\ (\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b})^\mathsf{T} & \boldsymbol{c}^\mathsf{T}\boldsymbol{y} + d \end{bmatrix} \succeq 0. \tag{2.11}$$

Polynomial sum-of-squares constraints. Let $\boldsymbol{x} \in \mathbb{R}^n$ and let $p(\boldsymbol{x})$ be a polynomial of degree $2d$, whose coefficients depend affinely on the optimisation variable $\boldsymbol{y} \in \mathbb{R}^m$. A sufficient condition for $p(\boldsymbol{x})$ to be non-negative for all $\boldsymbol{x}$ is that it can be written as a sum of squares of polynomials of degree no larger than d. The set of vectors $\boldsymbol{y}$ for which p admits such a sum-of-squares (SOS) decomposition can be represented using an LMI. Indeed, Parrilo (2003) demonstrated that $p(\boldsymbol{x})$ admits an SOS decomposition if and only if there exists a positive semidefinite matrix $\boldsymbol{Q} \in \mathbb{S}^s$ with $s := \binom{n+d}{d}$ such that

$$p(\boldsymbol{x}) = \boldsymbol{m}(\boldsymbol{x})^\mathsf{T}\boldsymbol{Q}\boldsymbol{m}(\boldsymbol{x}), \tag{2.12}$$

where $\boldsymbol{m}(\boldsymbol{x})$ is the vector of monomials of the form $x_1^{\alpha_1}x_2^{\alpha_2}\cdots x_n^{\alpha_n}$ of degree no larger than d. Comparing coefficients on both sides of (2.12) yields a set of equalities involving $\boldsymbol{y}$ and $\boldsymbol{Q}$, which can be used to express some of the entries of $\boldsymbol{Q}$ in terms of $\boldsymbol{y}$. Then, since $p(\boldsymbol{x})$ depends affinely on $\boldsymbol{y}$ by assumption, the requirement that $\boldsymbol{Q}$ is positive semidefinite becomes an LMI for $\boldsymbol{y}$ and any entries of $\boldsymbol{Q}$ that have not been eliminated. Consequently, one can optimise over SOS polynomials using LMIs.

Example 2.3. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2$ and consider the parametric polynomial

$$p(\boldsymbol{x}) = (y_1 + y_2) + 2y_1x_1 - 2y_2x_1x_2 + x_1^2 + (2 - y_2)x_2^2.$$

With $\boldsymbol{m}(\boldsymbol{x}) = [1, x_1, x_2]^\mathsf{T}$ and $\boldsymbol{Q} \in \mathbb{S}^3$, comparing coefficients on both sides of (2.12) yields

$$Q_{1,1} = y_1 + y_2, \quad Q_{1,2} = y_1 \quad Q_{1,3} = 0, \quad Q_{2,2} = 1, \quad Q_{2,3} = -y_2 \quad Q_{3,3} = 2 - y_2.$$

Consequently, for given values of the parameters y_1 and y_2, the polynomial $p(\boldsymbol{x})$ admits an SOS decomposition if

$$\begin{bmatrix} y_1 + y_2 & y_1 & 0 \\ y_1 & 1 & -y_2 \\ 0 & -y_2 & 2 - y_2 \end{bmatrix} \succeq 0.$$

Example 2.4. Let $x \in \mathbb{R}$, $\boldsymbol{y} \in \mathbb{R}^2$ and consider the parametric polynomial

$$p(x) = 1 - y_2 + 2y_1 x + (2 + y_2)x^2 + x^4.$$

With $\boldsymbol{m}(\boldsymbol{x}) = [1, x, x^2]^\mathsf{T}$ and $\boldsymbol{Q} \in \mathbb{S}^3$, comparing coefficients on both sides of (2.12) yields

$$Q_{1,1} = 1 - y_2, \qquad Q_{1,2} = y_1 \qquad Q_{2,2} + 2Q_{1,3} = 2 + y_2, \qquad Q_{2,3} = 0 \qquad Q_{3,3} = 1.$$

At this stage, one can choose which variables to eliminate using these equality constraints. For instance, upon rewriting all entries of $\boldsymbol{Q}$ in terms of y_1, y_2, and $Q_{1,3}$ one concludes that the polynomial $p(x)$ admits an SOS decomposition for given y_1 and y_2 if there exists $Q_{1,3}$ such that

$$\boldsymbol{F}_1(y_1, y_2, Q_{1,3}) := \begin{bmatrix} 1 - y_2 & y_1 & Q_{1,3} \\ y_1 & 2 + y_2 - 2Q_{1,3} & 0 \\ Q_{1,3} & 0 & 1 \end{bmatrix} \succeq 0.$$

If, instead, the entries of $\boldsymbol{Q}$ are expressed in terms of y_1, y_2, and $Q_{2,2}$ one has that for given y_1 and y_2 the polynomial $p(x)$ admits an SOS decomposition if there exists $Q_{2,2}$ such that

$$\boldsymbol{F}_2(y_1, y_2, Q_{2,2}) := \begin{bmatrix} 1 - y_2 & y_1 & \frac{2+y_2-Q_{2,2}}{2} \\ y_1 & Q_{2,2} & 0 \\ \frac{2+y_2-Q_{2,2}}{2} & 0 & 1 \end{bmatrix} \succeq 0.$$

As one would expect, these two LMIs are equivalent for the purposes of determining for which values of y_1 and y_2 the polynomial $p(x)$ admits an SOS decompositions. In fact, although their feasible sets

$$S_1 := \left\{(y_1, y_2, Q_{1,3}) \in \mathbb{R}^3 \,:\, \boldsymbol{F}_1(y_1, y_2, Q_{1,3}) \succeq 0\right\},$$
$$S_2 := \left\{(y_1, y_2, Q_{2,2}) \in \mathbb{R}^3 \,:\, \boldsymbol{F}_2(y_1, y_2, Q_{2,2}) \succeq 0\right\},$$

are different subsets of $\mathbb{R}^3$, their projections on the (y_1, y_2) plane coincide. This is illustrated in figure 2.2, which shows the projections of S_1 and S_2 onto the coordinate planes.

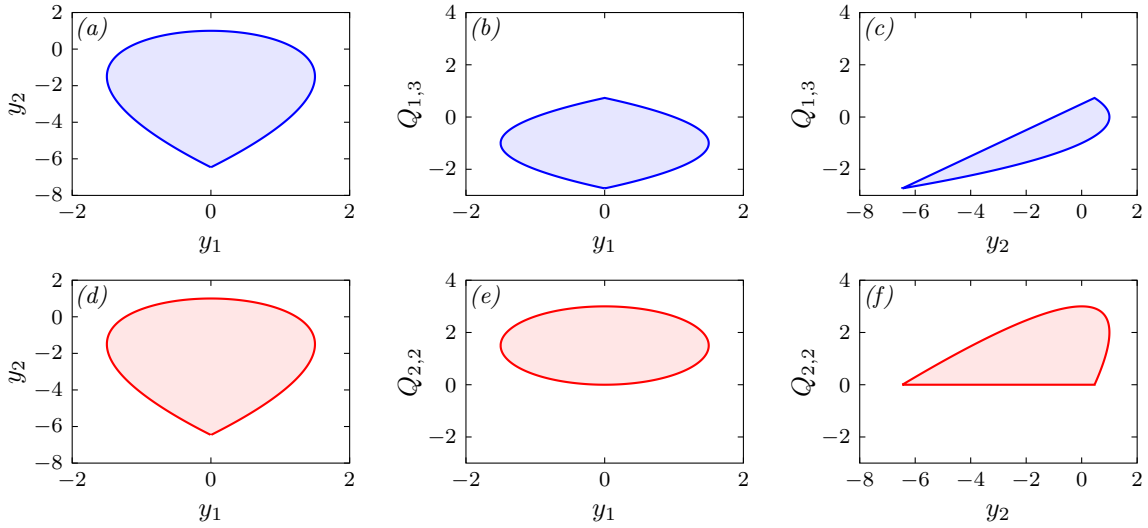


FIGURE 2.2: Panels *(a)–(c)*: projections on the coordinate planes of the feasible set S_1 of the LMI $\boldsymbol{F}_1(y_1, y_2, Q_{1,3}) \succeq 0$ in example 2.4. Panels *(d)–(f)*: projections on the coordinate planes of the feasible set S_2 of the LMI $\boldsymbol{F}_2(y_1, y_2, Q_{2,2}) \succeq 0$ in example 2.4.

2.4 Semidefinite programmes

Let $\boldsymbol{X} \in \mathbb{S}^n$ be a matrix optimisation variable with $n(n+1)/2$ independent degrees of freedom, and let $m \leq n(n+1)/2$. A semidefinite programme (SDP) is an optimisation problem of the form (known as *standard primal form*)

$$\begin{array}{ll} \min\limits_{\boldsymbol{X} \in \mathbb{S}^n} & \langle \boldsymbol{F}_0, \boldsymbol{X} \rangle \\ \text{s.t.} & \langle \boldsymbol{F}_i, \boldsymbol{X} \rangle = b_i, \quad i = 1, \ldots, m, \\ & \boldsymbol{X} \succeq 0, \end{array} \tag{2.13}$$

where $\boldsymbol{F}_i \in \mathbb{S}^n$, $i = 0, \ldots, m$, and $\boldsymbol{b} \in \mathbb{R}^m$ are given problem data. In other words, an SDP in standard primal form consists of the minimisation of a linear function of $\boldsymbol{X}$ subject to m affine equality constraints and the requirement that $\boldsymbol{X}$ is positive semidefinite. This can be done efficiently using a variety of algorithms (see Boyd *et al.*, 1994; Parrilo, 2013, and references therein), and many software packages that implement them are available open-source. The most common examples are SeDuMi (Sturm, 1999, 2002), SDPT3 (Toh *et al.*, 1999; Tütüncü *et al.*, 2003), and Mosek (Andersen *et al.*, 2009).

Since the constraint $\boldsymbol{X} \succeq 0$ is a particular type of LMI, it is perhaps not surprising that SDPs are closely related to optimisation problems with LMI-representable constraints. The link comes from Lagrangian duality: the dual problem associated with an SDP in standard primal form is a problem with an LMI constraint. To derive the dual problem, one augments the objective in (2.13) with the constraints using a vector of Lagrange multipliers $\boldsymbol{y} \in \mathbb{R}^m$

for the equality constraints, and a matrix Lagrange multiplier $\boldsymbol{Z}$ for the LMI $\boldsymbol{X} \succeq 0$. The Lagrangian for the SDP (2.13) is

$$\begin{aligned} L(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z}) &:= \langle \boldsymbol{F}_0, \boldsymbol{X} \rangle - \sum_{i=1}^{m} y_i \left(\langle \boldsymbol{F}_i, \boldsymbol{X} \rangle - b_i \right) - \langle \boldsymbol{X}, \boldsymbol{Z} \rangle \\ &= \left\langle \boldsymbol{F}_0 - \sum_{i=1}^{m} y_i \boldsymbol{F}_i - \boldsymbol{Z}, \boldsymbol{X} \right\rangle + \boldsymbol{b}^{\mathsf{T}} \boldsymbol{y}, \end{aligned} \tag{2.14}$$

and the dual function is

$$g(\boldsymbol{y}, \boldsymbol{Z}) = \inf_{\boldsymbol{X}} \; L(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z}) = \begin{cases} \boldsymbol{b}^{\mathsf{T}} \boldsymbol{y} & \text{if } \boldsymbol{F}_0 - \sum_{i=1}^{m} y_i \boldsymbol{F}_i - \boldsymbol{Z} = \boldsymbol{0}, \\ -\infty, & \text{otherwise.} \end{cases} \tag{2.15}$$

As in section 2.1, the dual function yields a lower bound on the optimal value $p^\star$ of (2.13). To see this, note that $\langle \boldsymbol{X}, \boldsymbol{Z} \rangle \geq 0$ for any $\boldsymbol{X}, \boldsymbol{Z} \succeq 0$: upon writing $\boldsymbol{X} = \sum_{i=1}^{n} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^{\mathsf{T}}$, where $\lambda_i \geq 0$ is the i-th eigenvalue and $\boldsymbol{v}_i$ is the corresponding eigenvector, one has

$$\langle \boldsymbol{X}, \boldsymbol{Z} \rangle = \operatorname{tr} \left(\sum_{i=1}^{n} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^{\mathsf{T}} \boldsymbol{Z} \right) = \sum_{i=1}^{n} \lambda_i \operatorname{tr} \left(\boldsymbol{v}_i \boldsymbol{v}_i^{\mathsf{T}} \boldsymbol{Z} \right) = \sum_{i=1}^{n} \lambda_i \boldsymbol{v}_i^{\mathsf{T}} \boldsymbol{Z} \boldsymbol{v}_i \geq 0. \tag{2.16}$$

Then, it is relatively straightforward to check that for any vector $\boldsymbol{y}$ and any matrix $\boldsymbol{Z} \succeq 0$

$$g(\boldsymbol{y}, \boldsymbol{Z}) = \inf_{\boldsymbol{X}} \; L(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z}) \leq \inf_{\substack{\boldsymbol{X} \succeq 0, \\ \langle \boldsymbol{F}_i, \boldsymbol{X} \rangle = b_i}} \; L(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z}) \leq p^\star. \tag{2.17}$$

Consequently, the dual problem associated with (2.13) is

$$\begin{aligned} \max_{\boldsymbol{y} \in \mathbb{R}^m, \boldsymbol{Z} \in \mathbb{S}^n} \quad & \boldsymbol{b}^{\mathsf{T}} \boldsymbol{y} \\ \text{s.t.} \quad & \boldsymbol{F}_0 - \sum_{i=1}^{m} y_i \boldsymbol{F}_i = \boldsymbol{Z}, \\ & \boldsymbol{Z} \succeq 0, \end{aligned} \tag{2.18}$$

from which the matrix $\boldsymbol{Z}$ can be eliminated to arrive at the LMI-constrained optimisation problem

$$\begin{aligned} \max_{\boldsymbol{y} \in \mathbb{R}^m} \quad & \boldsymbol{b}^{\mathsf{T}} \boldsymbol{y} \\ \text{s.t.} \quad & \boldsymbol{F}_0 - \sum_{i=1}^{m} y_i \boldsymbol{F}_i \succeq 0. \end{aligned} \tag{2.19}$$

Note that the optimal value of (2.19) is a strict lower bound on the optimal value of (2.13) unless strong duality holds. Slater's condition guarantees that this is the case, and moreover

that the optimal value of (2.19) is attained by an optimal point $\boldsymbol{y}^\star$, if there exists $\boldsymbol{X}_0 \succ 0$ satisfying the equality constraints in (2.13).

The argument outlined above can be reversed to show that the dual of optimisation problems with an LMI (or LMI-representable) constraint in the form (2.19) is an SDP in the standard primal form (2.13). Similarly, the relevant version of Slater's condition states that strong duality holds and there exists an optimal point $\boldsymbol{X}^\star$ that attains the optimal value of (2.13) if there exists $\boldsymbol{y} \in \mathbb{R}^m$ such that $\boldsymbol{F}_0 - \sum_{i=1}^m y_i \boldsymbol{F}_i \succ 0$.

Given the strong link between SDPs in standard primal form and LMI-constrained problems, the latter are often also referred to as SDPs. Indeed, problem (2.19) is known in the literature as an SDP in *standard dual form*. To further simplify the presentation throughout the rest of this thesis, moreover, it will be convenient to call an SDP any optimisation problem of the form

$$\begin{aligned} \max_{\boldsymbol{y} \in \mathbb{R}^m} \quad & \boldsymbol{b}^\mathsf{T} \boldsymbol{y} \\ \text{s.t.} \quad & \boldsymbol{B}\boldsymbol{y} = \boldsymbol{c}, \\ & \boldsymbol{F}_i(\boldsymbol{y}) \succeq 0, \quad i = 1, \ldots, q, \end{aligned} \tag{2.20}$$

where $\boldsymbol{b} \in \mathbb{R}^m$ is the cost vector, $\boldsymbol{B} \in \mathbb{R}^{p \times m}$ and $\boldsymbol{c} \in \mathbb{R}^p$ define p affine equality constraints (it is assumed that $\operatorname{rank}(\boldsymbol{B}) = p < m$, so there are no redundant equalities and the problem is not over-constrained), and $\boldsymbol{F}_i(\boldsymbol{y}) \succeq 0$, $i = 1, \ldots, q$, are LMIs. Problems with LMI-representable constraints will also be called SDPs. This slight abuse of terminology is justified because multiple LMIs $\boldsymbol{F}_1(\boldsymbol{y}) \succeq 0, \ldots, \boldsymbol{F}_q(\boldsymbol{y}) \succeq 0$ are equivalent to the block-diagonal LMI

$$\begin{bmatrix} \boldsymbol{F}_1(\boldsymbol{y}) & & \\ & \ddots & \\ & & \boldsymbol{F}_q(\boldsymbol{y}) \end{bmatrix} \succeq 0.$$

Moreover, the affine equality constraints on $\boldsymbol{y}$ can be eliminated upon considering the change of variable $\boldsymbol{y} = \boldsymbol{v} + \boldsymbol{D}\tilde{\boldsymbol{y}}$, where $\boldsymbol{v}$ is any vector satisfying $\boldsymbol{B}\boldsymbol{v} = \boldsymbol{c}$, $\boldsymbol{D} \in \mathbb{R}^{m \times (m-p)}$ is a matrix such that $\operatorname{img}(\boldsymbol{D}) \equiv \ker(\boldsymbol{B})$, and $\tilde{\boldsymbol{y}} \in \mathbb{R}^{m-p}$ is the new optimisation variable. Consequently, optimisation problems of the general form (2.20) can always be rewritten as single-LMI, equality-free problems, and therefore also as dual-standard-form SDPs.

2.5 A useful complementarity result

Suppose that the SDP (2.13) and the LMI-constrained problem (2.19) are strongly dual, so their optimal values coincide, and assume that these are attained by optimal points $\boldsymbol{X}^\star$

and $\boldsymbol{y}^\star$. According to Slater's condition, this is guaranteed if there exist a strictly positive definite $\boldsymbol{X}$ satisfying the equality constraints in (2.13) and a vector $\boldsymbol{y}$ in (2.19) such that $\boldsymbol{F}_0 - \sum_{i=1}^m y_i \boldsymbol{F}_i \succ 0$. Then, the following complementarity result holds.

Proposition 2.2. *Assume that strong duality holds for* (2.13) *and* (2.19)*, and that their common optimal value is attained by optimal points $\boldsymbol{X}^\star$ and $\boldsymbol{y}^\star$. Then, the complementarity condition $\langle \boldsymbol{F}(\boldsymbol{y}^\star), \boldsymbol{X}^\star \rangle = 0$ holds. Moreover, if $\boldsymbol{F}(\boldsymbol{y}^\star)$ has rank $n - r$ for some integer r, $0 \leq r \leq n$, then* $\operatorname{rank}(\boldsymbol{X}^\star) \leq r$.

Proof. Since $\boldsymbol{X}^\star$ and $\boldsymbol{y}^\star$ are optimal solutions of (2.13) and (2.19) and strong duality holds, $\boldsymbol{b}^\mathsf{T}\boldsymbol{y}^\star = \langle \boldsymbol{F}_0, \boldsymbol{X}^\star \rangle$. Moreover, since $\boldsymbol{X}^\star \succeq 0$ and $\boldsymbol{F}(\boldsymbol{y}^\star) \succeq 0$, one has $\langle \boldsymbol{F}(\boldsymbol{y}^\star), \boldsymbol{X}^\star \rangle \geq 0$. Then,

$$\boldsymbol{b}^\mathsf{T}\boldsymbol{y}^\star \geq \boldsymbol{b}^\mathsf{T}\boldsymbol{y}^\star - \langle \boldsymbol{F}(\boldsymbol{y}^\star), \boldsymbol{X}^\star \rangle = \boldsymbol{b}^\mathsf{T}\boldsymbol{y}^\star - \langle \boldsymbol{F}_0, \boldsymbol{X}^\star \rangle + \sum_{i=1}^m y_i^\star \underbrace{\langle \boldsymbol{F}_i, \boldsymbol{X}^\star \rangle}_{=b_i} = \boldsymbol{b}^\mathsf{T}\boldsymbol{y}^\star.$$

The first inequality must therefore be an equality, which implies $\langle \boldsymbol{F}(\boldsymbol{y}^\star), \boldsymbol{X}^\star \rangle = 0$. Moreover, if $\boldsymbol{F}(\boldsymbol{y}^\star)$ has rank $n - r$, then it can be written as $\boldsymbol{F}(\boldsymbol{y}^\star) = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\mathsf{T}$, where $\boldsymbol{\Lambda} \in \mathbb{S}^{n-r}$ is the diagonal matrix of strictly positive eigenvalues and $\boldsymbol{U} \in \mathbb{R}^{n \times (n-r)}$ is the matrix of corresponding eigenvectors. Similarly, if $\operatorname{rank}(\boldsymbol{X}^\star) = p \leq n$, then $\boldsymbol{X}^\star = \boldsymbol{V}\boldsymbol{\Gamma}\boldsymbol{V}^\mathsf{T}$ with $\boldsymbol{\Gamma} \in \mathbb{S}^p$ the diagonal matrix of strictly positive eigenvalues and $\boldsymbol{V} \in \mathbb{R}^{n \times p}$ the matrix of corresponding eigenvectors. Consequently, the properties of the trace inner product imply

$$0 = \langle \boldsymbol{F}(\boldsymbol{y}^\star), \boldsymbol{X}^\star \rangle = \left\langle \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\mathsf{T}, \boldsymbol{V}\boldsymbol{\Gamma}\boldsymbol{V}^\mathsf{T} \right\rangle = \left\langle \boldsymbol{\Lambda}, \left(\boldsymbol{V}^\mathsf{T}\boldsymbol{U}\right)^\mathsf{T} \boldsymbol{\Gamma} \left(\boldsymbol{V}^\mathsf{T}\boldsymbol{U}\right) \right\rangle.$$

This means that $\boldsymbol{V}^\mathsf{T}\boldsymbol{U} = \boldsymbol{0}$, *i.e.*, the p-dimensional eigenspace of $\boldsymbol{X}^\star$ must be orthogonal to the $(n - r)$-dimensional eigenspace of $\boldsymbol{F}(\boldsymbol{y}^\star)$. Therefore, $p = \operatorname{rank}(\boldsymbol{X}^\star) \leq r$. □

Proposition 2.2 enables a useful low-rank representation of the Lagrangian of the LMI-constrained problem (2.19) when a lower bound on the rank of its optimal solution is available. In fact, if it can be established that $\boldsymbol{F}(\boldsymbol{y}^\star)$ has rank at least $n - r$, then the Lagrange multiplier $\boldsymbol{X}$ for the LMI constraint in (2.19) may be assumed to take the rank-r form $\boldsymbol{X} = \sum_{i=1}^r \boldsymbol{v}_i \boldsymbol{v}_i^\mathsf{T}$ for some vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r$. Note that these vectors need not be orthonormal, and that although the rank-r form forces $\boldsymbol{X} \succeq 0$ there is no loss of generality because to derive the dual of (2.19) one eventually restricts attention to positive semidefinite Lagrange multipliers. Consequently, the Lagrangian for (2.19) can be written as

$$L(\boldsymbol{y}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_r) = \boldsymbol{b}^\mathsf{T}\boldsymbol{y} - \sum_{i=1}^r \boldsymbol{v}_i^\mathsf{T} \boldsymbol{F}(\boldsymbol{y}) \boldsymbol{v}_i. \tag{2.21}$$

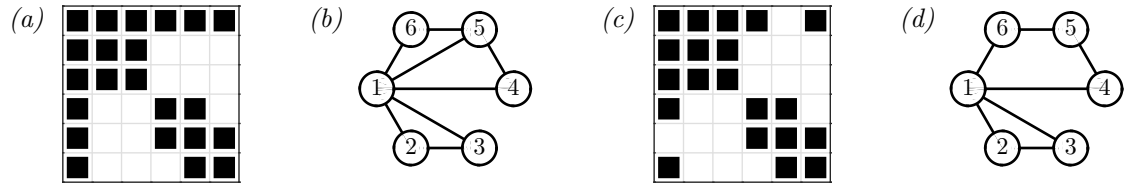


FIGURE 2.3: *(a)* A possible sparsity pattern for a 6×6 symmetric matrix. *(b)* Graph representation of the sparsity pattern in panel *(a)*. *(c)* A different sparsity pattern for a 6×6 symmetric matrix. *(d)* Graph representation of the sparsity pattern in panel *(c)*.

When $r \ll n$, the low-rank representation of the positive semidefinite Lagrange multiplier $\boldsymbol{X}$ can be exploited to reduce the computational resources required to solve extremely large SDPs (Burer *et al.*, 2002; Burer & Monteiro, 2003, 2005; Burer & Choi, 2006). In addition, the low-rank form (2.21) of the Lagrangian will become useful in chapter 3.

2.6 Chordal decomposition of LMIs

A challenge for the practical solution of SDPs is that while state-of-the-art software packages can handle multiple LMIs very efficiently, the computational cost of a single LMI grows non-linearly as a function of its size. This section reviews the method of *chordal decomposition*, developed by Fukuda *et al.* (2000), Nakata *et al.* (2003), and Kim *et al.* (2011) to replace a large, sparse LMI with a set of multiple, smaller LMIs.

The method begins with the realisation that the sparsity pattern of a matrix $\boldsymbol{X} \in \mathbb{S}^n$ can be represented by a graph $\mathcal{G}(V, E)$, where $V = \{1, \ldots, n\}$ is the set of vertices in the graph and $E \subseteq V \times V$ is a set of edges (or connections) between vertices such that $(i, j) \in E$ if $\boldsymbol{X}_{i,j} \neq 0$. For instance, the graphs shown in figures 2.3*(b,d)* correspond to 6×6 symmetric matrices with sparsity patterns illustrated in figures 2.3*(a,c)*. To exploit the link between matrices and graphs, it is necessary to introduce some terminology used in graph theory.

Consider a graph $\mathcal{G}(V, E)$ with vertices $V = \{1, \ldots, n\}$ and edges $E \subseteq V \times V$. A vertex $i \in V$ is called *simplicial* if all of its neighbours are pairwise connected, meaning that if $j, k \neq i$ are any two distinct vertices such that $(i, j) \in E$ and $(i, k) \in E$, then $(j, k) \in E$. A subset of vertices $C \subseteq V$ such that $(i, j) \in E$ for any distinct vertices $i, j \in C$ is called a *clique*. The number of vertices in C is denoted by $|C|$, and if C is not a subset of any other clique it is called a *maximal clique*. It can be shown (Blair & Peyton, 1993, Lemma 3) that a simplicial vertex belongs to one and only one maximal clique. A *cycle* of length k is a set of pairwise distinct vertices $\{v_1, \ldots, v_k\} \subseteq V$ such that $(v_k, v_1) \in E$ and $(v_i, v_{i+1}) \in E$ for all $i = 1, \ldots, k-1$, while a *chord* is an edge joining two non-consecutive vertices in a cycle. For example, the graph in figure 2.3*(b)* has maximal cliques $C_1 = \{1, 2, 3\}$, $C_2 = \{1, 4, 5\}$,

and $C_3 = \{1, 5, 6\}$, which are also cycles of length 3. Vertices 2, 3, 4, and 6 are simplicial and belong to one and only one clique, while vertices 1 and 5 are not simplicial. The cycle $\{1, 4, 5, 6\}$ has length 4 and has a chord, (1, 5). Similarly, the graph in figure 2.3*(d)* has maximal cliques $C_1 = \{1, 2, 3\}$, $C_2 = \{1, 4\}$, $C_3 = \{1, 6\}$, $C_4 = \{4, 5\}$, and $C_5 = \{5, 6\}$. Clique C_1 is a cycle of length 3, while $\{1, 4, 5, 6\}$ is a cycle of length 4 and has no chords. Vertices 2 and 3 are simplicial and belong only to clique C_1, while vertices 1, 4, 5, and 6 are not simplicial.

A graph is called *chordal* if every cycle of length larger than 3 has at least one chord, and the sparsity pattern of a matrix is said to be chordal if its associated graph is so. Thus, the sparsity pattern illustrated in figure 2.3*(a)* is chordal, since the only cycle of length 4 of the corresponding graph in figure 2.3*(b)* has the chord (1, 5). Instead, since the cycle {1, 4, 5, 6} of the graph in figure 2.3*(d)* has no chords, the sparsity pattern in figure 2.3*(c)* is not chordal. Chordality is an extremely useful property because the question of whether a matrix with chordal sparsity pattern is positive semidefinite can be answered utilising the maximal cliques of the underlying graph (Agler *et al.*, 1988). To make this idea precise, let $C_1, \ldots, C_p$ be the maximal cliques of a chordal graph and, for each $k = 1, \ldots, p$, define $\boldsymbol{E}_k \in \mathbb{R}^{|C_k| \times n}$ according to

$$(\boldsymbol{E}_k)_{i,j} = \begin{cases} 1, & \text{if } C_k(i) = j, \\ 0, & \text{otherwise.} \end{cases} \tag{2.22}$$

Here, $C_k(i)$ denotes the i-th vertex in the clique C_k, when the vertices are sorted in the natural ordering. In other words, the matrix $\boldsymbol{E}_k$ is such that, given a $|C_k| \times |C_k|$ matrix $\boldsymbol{Y}$, the operation $\boldsymbol{E}_k^{\mathsf{T}} \boldsymbol{Y} \boldsymbol{E}_k$ creates a sparse $n \times n$ matrix such that $\boldsymbol{Y}$ is the principal submatrix identified by C_k, while all other entries are zero. For instance, consider the chordal graph in figure 2.3*(b)*, whose maximal cliques are $C_1 = \{1, 2, 3\}$, $C_2 = \{1, 4, 5\}$, and $C_3 = \{1, 5, 6\}$. For $k = 2$ and $\boldsymbol{Y} \in \mathbb{S}^3$ one has

$$\boldsymbol{E}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \qquad \boldsymbol{E}_2^{\mathsf{T}} \boldsymbol{Y} \boldsymbol{E}_2 = \begin{bmatrix} Y_{1,1} & 0 & 0 & Y_{1,2} & Y_{1,3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ Y_{1,2} & 0 & 0 & Y_{2,2} & Y_{2,3} & 0 \\ Y_{1,3} & 0 & 0 & Y_{2,3} & Y_{3,3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Using this notation, the key result due to Agler *et al.* (1988) may be stated as follows.

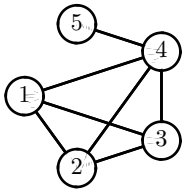


FIGURE 2.4: Graph representation of the sparsity pattern of LMI (2.23) in example 2.5. The graph is chordal and has two maximal cliques, $C_1 = \{1, 2, 3, 4\}$ and $C_2 = \{4, 5\}$.

Theorem 2.3 (Agler *et al.*, 1988)**.** *Assume that $\boldsymbol{X} \in \mathbb{S}^n$ has a chordal sparsity pattern and that the associated graph $\mathcal{G}\{V, E\}$ has p maximal cliques $C_1, \ldots, C_p$. Then, $\boldsymbol{X} \succeq 0$ if and only if there exist symmetric matrices $\boldsymbol{Y}_k \in \mathbb{S}^{|C_k|}$, $k = 1, \ldots, p$, such that $\boldsymbol{Y}_k \succeq 0$ for all $k = 1, \ldots, p$ and $\boldsymbol{X} = \sum_{k=1}^{p} \boldsymbol{E}_k^{\mathsf{T}} \boldsymbol{Y}_k \boldsymbol{E}_k$.*

Note that the "if" part of the theorem is immediate and does not rely on chordality. Instead, the "only if" part relies on the facts that chordal graphs have at least one simplicial vertex (Blair & Peyton, 1993, Lemma 1), which belongs to one and only one maximal clique, and that any graph obtained by removing a simplicial vertex and all edges connecting to it is also chordal (Blair & Peyton, 1993). The interested reader can find a complete and relatively simple proof of Theorem 2.3 in a recent work by Kakimura (2010).

In the context of SDPs, Theorem 2.3 means that when problem (2.19) has a large and sparse LMI with chordal sparsity pattern, it can be substituted with an equivalent set of smaller LMIs (each as large as the corresponding maximal clique) plus a set of affine equality constraints. This procedure can be automated using the MATLAB package SparseCoLO (Fujisawa *et al.*, 2009), and it can substantially improve computational efficiency if the maximal cliques of the chordal graph are small (Nakata *et al.*, 2003; Kim *et al.*, 2011).

Example 2.5. To illustrate how the chordal decomposition method described above is implemented in practice, consider the 5×5 LMI from example 2.2,

$$\boldsymbol{F}(\boldsymbol{y}) = \begin{bmatrix} 3y_1 & y_2 - y_1 & 2y_2 & y_2 - 1 & 0 \\ y_2 - y_1 & 5 - y_2 & -y_2 & y_1 & 0 \\ 2y_2 & -y_2 & 2 - y_1 & y_1 + y_2 & 0 \\ y_2 - 1 & y_1 & y_1 + y_2 & 2 + y_2 & -1 \\ 0 & 0 & 0 & -1 & 5 \end{bmatrix} \succeq 0. \tag{2.23}$$

The graph representing its sparsity pattern, shown in figure 2.4, has two maximal cliques, $C_1 = \{1, 2, 3, 4\}$ and $C_2 = \{4, 5\}$, and is chordal because the only cycle of length 4 has one

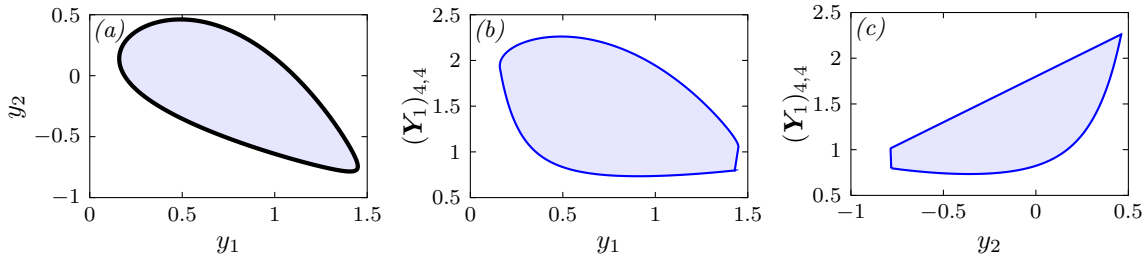


FIGURE 2.5: Projections on the coordinate planes (y_1, y_2), $(y_1, (\boldsymbol{Y}_2)_{4,4})$, and $(y_2, (\boldsymbol{Y}_2)_{4,4})$ of the joint feasible set of the two LMIs in (2.24). Also shown in panel *(a)* is the boundary of the feasible set of the original 5×5 LMI (2.23) (—).

chord. Consequently, Theorem 2.3 implies that (2.23) holds if there exists matrices

$$\boldsymbol{Y}_1 := \begin{bmatrix} (\boldsymbol{Y}_1)_{1,1} & (\boldsymbol{Y}_1)_{1,2} & (\boldsymbol{Y}_1)_{1,3} & (\boldsymbol{Y}_1)_{1,4} \\ (\boldsymbol{Y}_1)_{1,2} & (\boldsymbol{Y}_1)_{2,2} & (\boldsymbol{Y}_1)_{2,3} & (\boldsymbol{Y}_1)_{2,4} \\ (\boldsymbol{Y}_1)_{1,3} & (\boldsymbol{Y}_1)_{2,3} & (\boldsymbol{Y}_1)_{3,3} & (\boldsymbol{Y}_1)_{3,4} \\ (\boldsymbol{Y}_1)_{1,4} & (\boldsymbol{Y}_1)_{2,4} & (\boldsymbol{Y}_1)_{3,4} & (\boldsymbol{Y}_1)_{4,4} \end{bmatrix} \succeq 0, \qquad \boldsymbol{Y}_2 := \begin{bmatrix} (\boldsymbol{Y}_2)_{1,1} & (\boldsymbol{Y}_2)_{1,2} \\ (\boldsymbol{Y}_2)_{1,2} & (\boldsymbol{Y}_2)_{2,2} \end{bmatrix} \succeq 0,$$

such that

$$\begin{bmatrix} (\boldsymbol{Y}_1)_{1,1} & (\boldsymbol{Y}_1)_{1,2} & (\boldsymbol{Y}_1)_{1,3} & (\boldsymbol{Y}_1)_{1,4} & 0 \\ (\boldsymbol{Y}_1)_{1,2} & (\boldsymbol{Y}_1)_{2,2} & (\boldsymbol{Y}_1)_{2,3} & (\boldsymbol{Y}_1)_{2,4} & 0 \\ (\boldsymbol{Y}_1)_{1,3} & (\boldsymbol{Y}_1)_{2,3} & (\boldsymbol{Y}_1)_{3,3} & (\boldsymbol{Y}_1)_{3,4} & 0 \\ (\boldsymbol{Y}_1)_{1,4} & (\boldsymbol{Y}_1)_{2,4} & (\boldsymbol{Y}_1)_{3,4} & (\boldsymbol{Y}_1)_{4,4} + (\boldsymbol{Y}_2)_{1,1} & (\boldsymbol{Y}_2)_{1,2} \\ 0 & 0 & 0 & (\boldsymbol{Y}_2)_{1,2} & (\boldsymbol{Y}_2)_{2,2}. \end{bmatrix} = \boldsymbol{F}(\boldsymbol{y}).$$

Upon eliminating all entries of $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ except for $(\boldsymbol{Y}_1)_{4,4}$, one concludes that the original LMI (2.23) holds for given values y_1 and y_2 if there exists $(\boldsymbol{Y}_1)_{4,4}$ such that

$$\begin{bmatrix} 3y_1 & y_2 - y_1 & 2y_2 & y_2 - 1 \\ y_2 - y_1 & 5 - y_2 & -y_2 & y_1 \\ 2y_2 & -y_2 & 2 - y_1 & y_1 + y_2 \\ y_2 - 1 & y_1 & y_1 + y_2 & (\boldsymbol{Y}_1)_{4,4} \end{bmatrix} \succeq 0, \qquad \begin{bmatrix} 2 + y_2 - (\boldsymbol{Y}_1)_{4,4} & -1 \\ -1 & 5 \end{bmatrix} \succeq 0. \tag{2.24}$$

Figure 2.5 shows the joint feasible set of these two LMIs. As one would expect, the projection on the (y_1, y_2) plane coincides with the feasible set of the original LMI (2.23), which is plotted in figure 2.1*(b)* and whose boundary is shown in figure 2.5 to ease the comparison. This confirms that the LMIs in (2.24) are equivalent to (2.23). So, the size of the largest LMI can be reduced by introducing an extra variable and increasing the number of LMIs.

Chapter 3

Asymptotic energy bounds for the Kuramoto–Sivashinsky equation using time-marching methods

The Kuramoto–Sivashinsky (KS) equation is a PDE with a hydrodynamic-type nonlinearity that describes the weakly nonlinear dynamics of reaction-diffusion systems (Kuramoto & Tsuzuki, 1975, 1976), flame front oscillations (Sivashinsky, 1977, 1980; Michelson & Sivashinsky, 1977), and waves on the surface of thin liquid films (Sivashinsky & Michelson, 1980). This chapter studies the one-dimensional version of the KS equation,

$$\partial_t u + u\,\partial_x u + \partial_x^2 u + \partial_x^4 u = 0, \tag{3.1}$$

considered on $[-\ell, \ell]$ with periodic boundary conditions (BCs) and zero-average initial conditions (ICs):

$$\partial_x^\alpha u(-\ell) = \partial_x^\alpha u(\ell), \quad \alpha = 0, \ldots, 3, \tag{3.2a}$$

$$u(x,0) = u_0(x), \quad \int_{-\ell}^{\ell} u_0(x)\,\mathrm{d}x = 0. \tag{3.2b}$$

With such boundary and initial conditions, (3.1) has a unique solution at all times in the appropriate function space (see, for example, Robinson, 2001, chapter 17). The domain half-size ℓ is the governing parameter and the trivial solution $u(x,t) = 0$ is globally asymptotically stable when $\ell < \pi$. Windows of chaotic dynamics are observed as ℓ is increased through π (Hyman & Nicolaenko, 1986; Hyman *et al.*, 1986), making the KS equation (3.1) a paradigm for chaotic systems.

It is a long-standing conjecture (Wittenberg, 2002) that, at large ℓ, the asymptotic kinetic energy of solutions of the KS equation grows proportionally to ℓ,

$$\mathcal{E} := \limsup_{t\to\infty} \|u(x,t)\|_2^2 \sim \ell. \tag{3.3}$$

Many researchers have tried to prove this fact by deriving upper bounds on $\mathcal{E}$. Nicolaenko *et al.* (1985) proved that $\mathcal{E} \lesssim \ell^5$ for odd solutions of (3.1). Their argument relied on the application of the so-called background method: they considered fluctuations $v(x,t) = u(x,t) - \phi(x)$ from a steady background field $\phi(x)$, chosen subject to certain constraints that ensure boundedness of $\|v\|_2^2$ as $t \to \infty$ (see section 3.1 for more details). The restriction to odd solutions is justified because (3.1) is invariant under the transformation $u(x,t) \to -u(-x,t)$, so if the initial condition is odd, then the solution remains odd at all times. Goodman (1994) subsequently extended the result to generic solutions, while Collet *et al.* (1993) improved the bound to $\mathcal{E} \lesssim \ell^{16/5}$ with a more careful choice of background field. More recently, Bronski & Gambill (2006) proved that $\mathcal{E} \lesssim \ell^3$, which to this date remains the best result obtained with the background method.[1] Bronski & Gambill (2006) also showed that the scaling of their bound with ℓ cannot be improved if the form of the background field is restricted such that $\phi'(x) = c + q(x)$, where c is a constant and $q(x)$ a function with compact support on $(-\ell, \ell)$. This partially proved the hypothesis, already put forward by Wittenberg (2002), that a bound proportional to ℓ^3 is optimal within the background method. Strong numerical evidence of this fact has recently been provided by Fantuzzi & Wynn (2015), who employed semidefinite programming to optimise ϕ.

In this chapter, the optimisation of the background field for the KS equation will be repeated using a time-marching method proposed by Wen *et al.* (2013, 2015). The essence of this approach is to find a saddle point of the Lagrangian of the variational problem for the optimal ϕ by solving a time-dependent version of the Euler–Lagrange (EL) equations that describe stationary points of the Lagrangian. It will be shown that although this strategy has been employed successfully by Wen *et al.* (2013, 2015) to solve some optimal background field problems arising in fluid dynamics, its application to the KS equation is not straightforward. The main difficulty, showcased in section 3.3, is that convergence to an incorrect solution can occur when the Lagrangian is constructed as inferred from the examples given by Wen *et al.* (2013, 2015). Precisely, the computed ϕ solves the EL equations, but is not the optimal background field because it does not satisfy the constraints arising from the background method analysis. Section 3.4 demonstrates that this problem seems to be resolved by considering a different Lagrangian, but this requires additional non-trivial insight on the constraints on ϕ. Another issue is that convergence of the time-marching method to the optimal background field cannot be guaranteed even when the correct Lagrangian is used. This is true also for the problems considered by Wen *et al.* (2013, 2015), but they could

[1] More sophisticated mathematical arguments have resulted in better bounds. For example, by carefully comparing solutions of the KS equation to so-called *entropy solutions* of the inviscid Burger's equation, Giacomelli & Otto (2005) proved that $\mathcal{E} = o(\ell^3)$, in the sense that $\lim_{\ell \to +\infty} \ell^{-3}\mathcal{E} = 0$.

at least establish that if a steady solution of the time-dependent EL equation is obtained, then the computed background field is necessarily the optimal one. Their proof, however, does not extend to the KS equation, so convergence to a background field that does not satisfy the required constraints remains a theoretical possibility. Similar difficulties—both in constructing the correct Lagrangian functional and in proving theoretical convergence guarantees—may arise also when optimising background fields beyond the KS equation. This motivates the development of alternative strategies for the computation of optimal background fields, which can be applied when time-marching methods fail.

3.1 Background method analysis

An upper bound on the asymptotic energy of the solution of (3.1) can be derived with the background method. This section describes the argument used by Collet *et al.* (1993), tuned to arrive at the same variational problem for the optimal background field solved by Fantuzzi & Wynn (2015). It is assumed that the initial condition for (3.1) is odd, so $u(x,t)$ remains odd at all times. Bounds obtained for odd solutions can be extended to general solutions using standard arguments (Collet *et al.*, 1993; Goodman, 1994).

The analysis begins by decomposing $u(x,t) = \phi(x) + v(x,t)$, with ϕ and v odd and periodic, so (3.1) becomes

$$\partial_t v + \phi\,\partial_x\phi + v\,\partial_x\phi + \phi\,\partial_x v + v\,\partial_x v + \partial_x^2 v + \partial_x^4 v + \partial_x^2\phi + \partial_x^4\phi = 0. \tag{3.4}$$

Multiplying this equation by v, integrating by parts over $[-\ell,\ell]$, and rearranging yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\|v\|_2^2}{2} = -\int \left|\partial_x^2 v\right|^2 - |\partial_x v|^2 + \frac{1}{2}\phi'\,v^2\,\mathrm{d}x - \int \partial_x^2 v\,\phi'' - \partial_x v\,\phi' + \phi'\,\phi\,v\,\mathrm{d}x. \tag{3.5}$$

Here and in what follows the limits of integration are omitted to lighten the notation. Equation (3.5) is well defined if ϕ and v belong to the space $H_{\mathrm{p,o}}$ of square-integrable, odd, and periodic functions on $[-\ell,\ell]$ with two square-integrable and periodic derivatives,

$$H_{\mathrm{p,o}} := \big\{ v \in L^2(-\ell,\ell) \,:\, v(-x) = -v(x),\ v(-\ell) = v(\ell), \\ \text{and } \partial^\alpha v \in L^2(-\ell,\ell),\ \partial^\alpha v(-\ell) = \partial^\alpha v(\ell) \text{ for } \alpha \in \{1,2\} \big\}. \tag{3.6}$$

In fact, any functions $\phi, v \in H_{\mathrm{p,o}}$ must be at least continuously differentiable even if differentiation is understood in the weak sense (see, for instance, Theorem 8.2 in Brezis, 2010).

In particular one concludes that $\phi' \in L^\infty(-\ell, \ell)$ and, therefore, the right-hand side of (3.5) is bounded for all $\phi, v \in H_{\mathrm{p,o}}$.

Suppose now that ϕ is such that, for all $w \in H_{\mathrm{p,o}}$ and some $\epsilon > 0$,

$$\int \left|w''\right|^2 - \left|w'\right|^2 + \phi' w^2 \,\mathrm{d}x \geq \epsilon \left\|w\right\|_2^2 . \tag{3.7}$$

Then, the symmetric bilinear form

$$\mathcal{B}\{f, g\} := \int f''g'' - f'g' + \phi'\, f\, g \,\mathrm{d}x \tag{3.8}$$

is positive definite, and so it defines an inner product on $H_{\mathrm{p,o}}$. Consequently, the Cauchy–Schwarz inequality implies that $|\mathcal{B}\{f, g\}| \leq (\mathcal{B}\{f, f\})^{1/2} (\mathcal{B}\{g, g\})^{1/2}$ and one has

$$\begin{aligned} |\mathcal{B}\{v, \phi\}| &\leq \left(\int \left|\partial_x^2 v\right|^2 - \left|\partial_x v\right|^2 + \phi'\, v^2 \,\mathrm{d}x \right)^{\frac{1}{2}} \left(\int \left|\phi''\right|^2 - \left|\phi'\right|^2 + \phi' \phi^2 \,\mathrm{d}x \right)^{\frac{1}{2}} \\ &\leq \frac{1}{4} \int \left|\partial_x^2 v\right|^2 - \left|\partial_x v\right|^2 + \phi'\, v^2 \,\mathrm{d}x + \int \left|\phi''\right|^2 - \left|\phi'\right|^2 \,\mathrm{d}x, \end{aligned} \tag{3.9}$$

where the elementary inequality $ab \leq a^2/4 + b^2$ was used to obtain the second inequality and the term $\int \phi' \phi^2 \,\mathrm{d}x$ vanishes by periodicity. Using (3.9) to estimate the second integral on the right-hand side of (3.5) and rearranging yields

$$\frac{\mathrm{d}}{\mathrm{d}t} \frac{\left\|v\right\|_2^2}{2} \leq -\frac{3}{4} \int \left|\partial_x^2 v\right|^2 - \left|\partial_x v\right|^2 + \frac{1}{3} \phi'\, v^2 \,\mathrm{d}x + \int \left|\phi''\right|^2 - \left|\phi'\right|^2 \,\mathrm{d}x. \tag{3.10}$$

At this stage, suppose that, in addition to satisfying (3.7), the background field ϕ is such that all $v \in H_{\mathrm{p,o}}$ satisfy

$$\int \left|v''\right|^2 - \left|v'\right|^2 + \frac{1}{3} \phi'\, v^2 \,\mathrm{d}x \geq \frac{1}{2} \left\|v\right\|_2^2 . \tag{3.11}$$

Then, after bounding the first term on the right-hand side of (3.10), Gronwall's inequality (Doering & Gibbon, 1995, chapter 2) implies

$$\limsup_{t \to \infty} \left\|v\right\|_2^2 \leq \frac{8}{3} \int \left|\phi''\right|^2 - \left|\phi'\right|^2 \,\mathrm{d}x. \tag{3.12}$$

Finally, combining this result with the elementary estimate $\left\|u\right\|_2^2 \leq 2 \left\|\phi\right\|_2^2 + 2 \left\|v\right\|_2^2$ yields a bound on the asymptotic energy $\mathcal{E}$ of the KS equation:

$$\mathcal{E} \leq 2 \int \frac{8}{3} \left|\phi''\right|^2 - \frac{8}{3} \left|\phi'\right|^2 + \left|\phi\right|^2 \,\mathrm{d}x. \tag{3.13}$$

Recalling that the background field ϕ must satisfy conditions (3.7) and (3.11), the best bound on $\mathcal{E}$ obtainable within this bounding framework is given by

$$\begin{aligned} \inf_{\phi(x),\,\epsilon} \quad & 2\int \frac{8}{3}\left|\phi''\right|^2 - \frac{8}{3}\left|\phi'\right|^2 + |\phi|^2 \,\mathrm{d}x \\ \text{s.t.} \quad & \int \left|v''\right|^2 - \left|v'\right|^2 + \left(\frac{1}{3}\phi' - \frac{1}{2}\right) v^2 \,\mathrm{d}x \geq 0 \qquad \forall v \in H_{\mathrm{p,o}}, \\ & \int \left|w''\right|^2 - \left|w'\right|^2 + \left(\phi' - \epsilon\right) w^2 \,\mathrm{d}x \geq 0 \qquad \forall w \in H_{\mathrm{p,o}}, \\ & \epsilon > 0. \end{aligned} \tag{3.14}$$

In fact, noticing that the infimum over $\epsilon > 0$ coincides with the case $\epsilon = 0$, after dropping a factor of 2 from the objective it suffices to solve the variational problem

$$\begin{aligned} \inf_{\phi(x)} \quad & \int \frac{8}{3}\left|\phi''\right|^2 - \frac{8}{3}\left|\phi'\right|^2 + |\phi|^2 \,\mathrm{d}x \\ \text{s.t.} \quad & \mathcal{Q}_1\{v\} := \int \left|v''\right|^2 - \left|v'\right|^2 + \left(\frac{1}{3}\phi' - \frac{1}{2}\right) v^2 \,\mathrm{d}x \geq 0 \qquad \forall v \in H_{\mathrm{p,o}}, \\ & \mathcal{Q}_2\{w\} := \int \left|w''\right|^2 - \left|w'\right|^2 + \phi' \, w^2 \,\mathrm{d}x \geq 0 \qquad \forall w \in H_{\mathrm{p,o}}. \end{aligned} \tag{3.15}$$

It is an exercise in the calculus of variations to show that the objective function of this variational problem is strictly convex, meaning that its second variation with respect to ϕ is a positive definite functional. Consequently, the optimal value of (3.15) is attained by a unique optimal background field. The rest of this chapter will focus on computing the optimal ϕ for a given domain half-size ℓ using the time-marching method employed by Wen *et al.* (2013, 2015) to solve optimal background field problems arising in fluid dynamics.

3.2 A problem with spectral constraints

In the language of the background method, (3.15) is a variational problem with two *spectral constraints*. This terminology reflects the fact that each constraint requires a ϕ-dependent linear operator to have non-negative eigenvalues. In fact, the constraints in (3.15) can be replaced by

$$\inf_{\substack{v \in H_{\mathrm{p,o}} \\ \|v\|_2 = 1}} \mathcal{Q}_1\{v\} = \inf_{\substack{v \in H_{\mathrm{p,o}} \\ \|v\|_2 = 1}} \int \left|v''\right|^2 - \left|v'\right|^2 + \left(\frac{1}{3}\phi' - \frac{1}{2}\right) v^2 \,\mathrm{d}x \geq 0, \tag{3.16a}$$

$$\inf_{\substack{w \in H_{\mathrm{p,o}} \\ \|w\|_2 = 1}} \mathcal{Q}_2\{w\} = \inf_{\substack{w \in H_{\mathrm{p,o}} \\ \|w\|_2 = 1}} \int \left|w''\right|^2 - \left|w'\right|^2 + \phi' \, w^2 \,\mathrm{d}x \geq 0. \tag{3.16b}$$

The restriction to unit-norm functions is justified because $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are homogeneous functionals and $\mathcal{Q}_1\{0\} = \mathcal{Q}_2\{0\} = 0$. Note also that $\phi' \in L^\infty(-\ell, \ell)$ because $\phi \in H_{\mathrm{p,o}}$

implies that the background field must be at least continuously differentiable on $(-\ell, \ell)$. Then, the following result applies.

Theorem 3.1. *Let $f \in L^\infty(-\ell, \ell)$ be periodic, let D be the linear operator defined by $Du := \partial^4 u + \partial^2 u + f\,u$, and consider the eigenvalue problem*

$$\begin{cases} Du = \sigma u, & x \in (-\ell, \ell), \\ u(-x) = -u(x), & x \in (-\ell, \ell), \\ \partial^\alpha u(-\ell) = \partial^\alpha u(\ell), & \alpha \in \{0,\, 1,\, 2,\, 3\}. \end{cases} \tag{3.17}$$

(i) The eigenvalues of D are real and form at most a countable ordered sequence, $\{\sigma_k\}_{k\geq 0}$, such that $\sigma_0 > -\infty$ and $\lim_{k\to+\infty} \sigma_k = +\infty$.

(ii) The minimum eigenvalue σ_0 satisfies

$$\sigma_0 = \min_{\substack{u\in H_{\mathrm{p,o}} \\ \|u\|_2=1}} \int_{-\ell}^{\ell} \left|u''\right|^2 - \left|u'\right|^2 + f\,u^2 \,\mathrm{d}x. \tag{3.18}$$

Proof. See appendix A.1. □

Theorem 3.1 guarantees that the infima in (3.16a) and (3.16b) are achieved and correspond, respectively, to the minimum eigenvalues of the eigenvalue problems

$$\partial^4 v + \partial^2 v + \left(\frac{1}{3}\partial\phi - \frac{1}{2}\right) v = \lambda v, \tag{3.19a}$$

$$\partial^4 w + \partial^2 w + \partial\phi\, w = \mu w. \tag{3.19b}$$

The constraints in (3.15), therefore, require that the eigenvalues of the linear operators on the left-hand sides of (3.19a) and (3.19b) are non-negative. Using established terminology, the minimum eigenvalues λ_0 and μ_0 will be referred to as *ground-state eigenvalues*, and the corresponding eigenfunctions v_0 and w_0 will be referred to as *ground-state eigenfunctions*.

Remark 3.1. In addition to clarifying the nature of the constraints in (3.15), the eigenvalue problems (3.19a) and (3.19b) provide an easy way to test if a candidate background field is feasible. Indeed, one can simply employ one's preferred numerical scheme to compute the ground-state eigenvalues, and check their sign.

Remark 3.2. In section 3.4 it will be important to consider the multiplicity of the ground-state eigenvalue for problems (3.19a) and (3.19b). A bound on this can be obtained by noticing that if the background field ϕ is odd and periodic, then solving (3.19a) and (3.19b) for odd and periodic eigenfunctions on the domain $[-\ell, \ell]$ is the same as solving them on $[0, \ell]$

with homogeneous Dirichlet conditions on v, w, and their second derivatives. In fact, the second derivative of an odd, periodic function is so too, and any function that is both odd and periodic on $[-\ell, \ell]$ must vanish at $x = 0$ and at $x = \ell$. Then, a result by Everitt (1957) on fourth-order Sturm–Liouville operators guarantees that each eigenvalue of problems (3.19a) and (3.19b) is repeated at most twice, and one has

$$-\infty < \lambda_0 \leq \lambda_1 < \lambda_2 \leq \lambda_3 < \dots \quad \text{and} \quad -\infty < \mu_0 \leq \mu_1 < \mu_2 \leq \mu_3 < \dots . \tag{3.20}$$

3.3 The original time-marching optimisation method

Background fields that satisfy the spectral constraints can be constructed analytically (see for instance Collet *et al.*, 1993) and, as explained in remark 3.2, the feasibility of a candidate background field can be tested by solving the eigenvalue problems (3.19a) and (3.19b). The challenge, both theoretical and computational, is to optimise the background field and solve the full variational problem (3.15). This section will attempt to do so using a time-marching optimisation method proposed by Wen *et al.* (2013, 2015). Note that this method is described only through examples, from which a general approach must be inferred.

3.3.1 Formulation of a time-dependent problem

The starting point of the method proposed by Wen *et al.* (2013, 2015) is the derivation of suitable Euler–Lagrange (EL) equations characterising the optimal background field. The optimal value $p^\star$ of (3.15) satisfies

$$p^\star = \min_{\phi} \max_{v,w \in H_{\mathrm{p,o}}} \mathcal{L}\{\phi, v, w\}, \tag{3.21}$$

where

$$\mathcal{L}\{\phi, v, w\} := \int \frac{8}{3} \left|\phi''\right|^2 - \frac{8}{3} \left|\phi'\right|^2 + \phi^2 \,\mathrm{d}x - \mathcal{Q}_1\{v\} - \mathcal{Q}_2\{w\} \tag{3.22}$$

is the Lagrangian of the problem. Non-negative Lagrange multipliers for the constraints $\mathcal{Q}_1\{v\} \geq 0$ and $\mathcal{Q}_2\{w\} \geq 0$ are not necessary when forming the Lagrangian: they can always be eliminated by rescaling v and w, taking advantage of the fact that $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are homogeneous functionals. Moreover, under the reasonable assumption that strong duality holds for (3.15), the order of minimisation/maximisation in (3.21) is unimportant and the min-max procedure can be carried out simultaneously (cf. section 2.1).[2] Then, the optimal background field ϕ and the optimal functions v, w are a saddle point of the Lagrangian,

[2] The assumption of strong duality is never mentioned explicitly by Wen *et al.* (2013, 2015), but it seems essential to ensure that minimisation over ϕ and maximisation over v and w can be carried out concurrently.

and the characterising EL equations are found upon setting to zero the variations of $\mathcal{L}$ with respect to ϕ, v, and w. One obtains

$$\frac{1}{2}\frac{\delta\mathcal{L}}{\delta\phi} := \frac{8}{3}\partial^4\phi + \frac{8}{3}\partial^2\phi + \phi + \frac{1}{3}v\,\partial v + w\,\partial w = 0, \tag{3.23a}$$

$$\frac{1}{2}\frac{\delta\mathcal{L}}{\delta v} := -\partial^4 v - \partial^2 v - \left(\frac{1}{3}\partial\phi - \frac{1}{2}\right)v = 0, \tag{3.23b}$$

$$\frac{1}{2}\frac{\delta\mathcal{L}}{\delta w} := -\partial^4 w - \partial^2 w - \partial\phi\, w = 0. \tag{3.23c}$$

These equations are to be solved on $[-\ell, \ell]$ for odd, periodic functions ϕ, v and w, whilst ensuring that ϕ also satisfies the spectral constraints.

The key idea put forward by Wen *et al.* (2013, 2015) at this stage is to drop the spectral constraints, consider all variables to be time-dependent, add appropriate time derivatives to (3.23a)–(3.23c), and solve the resulting time-dependent equations starting from suitable initial guesses for ϕ, v, and w until convergence to a steady state (assuming that this occurs). Such a steady-state solution clearly satisfies (3.23a)–(3.23c) and one obtains the optimal solution of (3.15) if the steady-state background field ϕ satisfies the spectral constraints.[3]

To construct the time-dependent version of (3.23a)–(3.23c) one considers the rate of change of the Lagrangian (3.22) over time. In fact, when time dependence is introduced one has

$$\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}t} = \int \frac{\delta\mathcal{L}}{\delta\phi}\,\partial_t\phi + \frac{\delta\mathcal{L}}{\delta v}\,\partial_t v + \frac{\delta\mathcal{L}}{\delta w}\,\partial_t w\,\mathrm{d}x. \tag{3.24}$$

Recalling (3.21), to minimise $\mathcal{L}$ with respect to ϕ and maximise it over v and w one should choose $\partial_t\phi = -\alpha\frac{\delta\mathcal{L}}{\delta\phi}$, $\partial_t v = \beta\frac{\delta\mathcal{L}}{\delta v}$, and $\partial_t w = \gamma\frac{\delta\mathcal{L}}{\delta w}$ for some $\alpha, \beta, \gamma > 0$, where the relative values of these parameters determine the importance of optimising $\mathcal{L}$ with respect to each of its arguments. For instance, setting $\alpha \gg \beta, \gamma$ would prioritise minimisation over ϕ. The choice $\alpha = \beta = \gamma = 1/2$ is made here for simplicity, yielding the time-dependent equations

$$\partial_t\phi + \frac{8}{3}\partial_x^4\phi + \frac{8}{3}\partial_x^2\phi + \phi + \frac{1}{3}v\,\partial_x v + w\,\partial_x w = 0, \tag{3.25a}$$

$$\partial_t v + \partial_x^4 v + \partial_x^2 v + \left(\frac{1}{3}\partial_x\phi - \frac{1}{2}\right)v = 0, \tag{3.25b}$$

$$\partial_t w + \partial_x^4 w + \partial_x^2 w + \partial_x\phi\, w = 0. \tag{3.25c}$$

The linear terms in these equations are the same as in the KS equation (3.1) with the addition of the stabilising term ϕ in (3.25a), and the nonlinear terms have a similar structure to the

[3]This can be guaranteed for the variational problems considered by Wen *et al.* (2015), but a general result is not available. One may of course doubt that convergence to the correct solution is generic, and one would be right: the results presented in this section demonstrate that solving the time-dependent version of (3.23a)–(3.23c) does not yield the optimal solution of (3.15).

nonlinearity of the KS equation. It is therefore reasonable to expect that, as for the KS equation, the solution of (3.25a)–(3.25c) on $[-\ell, \ell]$ with periodic BCs is unique for given ICs. Under this assumption, it is not difficult to verify that when $\phi(x,0)$, $v(x,0)$, and $w(x,0)$ are odd, the functions $\phi(x,t)$, $v(x,t)$, and $w(x,t)$ remain odd at all instants in time. Thus, to compute odd and periodic steady states it suffices to consider ICs that are so.

3.3.2 Implementation and results

In order to solve (3.25a)–(3.25c) numerically, the fields ϕ, v and w were discretised using the N-dimensional sine series expansions

$$\begin{bmatrix} \phi(x,t) \\ v(x,t) \\ w(x,t) \end{bmatrix} = \sum_{n=1}^{N} \frac{1}{\sqrt{\ell}} \begin{bmatrix} \hat{\phi}_n(t) \\ \hat{v}_n(t) \\ \hat{w}_n(t) \end{bmatrix} \sin\left(\frac{n\pi x}{\ell}\right). \tag{3.26}$$

Substituting these expansions into (3.25a)–(3.25c), multiplying by $\ell^{-1/2}\sin(m\pi\ell^{-1}x)$ for each $m = 1, \ldots, N$ in turn, and integrating over $[-\ell, \ell]$ yields a system of N ordinary differential equations of the form $\frac{\mathrm{d}\boldsymbol{a}}{\mathrm{d}t} = \boldsymbol{f}(\boldsymbol{a})$, where the state vector $\boldsymbol{a}$ contains the expansion coefficients in (3.26). For simplicity, these were solved in MATLAB using the built-in function `ode113` for $\ell = 6\pi$, 8π, 10π, and 12π using $N = 100$; in all cases the numerical results change by less than 1% when N is increased to 120. To speed up the computation of steady states, `ode113` was stopped if $\|\boldsymbol{f}(\boldsymbol{a})\| \leq 0.1$ and Newton's iterations were employed to find a zero of the vector field $\boldsymbol{f}(\boldsymbol{a})$. Finally, the eigenvalue problems (3.19a) and (3.19b) were solved using a Galerkin projection method based on the expansions (3.26) to check whether the steady-state background field satisfied the spectral constraints. Results were qualitatively similar for all tested values of ℓ, so only those obtained with $\ell = 10\pi$ are presented below for the sake of brevity.

Three sets of ICs, illustrated in figure 3.1, were considered. The first, referred to as IC1, consisted of unit-amplitude sinusoidal ICs $\phi(x,0) = v(x,0) = w(x,0) = \sin(\pi\ell^{-1}x)$. The other two sets of ICs, referred to as IC2 and IC3, were randomly generated profiles of the form

$$\begin{bmatrix} \phi(x,0) \\ v(x,0) \\ w(x,0) \end{bmatrix} = \sum_{n=1}^{20} \frac{1}{\sqrt{\ell}} \begin{bmatrix} \hat{\phi}_n \\ \hat{v}_n \\ \hat{w}_n \end{bmatrix} \sin\left(\frac{n\pi x}{\ell}\right), \tag{3.27}$$

with expansion coefficients drawn from the uniform distribution on $[-1, 1]$. This choice was made so no "preferential" Fourier mode is artificially enforced by the ICs.

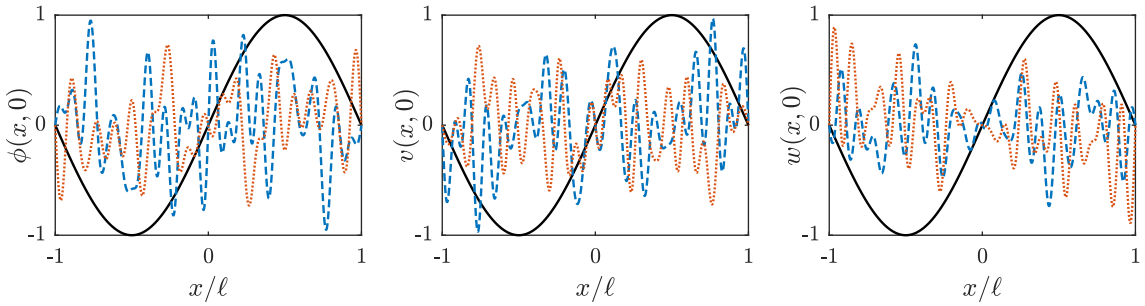


FIGURE 3.1: Initial conditions $\phi(x,0)$, $v(x,0)$, and $w(x,0)$ used to solve (3.25a)–(3.25c) with $\ell = 10\pi$. IC1: unit-amplitude sinusoidal initial condition (——). IC2: random initial condition of the form (3.27) (- - -). IC3: random initial condition of the form (3.27) (·····).

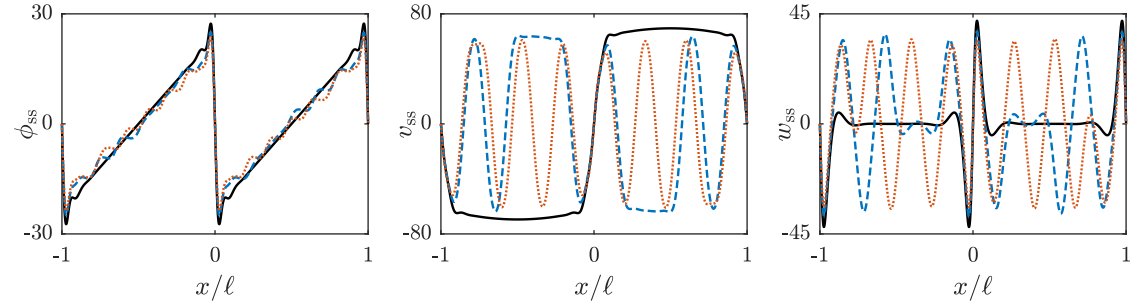


FIGURE 3.2: Steady-state solutions ϕ_{ss}, v_{ss}, and w_{ss} of (3.25a)–(3.25c) for $\ell = 10\pi$, computed using initial conditions IC1 (——), IC2 (- - -), and IC3 (·····).

The steady state fields $\phi_{ss}(x)$, $v_{ss}(x)$, and $w_{ss}(x)$ obtained using initial conditions IC1, IC2, and IC3 are plotted in figure 3.2. It is immediately apparent that the time-marching solution method described above suffers from lack of robustness: different ICs yield different steady states. In addition, table 3.1 reveals that none of the computed steady-state background fields corresponds to the optimal solution of (3.15) because in all cases at least one of problems (3.19a) and (3.19b) has negative eigenvalues, meaning that at least one of the spectral constraints is violated. (Note, however, that the spectral constraint $\mathcal{Q}_2\{w\} \geq 0$ can be considered satisfied for initial conditions IC1 and IC2 since all tabulated eigenvalues μ_n are either positive or zero within reasonable numerical tolerances.)

Convergence to "spurious" solutions of the EL equations (3.23a)–(3.23c), which do not satisfy the spectral constraints, is in stark contrast with the successful application of the time-marching solution method reported by Wen *et al.* (2013, 2015). Such spurious solutions are provably linearly unstable in the examples they considered. This is not the case here and it can be confirmed by linearising (3.25a)–(3.25c) around each of the computed steady-state solutions. Upon considering infinitesimal normal-mode perturbations from the steady states of the form $\tilde{\phi}(x)\mathrm{e}^{\zeta t}$, $\tilde{v}(x)\mathrm{e}^{\zeta t}$, and $\tilde{w}(x)\mathrm{e}^{\zeta t}$, with $\zeta \in \mathbb{C}$, equations (3.25a)–(3.25c) become a

TABLE 3.1: First 10 eigenvalues for (3.19a) and (3.19b) when $\phi = \phi_{\text{ss}}$ is the steady state computed with each of the three sets of initial conditions IC1, IC2, and IC3. Negative values of λ_n and μ_n indicate violation of the spectral constraints $\mathcal{Q}_1\{v\} \geq 0$ and $\mathcal{Q}_2\{w\} \geq 0$, respectively.

	ϕ_{ss} from IC1		ϕ_{ss} from IC2		ϕ_{ss} from IC3	
n	λ_n	μ_n	λ_n	μ_n	λ_n	μ_n
0	-2.38×10^{-1}	-4.39×10^{-8}	-6.13×10^{-1}	-2.12×10^{-15}	-6.35×10^{-1}	-1.25×10^{-2}
1	-2.27×10^{-1}	-1.88×10^{-15}	-5.12×10^{-1}	1.07×10^{-4}	-5.64×10^{-1}	8.12×10^{-15}
2	-2.02×10^{-1}	1.27	-4.59×10^{-1}	7.97×10^{-2}	-4.89×10^{-1}	3.83×10^{-2}
3	-1.59×10^{-1}	1.28	-2.57×10^{-1}	3.14×10^{-1}	-4.02×10^{-1}	1.58×10^{-1}
4	-1.34×10^{-1}	1.33	-1.41×10^{-1}	3.66×10^{-1}	-2.75×10^{-1}	3.53×10^{-1}
5	-9.75×10^{-2}	1.36	-9.16×10^{-2}	6.06×10^{-1}	-1.83×10^{-1}	5.24×10^{-1}
6	-4.70×10^{-2}	1.42	-1.91×10^{-2}	1.38	-7.89×10^{-15}	5.80×10^{-1}
7	-1.30×10^{-2}	1.44	3.07×10^{-15}	1.39	3.91×10^{-3}	7.50×10^{-1}
8	-1.63×10^{-16}	1.49	6.11×10^{-2}	1.62	4.94×10^{-2}	1.54
9	2.74×10^{-2}	1.58	1.98×10^{-1}	1.71	1.92×10^{-1}	1.79

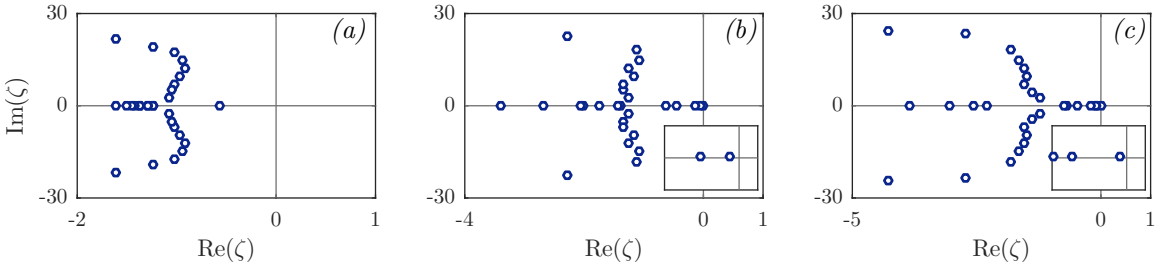


FIGURE 3.3: Eigenvalues of the linear stability problem (3.28a)–(3.28c) (only the 20 eigenvalues with largest real part are shown). Panels *(a)–(c)* refer to the steady states computed with initial conditions IC1, IC2, and IC3, respectively. Inserts in *(b)* and *(c)* show a detailed view of the region $-0.1 \leq \text{Re}(\zeta) \leq 0.025$, $-0.1 \leq \text{Im}(\zeta) \leq 0.1$. All eigenvalues shown lie in the open left half-plane. Solid lines indicate the real and imaginary axes.

linear eigenvalue problem of the form

$$\frac{8}{3}\partial^4\tilde{\phi} + \frac{8}{3}\partial^2\tilde{\phi} + \tilde{\phi} + \frac{1}{3}v_{\text{ss}}\,\partial\tilde{v} + \frac{1}{3}\tilde{v}\,\partial v_{\text{ss}} + w_{\text{ss}}\,\partial\tilde{w} + \tilde{w}\,\partial w_{\text{ss}} = -\zeta\tilde{\phi}, \tag{3.28a}$$

$$\partial^4\tilde{v} + \partial^2\tilde{v} + \frac{1}{3}\tilde{v}\,\partial\phi_{\text{ss}} + \frac{1}{3}v_{\text{ss}}\,\partial\tilde{\phi} - \frac{1}{2}\tilde{v} = -\zeta\tilde{v}, \tag{3.28b}$$

$$\partial^4\tilde{w} + \partial^2\tilde{w} + \partial\phi_{\text{ss}}\,\tilde{w} + w_{\text{ss}}\,\partial\tilde{\phi} = -\zeta\tilde{w}. \tag{3.28c}$$

The corresponding steady states ϕ_{ss}, v_{ss}, and w_{ss} are linearly unstable if there exists an eigenvalue ζ with $\text{Re}(\zeta) > 0$, while they are linearly stable if $\text{Re}(\zeta) < 0$ for all eigenvalues. Figure 3.3 illustrates the 20 eigenvalues with largest real part obtained for each of the three different steady states plotted in figure 3.2 (pairs of complex conjugate eigenvalues are counted as a one). In each case, all eigenvalues have strictly negative real part, confirming that each of the computed steady-state solutions of (3.25a)–(3.25c) is linearly stable.

It should be remarked that the existence of steady solutions that are linearly stable but do not satisfy at least one of the spectral constraints does not mean that solving the time-dependent equations (3.25a)–(3.25c) may never return the optimal solution of problem (3.15). Indeed, if the corresponding solution of the EL equations were a linearly stable and locally attracting equilibrium for (3.25a)–(3.25c), it would suffice to consider an IC within its local basin of attraction. However, proving that the desired unknown steady state is a local attractor may not be straightforward. Moreover, even if it were possible to prove at least linear stability, estimating the local basin of attraction of an unknown equilibrium seems a formidable challenge. In practice, therefore, the existence of linearly stable but spurious solutions of the EL equations prohibits a successful application of the time-marching solution method described above. Fortunately, a simple modification discussed in the next section allows for the resolution of this problem.

3.4 A modified time-marching optimisation method

One possible way to resolve the failure of the time-marching solution method described in the previous section is suggested by an informal analogy between problem (3.15) and optimisation problems with LMIs (cf. chapter 2). This is motivated by the fact that both a spectral constraint and an LMI require that the eigenvalues of a certain linear operator are non-negative, the difference being the dimension of the space on which this linear operator is defined. It is therefore useful to consider the finite-dimensional equivalent of (3.15), meaning an LMI-constrained optimisation problem of the form

$$\begin{aligned} \min_{\boldsymbol{y}\in\mathbb{R}^m} \quad & c(\boldsymbol{y}) \\ \text{s.t.} \quad & \boldsymbol{F}_1(\boldsymbol{y}) \succeq 0, \\ & \boldsymbol{F}_2(\boldsymbol{y}) \succeq 0, \end{aligned} \tag{3.29}$$

where the matrices $\boldsymbol{F}_1(\boldsymbol{y}), \boldsymbol{F}_2(\boldsymbol{y}) \in \mathbb{S}^n$ depend affinely on $\boldsymbol{y} \in \mathbb{R}^m$ and $c : \mathbb{R}^m \to \mathbb{R}$ is a convex cost function. The exact form of c will not be important in the following discussion, but to make the analogy with (3.15) evident one may consider $c(\boldsymbol{y}) = \boldsymbol{y}^\mathsf{T}\boldsymbol{C}\boldsymbol{y}$ for some positive definite matrix $\boldsymbol{C} \in \mathbb{S}^m$.

Suppose that strong duality holds for (3.29) and that the minimum cost is achieved by an optimal solution $\boldsymbol{y}^\star$. Suppose also that one can determine *a priori* that each $\boldsymbol{F}_i(\boldsymbol{y}^\star)$, $i = 1, 2$, has at most r zero eigenvalues, the other ones being positive. Then, the rank of each $\boldsymbol{F}_i(\boldsymbol{y}^\star)$, $i = 1, 2$, is no less than $n - r$, and an argument similar to that described in

section 2.5 shows that the Lagrangian for (3.29) can be written in the low-rank form

$$L(\boldsymbol{y},\, \boldsymbol{v}_1,\, \ldots,\, \boldsymbol{v}_r, \boldsymbol{w}_1,\, \ldots,\, \boldsymbol{w}_r) = c(\boldsymbol{y}) - \sum_{i=1}^{r} \boldsymbol{v}_i^{\mathsf{T}} \boldsymbol{F}_1(\boldsymbol{y}) \boldsymbol{v}_i - \sum_{i=1}^{r} \boldsymbol{w}_i^{\mathsf{T}} \boldsymbol{F}_2(\boldsymbol{y}) \boldsymbol{w}_i. \tag{3.30}$$

In this expression, the vectors $\boldsymbol{v}_i$, …, $\boldsymbol{v}_r$ and $\boldsymbol{w}_i$, …, $\boldsymbol{w}_r$ are dual variables, with respect to which the Lagrangian should be maximised.

Just like (3.29) is the finite-dimensional equivalent of (3.15), when $r = 1$ the Lagrangian function (3.30) is the finite-dimensional equivalent of the Lagrangian (3.22). To see this more clearly, let D_1 and D_2 be the ϕ-dependent differential operators on the left-hand sides of (3.19a) and (3.19b), respectively. In other words, D_1 and D_2 satisfy

$$D_1 v = \partial^4 v + \partial^2 v + \left(\frac{1}{3}\partial\phi - \frac{1}{2}\right) v, \qquad D_2 w = \partial^4 w + \partial^2 w + \partial\phi\, w. \tag{3.31}$$

Then, integration by parts using periodicity yields

$$\mathcal{Q}_1\{v\} = \int v\, D_1 v \,\mathrm{d}x, \qquad \mathcal{Q}_2\{w\} = \int w\, D_2 w \,\mathrm{d}x, \tag{3.32}$$

so when $r = 1$ the terms $\mathcal{Q}_1\{v\}$ and $\mathcal{Q}_2\{w\}$ in (3.22) are, respectively, the infinite-dimensional version of the matrix-vector products $\boldsymbol{v}_1^{\mathsf{T}} \boldsymbol{F}_1(\boldsymbol{y}) \boldsymbol{v}_1$ and $\boldsymbol{w}_1^{\mathsf{T}} \boldsymbol{F}_2(\boldsymbol{y}) \boldsymbol{w}_1$ in (3.30).

This correspondence suggests that considering (3.22) as the Lagrangian for the variational problem (3.15) is tantamount to assuming that, when ϕ is the optimal background field, the operators D_1 and D_2 have at most one zero eigenvalue (all others being strictly positive). Since the optimality conditions (3.23b) and (3.23c) imply that D_1 and D_2 must have at least one zero eigenvalue when ϕ is the optimal background field, if one considers (3.22), then one is implicitly assuming that the ground-state eigenvalue of D_1 and D_2 for the optimal ϕ is a simple zero. This assumption, however, is unjustified: the eigenvalues of D_1 and D_2 may have multiplicity 2 (cf. remark 3.2 in section 3.2) and there is no reason to exclude that, for the optimal ϕ, the ground-state eigenvalues are repeated zeros.

These observations indicate that the failure of the time-marching optimisation method described in section 3.3 may be due to the Lagrangian (3.22) being incorrect, in the sense that it is not the most general Lagrangian satisfying (3.21). Instead, an analogy with (3.30) for $r = 2$ motivates one to consider a Lagrangian with two "copies" of each spectral constraint (as many as the largest possible multiplicity of the associated ground-state eigenvalues), *i.e.*,

$$\mathcal{L}\{\phi, v_1, v_2, w_1, w_2\} := \int \frac{8}{3}\left|\phi''\right|^2 - \frac{8}{3}\left|\phi'\right|^2 + \phi^2 \,\mathrm{d}x - \sum_{i=1}^{2} \mathcal{Q}_1\{v_i\} - \sum_{i=1}^{2} \mathcal{Q}_2\{w_i\}. \tag{3.33}$$

3.4.1 Modified time-dependent Euler–Lagrange equations

The EL equations characterising the optimal background field can be found by setting to zero the variations of the Lagrangian with respect to all of its arguments. When the modified Lagrangian (3.33) is considered, one obtains

$$\frac{8}{3}\partial^4\phi + \frac{8}{3}\partial^2\phi + \phi + \frac{1}{3}\left(v_1\,\partial v_1 + v_2\,\partial v_2\right) + w_1\,\partial w_1 + w_2\,\partial w_2 = 0, \tag{3.34a}$$

$$-\partial^4 v_1 - \partial^2 v_1 - \left(\frac{1}{3}\partial\phi - \frac{1}{2}\right) v_1 = 0, \tag{3.34b}$$

$$-\partial^4 v_2 - \partial^2 v_2 - \left(\frac{1}{3}\partial\phi - \frac{1}{2}\right) v_2 = 0, \tag{3.34c}$$

$$-\partial^4 w_1 - \partial^2 w_1 - \partial\phi\, w_1 = 0, \tag{3.34d}$$

$$-\partial^4 w_2 - \partial^2 w_2 - \partial\phi\, w_2 = 0. \tag{3.34e}$$

The corresponding time-dependent equations can be formulated by considering the total time derivative of the Lagrangian, recalling that this must be minimised with respect to ϕ, and maximised with respect to v_1, v_2, w_1, and w_2. One obtains

$$\partial_t\phi + \frac{8}{3}\partial_x^4\phi + \frac{8}{3}\partial_x^2\phi + \phi + \frac{1}{3}\left(v_1\,\partial_x v_1 + v_2\,\partial_x v_2\right) + w_1\,\partial_x w_1 + w_2\,\partial_x w_2 = 0, \tag{3.35a}$$

$$\partial_t v_1 + \partial_x^4 v_1 + \partial_x^2 v_1 + \left(\frac{1}{3}\partial_x\phi - \frac{1}{2}\right) v_1 = 0, \tag{3.35b}$$

$$\partial_t v_2 + \partial_x^4 v_2 + \partial_x^2 v_2 + \left(\frac{1}{3}\partial_x\phi - \frac{1}{2}\right) v_2 = 0, \tag{3.35c}$$

$$\partial_t w_1 + \partial_x^4 w_1 + \partial_x^2 w_1 + \partial_x\phi\, w_1 = 0, \tag{3.35d}$$

$$\partial_t w_2 + \partial_x^4 w_2 + \partial_x^2 w_2 + \partial_x\phi\, w_2 = 0. \tag{3.35e}$$

As in section 3.3.1, it is reasonable to assume that (3.35a)–(3.35e) have a unique solution when solved on $[-\ell, \ell]$ with periodic BCs and odd ICs. One can therefore compute odd and periodic stationary solutions by solving (3.35a)–(3.35e) with odd and periodic ICs.

3.4.2 Implementation and results

Equations (3.35a)–(3.35e) were solved for $\ell = 10\pi$ using the same numerical setup described in section 3.3.2 (computations for $\ell = 6\pi$, 8π, and 12π gave similar results and are not reported for brevity). Three different sets of ICs were considered, which are illustrated in figure 3.4. The first, denoted IC1 to parallel the notation of section 3.3, consisted of one-wavenumber sinusoidal ICs, $\phi(x,0) = v_1(x,0) = w_1(x,0) = \sin(\pi\ell^{-1}x)$ and $v_2(x,0) = w_2(x,0) = \sin(2\pi\ell^{-1}x)$. The ICs $v_2(x,0)$ and $w_2(x,0)$ must be linearly independent of

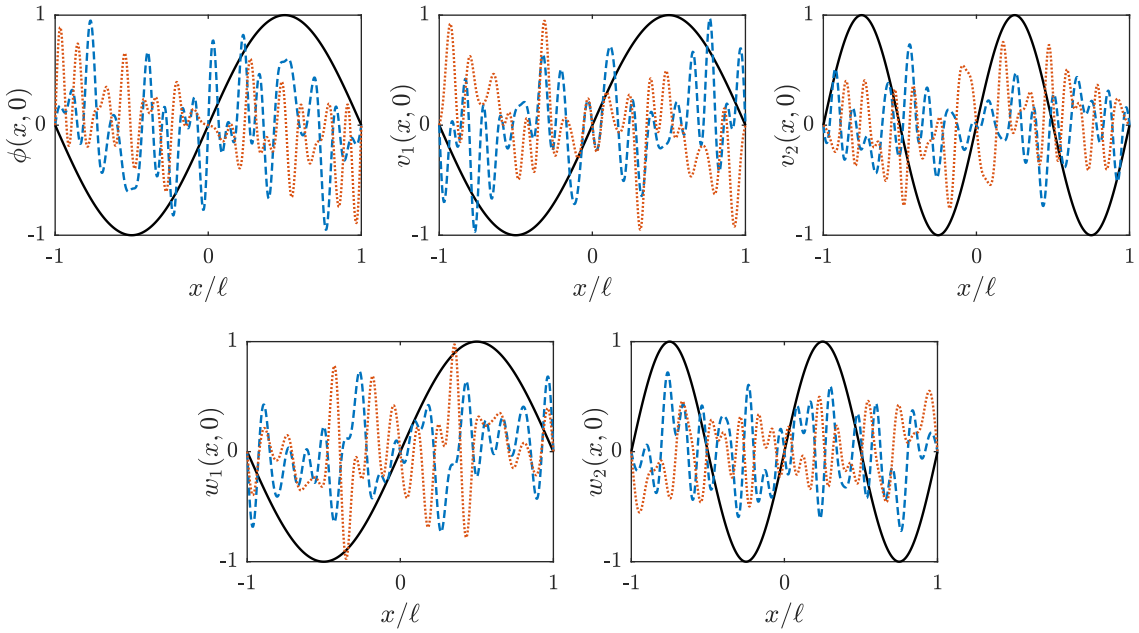


FIGURE 3.4: Initial conditions $\phi(x,0)$, $v_1(x,0)$, $v_2(x,0)$, $w_1(x,0)$, and $w_2(x,0)$ used to solve (3.35a)–(3.35e) with $\ell = 10\pi$. IC1: single-wavenumber sinusoidal initial condition (——). IC2: random initial condition of the form (3.36) (- - -). IC3: random initial condition of the form (3.36) (·····).

$v_1(x,0)$ and $w_1(x,0)$, respectively, otherwise (3.35a)–(3.35e) reduce to (3.25a)–(3.25c) and there would be no difference between the present computations and those of section 3.3. However, linear independence need not be maintained for $t > 0$. The other two sets of ICs, referred to as IC2 and IC3, consisted of random profiles of the form

$$\begin{bmatrix} \phi(x,0) \\ v_1(x,0) \\ v_2(x,0) \\ w_1(x,0) \\ w_2(x,0) \end{bmatrix} = \sum_{n=1}^{20} \frac{1}{\sqrt{\ell}} \begin{bmatrix} \hat{\phi}_n \\ \hat{v}_{1,n} \\ \hat{v}_{2,n} \\ \hat{w}_{1,n} \\ \hat{w}_{2,n} \end{bmatrix} \sin\left(\frac{n\pi x}{\ell}\right), \tag{3.36}$$

where, for the same reason given in section 3.3.2, the expansion coefficients were drawn from the uniform distribution on $[-1,1]$. The profiles generated for IC2 and IC3 satisfied the required linear independence conditions.

Figure 3.5 shows the steady states computed using each set of ICs. The difference with the results presented in section 3.3 is striking: although v_1, v_2, w_1, and w_2 converge to different steady states, the steady-state background fields ϕ_{ss} are the same in all cases (their Fourier coefficients agree up to 6 decimal places) and coincide with the "double shock" profile obtained by Fantuzzi & Wynn (2015). Inspection of the eigenvalues of (3.19a) and (3.19b),

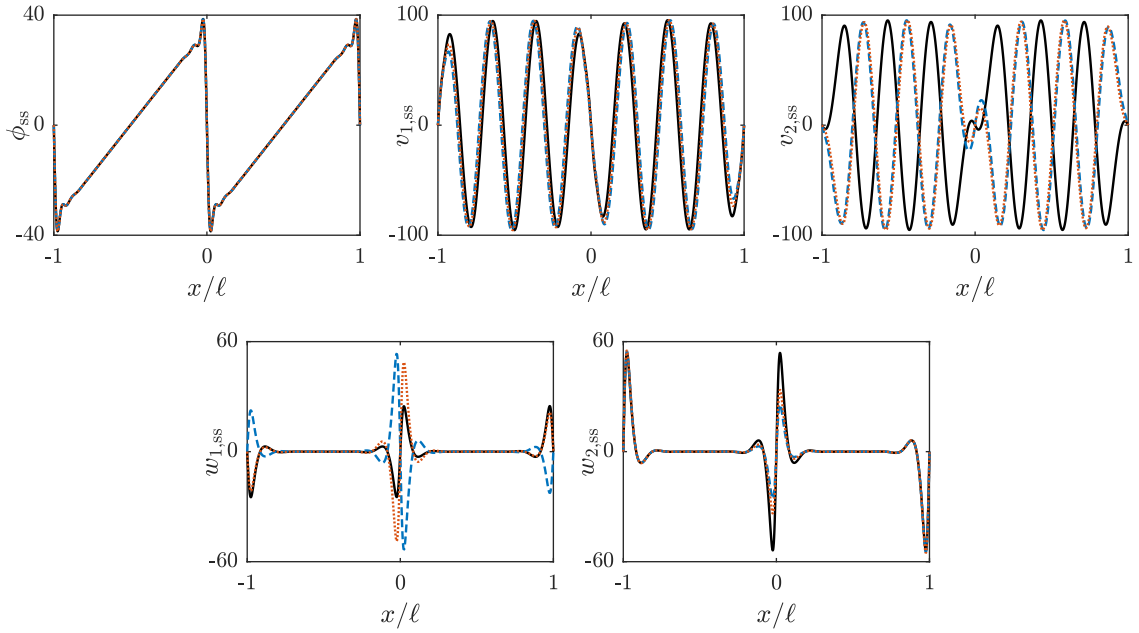


FIGURE 3.5: Steady-state solutions ϕ_{ss}, $v_{1,ss}$, $v_{2,ss}$, $w_{1,ss}$, and $w_{2,ss}$ of (3.35a)–(3.35e) for $\ell = 10\pi$ and initial conditions IC1 (——), IC2 (- - -), and IC3 (·····).

TABLE 3.2: First 10 eigenvalues for (3.19a) and (3.19b) when $\phi = \phi_{ss}$ is the steady state computed with each of the three sets of initial conditions IC1, IC2, and IC3 using the modified time-marching approach. Negative values of λ_n and μ_n indicate violation of the spectral constraints $\mathcal{Q}_1\{v\} \geq 0$ and $\mathcal{Q}_2\{w\} \geq 0$, respectively.

	ϕ_{ss} from IC1		ϕ_{ss} from IC2		ϕ_{ss} from IC3	
n	λ_n	μ_n	λ_n	μ_n	λ_n	μ_n
0	-5.75×10^{-14}	-3.22×10^{-8}	-9.25×10^{-14}	-8.67×10^{-8}	-6.21×10^{-14}	-7.06×10^{-8}
1	7.86×10^{-14}	1.52×10^{-7}	3.24×10^{-14}	9.75×10^{-8}	9.06×10^{-14}	1.14×10^{-7}
2	2.34×10^{-2}	1.88	2.34×10^{-2}	1.88	2.34×10^{-2}	1.88
3	2.60×10^{-2}	1.88	2.60×10^{-2}	1.88	2.60×10^{-2}	1.88
4	8.29×10^{-2}	2.00	8.29×10^{-2}	2.00	8.29×10^{-2}	2.00
5	9.27×10^{-2}	2.04	9.27×10^{-2}	2.04	9.27×10^{-2}	2.04
6	1.48×10^{-1}	2.05	1.48×10^{-1}	2.05	1.48×10^{-1}	2.05
7	1.92×10^{-1}	2.11	1.92×10^{-1}	2.11	1.92×10^{-1}	2.11
8	2.35×10^{-1}	2.18	2.35×10^{-1}	2.18	2.35×10^{-1}	2.18
9	2.49×10^{-1}	2.18	2.49×10^{-1}	2.18	2.49×10^{-1}	2.18

reported in table 3.2, confirms that ϕ_{ss} satisfies both spectral constraints up to reasonable numerical tolerances, so it is the optimal solution of (3.15).

Table 3.2 reveals also that when ϕ is the optimal background field, the ground-state eigenvalues of both problems (3.19a) and (3.19b) are repeated zeros (within reasonable tolerances). It is therefore hardly surprising that the computations of section 3.3 did not solve the variational problem (3.15) correctly: as described at the beginning of this section, the Lagrangian (3.22) implicitly assumed simple ground-state eigenvalues. The Lagrangian (3.33), instead, allows for ground-state eigenvalues with multiplicity 2.

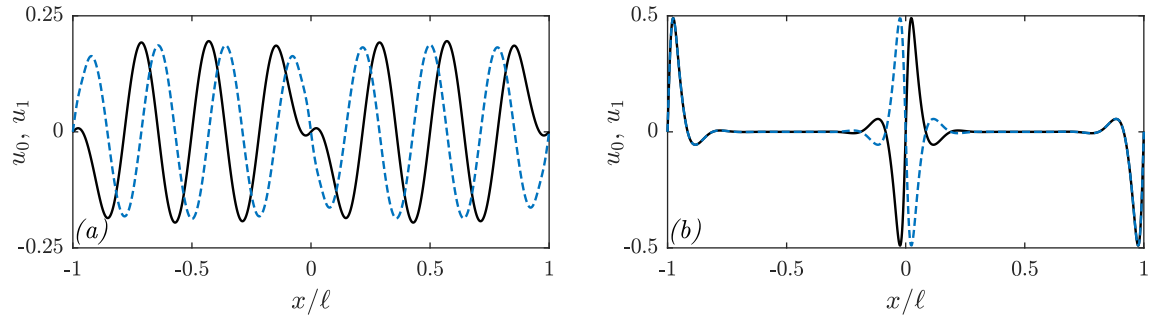


FIGURE 3.6: *(a)* Orthonormal ground-state eigenfunctions, denoted u_0 and u_1, of the eigenvalue problem (3.19a) when $\phi = \phi_{\text{ss}}$ is the optimal background field shown in figure 3.5. *(b)* Orthonormal ground-state eigenfunctions, also denoted u_0 and u_1, of the eigenvalue problem (3.19b) when $\phi = \phi_{\text{ss}}$ is the optimal background field shown in figure 3.5.

3.4.3 Non-uniqueness of steady states

The results presented in the previous section suggest that if the solution of (3.35a)–(3.35e) reaches a steady state, then the steady-state background field ϕ_{ss} is the optimal solution of (3.15). As discussed at the end of section 3.1, this is unique. On the other hand, figure 3.5 shows that different steady states $v_{1,\text{ss}}$, $v_{2,\text{ss}}$, $w_{1,\text{ss}}$, and $w_{2,\text{ss}}$ are possible, so solutions of (3.35a)–(3.35e) yielding the unique optimal ϕ are not unique themselves.

The existence of multiple steady solutions, all equally valid, is a consequence of the fact that, when ϕ is the optimal solution of (3.15), the ground-state eigenvalues for problems (3.19a) and (3.19b) are repeated zeros. For instance, to see why $v_{1,\text{ss}}$ and $v_{2,\text{ss}}$ are not unique, let u_0 and u_1 be the two orthonormal ground-state eigenfunctions for (3.19a), which are plotted in figure 3.6*(a)*. Since the corresponding ground-state eigenvalues are zero and the steady states $v_{1,\text{ss}}$ and $v_{2,\text{ss}}$ satisfy (3.34b) and (3.34c), there exist $\alpha_1, \ldots, \alpha_4 \in \mathbb{R}$ such that

$$v_{1,\text{ss}} = \alpha_1 u_0 + \alpha_2 u_1, \qquad v_{2,\text{ss}} = \alpha_3 u_0 + \alpha_4 u_1. \tag{3.37}$$

The contribution of $v_{1,\text{ss}}$ and $v_{2,\text{ss}}$ to (3.34a) can then be rewritten as

$$\frac{1}{3}\left(v_{1,\text{ss}}\,\partial v_{1,\text{ss}} + v_{2,\text{ss}}\,\partial v_{2,\text{ss}}\right) = \frac{1}{6}\partial\left(v_{1,\text{ss}}^2 + v_{2,\text{ss}}^2\right) = \frac{1}{6}\partial\left(A\,u_0^2 + 2\,B\,u_0\,u_1 + C\,u_1^2\right), \tag{3.38}$$

where $A := \alpha_1^2 + \alpha_3^2$, $B := \alpha_1\alpha_2 + \alpha_3\alpha_4$, and $C := \alpha_2^2 + \alpha_4^2$. An infinite family of equivalent steady states can therefore be constructed by varying the constants $\alpha_1, \ldots, \alpha_4$ in (3.37) whilst keeping the values A, B, and C constant. Table 3.3 demonstrates that this can indeed be done: the three steady-state profiles shown in figure 3.5, obtained with the three different sets of initial conditions IC1, IC2, and IC3, can be decomposed as in (3.37) with different choices of $\alpha_1, \ldots, \alpha_4$, all of which yield the same A, B, and C.

TABLE 3.3: Coefficients $\alpha_1, \ldots, \alpha_4$ and constants A, B, and C for the steady states $v_{1,\mathrm{ss}}$ and $v_{2,\mathrm{ss}}$ computed using each of the three sets of initial conditions IC1, IC2, and IC3.

	α_1	α_2	α_3	α_4	A	B	C
IC1	2.65	508.23	−487.07	2.77	237,245.75	4.72×10^{-8}	258,305.65
IC2	189.41	468.24	448.74	−197.64	237,245.75	2.25×10^{-8}	258,305.65
IC3	124.91	491.24	470.79	−130.34	237,245.75	-5.04×10^{-8}	258,305.65

TABLE 3.4: Coefficients $\beta_1, \ldots, \beta_4$ and constants D and E for the steady states $w_{1,\mathrm{ss}}$ and $w_{2,\mathrm{ss}}$ computed using each of the three sets of initial conditions IC1, IC2, and IC3.

	β_1	β_2	β_3	β_4	D	E
IC1	0.10	−50.61	109.82	0.04	14,623.00	-4.21×10^{-11}
IC2	−31.52	77.50	82.09	29.75	14,623.00	-1.96×10^{-11}
IC3	27.89	−71.60	90.74	22.01	14,623.00	5.02×10^{-11}

A very similar argument explains why the steady states $w_{1,\mathrm{ss}}$ and $w_{2,\mathrm{ss}}$ need not be unique. This time, let u_0 and u_1 be the two orthonormal ground-state eigenfunctions for (3.19b), which are illustrated in figure 3.6*(b)*. Observe that these eigenfunctions satisfy $|u_0(x)|^2 = |u_1(x)|^2$ pointwise. For any solution $w_{1,\mathrm{ss}}$ and $w_{2,\mathrm{ss}}$ of the EL equations (3.34c) and (3.34e), there exist real scalars $\beta_1, \ldots, \beta_4$ such that

$$w_{1,\mathrm{ss}} = \beta_1 u_0 + \beta_2 u_1, \qquad w_{2,\mathrm{ss}} = \beta_3 u_0 + \beta_4 u_1. \tag{3.39}$$

Recalling that $|u_0(x)|^2 = |u_1(x)|^2$, the contribution of $w_{1,\mathrm{ss}}$ and $w_{2,\mathrm{ss}}$ to (3.34a) becomes

$$w_{1,\mathrm{ss}}\,\partial w_{1,\mathrm{ss}} + w_{2,\mathrm{ss}}\,\partial w_{2,\mathrm{ss}} = \frac{1}{2}\partial\left(w_{1,\mathrm{ss}}^2 + w_{2,\mathrm{ss}}^2\right) = \frac{1}{2}\partial\left(D\,u_0^2 + 2\,E\,u_0\,u_1\right), \tag{3.40}$$

where $D := \sum_{i=1}^{4} \beta_i^2$ and $E := \beta_1\beta_2 + \beta_3\beta_4$. Then, an infinity of equally valid steady states can be constructed by varying $\beta_1, \ldots, \beta_4$ whilst keeping D and E constant. Table 3.4 confirms that the different profiles $w_{1,\mathrm{ss}}$ and $w_{2,\mathrm{ss}}$ plotted in figure 3.5, computed using initial conditions IC1, IC2, and IC3, can be constructed from (3.39) using different choices of $\beta_1, \ldots, \beta_4$, all resulting in the same values of D and E.

3.4.4 Linear stability of steady states

The ability to solve the variational problem (3.15) by evolving (3.35a)–(3.35e) from different sets of ICs is evidence that the time-marching method based on the Lagrangian (3.33) is robust. A proof of convergence to the correct steady state for any ICs seems difficult to obtain because equations (3.35a)–(3.35e) are nonlinear. A simpler task is to try to show that if a steady solution of (3.35a)–(3.35e) violates the spectral constraints, then it is not attracting

and, therefore, will not be computed in practice. Wen *et al.* (2015) proved statements of this kind for the examples considered in their work, and it is natural to wonder if their analysis can be adapted to the present case.

The argument proposed by Wen *et al.* (2015) relies on formulating a connection between the eigenvalue problems associated with the spectral constraints, (3.19a) and (3.19b), and the eigenvalue problem for the linear stability analysis of steady solutions of (3.35a)–(3.35e). The latter is derived by writing solutions of (3.35a)–(3.35e) as the sum of steady fields ϕ_{ss}, $v_{1,\mathrm{ss}}$, $v_{2,\mathrm{ss}}$, $w_{1,\mathrm{ss}}$ and $w_{2,\mathrm{ss}}$, which satisfy (3.34a)–(3.34e), and infinitesimal perturbations

$$\tilde{\phi}(x)\mathrm{e}^{\zeta t}, \qquad \tilde{v}_1(x)\mathrm{e}^{\zeta t}, \qquad \tilde{v}_2(x)\mathrm{e}^{\zeta t}, \qquad \tilde{w}_1(x)\mathrm{e}^{\zeta t}, \qquad \tilde{w}_2(x)\mathrm{e}^{\zeta t}, \tag{3.41}$$

where $\zeta \in \mathbb{C}$. Representing solutions in this way leads to the eigenvalue problem

$$\frac{8}{3}\partial^4\tilde{\phi} + \frac{8}{3}\partial^2\tilde{\phi} + \tilde{\phi} + \frac{1}{3}\partial\left(v_{1,\mathrm{ss}}\tilde{v}_1 + v_{2,\mathrm{ss}}\tilde{v}_2\right) + \partial\left(w_{1,\mathrm{ss}}\tilde{w}_1 + w_{2,\mathrm{ss}}\tilde{w}_2\right) = -\zeta\tilde{\phi}, \tag{3.42a}$$

$$\partial^4\tilde{v}_1 + \partial^2\tilde{v}_1 + \left(\frac{1}{3}\partial\phi_{\mathrm{ss}} - \frac{1}{2}\right)\tilde{v}_1 + \frac{1}{3}v_{1,\mathrm{ss}}\,\partial\tilde{\phi} = -\zeta\tilde{v}_1, \tag{3.42b}$$

$$\partial^4\tilde{v}_2 + \partial^2\tilde{v}_2 + \left(\frac{1}{3}\partial\phi_{\mathrm{ss}} - \frac{1}{2}\right)\tilde{v}_2 + \frac{1}{3}v_{2,\mathrm{ss}}\,\partial\tilde{\phi} = -\zeta\tilde{v}_2, \tag{3.42c}$$

$$\partial^4\tilde{w}_1 + \partial^2\tilde{w}_1 + \partial\phi_{\mathrm{ss}}\,\tilde{w}_1 + w_{1,\mathrm{ss}}\,\partial\tilde{\phi} = -\zeta\tilde{w}_1, \tag{3.42d}$$

$$\partial^4\tilde{w}_2 + \partial^2\tilde{w}_2 + \partial\phi_{\mathrm{ss}}\,\tilde{w}_2 + w_{2,\mathrm{ss}}\,\partial\tilde{\phi} = -\zeta\tilde{w}_2. \tag{3.42e}$$

To prove that a steady solution of (3.35a)–(3.35e) is linear unstable, hence not attracting, it suffices to find one eigenfunction for (3.42a)–(3.42e) with eigenvalue ζ satisfying $\mathrm{Re}(\zeta) > 0$.

Problem (3.42a)–(3.42e) can be linked to the eigenvalue problems associated with the spectral constraints, (3.19a) and (3.19b), by insisting that $\tilde{\phi} = 0$. With this choice, (3.42a)–(3.42e) reduce to

$$\frac{1}{3}\partial\left(v_{1,\mathrm{ss}}\tilde{v}_1 + v_{2,\mathrm{ss}}\tilde{v}_2\right) + \partial\left(w_{1,\mathrm{ss}}\tilde{w}_1 + w_{2,\mathrm{ss}}\tilde{w}_2\right) = 0, \tag{3.43a}$$

$$\partial^4\tilde{v}_1 + \partial^2\tilde{v}_1 + \left(\frac{1}{3}\partial\phi_{\mathrm{ss}} - \frac{1}{2}\right)\tilde{v}_1 = -\zeta\tilde{v}_1, \tag{3.43b}$$

$$\partial^4\tilde{v}_2 + \partial^2\tilde{v}_2 + \left(\frac{1}{3}\partial\phi_{\mathrm{ss}} - \frac{1}{2}\right)\tilde{v}_2 = -\zeta\tilde{v}_2, \tag{3.43c}$$

$$\partial^4\tilde{w}_1 + \partial^2\tilde{w}_1 + \partial\phi_{\mathrm{ss}}\,\tilde{w}_1 = -\zeta\tilde{w}_1, \tag{3.43d}$$

$$\partial^4\tilde{w}_2 + \partial^2\tilde{w}_2 + \partial\phi_{\mathrm{ss}}\,\tilde{w}_2 = -\zeta\tilde{w}_2. \tag{3.43e}$$

It is clear that (3.43b) and (3.43c) correspond to (3.19a) upon identifying $\zeta = -\lambda$, while (3.42d) and (3.42e) correspond to (3.19b) upon identifying $\zeta = -\mu$. This means that if the steady background field ϕ_{ss} violates the spectral constraint $\mathcal{Q}_1\{v\} \geq 0$, then there exists

$\tilde{v}_1$ that satisfies (3.43b) with $\zeta = -\lambda > 0$. If, instead, ϕ_{ss} violates the spectral constraint $\mathcal{Q}_2\{w\} \geq 0$, then there exists $\tilde{w}_1$ that satisfies (3.43d) with $\zeta = -\mu > 0$. The strategy proposed by Wen *et al.* (2015) is to use these observations to prove the following claim.

Claim 3.2. *If ϕ_{ss} does not satisfy at least one of the spectral constraints, then there exist $\zeta \in \mathbb{R}$, $\zeta > 0$, and functions $\tilde{v}_1$, $\tilde{v}_2$, $\tilde{w}_1$, and $\tilde{w}_2$ that satisfy* (3.43a)–(3.43e).

Unfortunately, it does not appear possible to prove this result unless one makes additional assumptions on ϕ_{ss}, $v_{1,\text{ss}}$, $v_{2,\text{ss}}$, $w_{1,\text{ss}}$ and $w_{2,\text{ss}}$. To see where the obstacle lies, suppose for definiteness that ϕ_{ss} violates the spectral constraint $\mathcal{Q}_1\{v\} \geq 0$. For simplicity, set $\tilde{w}_1 = \tilde{w}_2 = 0$ (a different choice does not seem to help) and recall that $v_{1,\text{ss}}$ and $v_{2,\text{ss}}$ are odd and periodic on the domain $[-\ell, \ell]$, so they must vanish at $x = 0$. Then, (3.43a) requires

$$v_{1,\text{ss}}(x)\tilde{v}_1(x) + v_{2,\text{ss}}(x)\tilde{v}_2(x) = 0 \quad \forall x \in [-\ell, \ell]. \tag{3.44}$$

The steady states $v_{1,\text{ss}}$ and $v_{2,\text{ss}}$ are linearly independent in general, so for (3.44) to hold $\tilde{v}_1$ and $\tilde{v}_2$ must be linearly independent, too. In addition, $\tilde{v}_1$ and $\tilde{v}_2$ must satisfy (3.43b) and (3.43c), respectively. Thus, they must be (a linear combination of) linearly independent eigenfunctions of (3.19a) with the *same* eigenvalue $\lambda = -\zeta$, so one must choose $\lambda = -\zeta$ to be a repeated eigenvalue of (3.19a). An immediate obstacle is that, without further assumptions, one cannot guarantee that (3.19a) has any repeated negative eigenvalues. Moreover, even when a degenerate negative eigenvalue exists, it is possible that its corresponding eigenfunctions cannot be linearly combined to construct $\tilde{v}_1$ and $\tilde{v}_2$ such that (3.44) holds.

The latter problem is crucial, and is born out of a fundamental difference between the analysis of this chapter and the proofs presented by Wen *et al.* (2015): equation (3.44) is a *pointwise* condition, while for the problems studied by Wen *et al.* (2015) one obtains *integral* conditions. For instance, when studying Rayleigh's two-dimensional model for convection one requires that[4]

$$\int_0^L W(x,z)\tilde{\theta}(x,z) + \theta(x,z)\tilde{W}(x,z)\,\mathrm{d}x = 0 \qquad \forall z \in [0,1], \tag{3.45}$$

where W and θ are the spurious steady solutions of the EL equations (whose instability is to be proven) and $\tilde{W}$, $\tilde{\theta}$ are the corresponding perturbations. Since all functions are periodic in x, one can satisfy (3.45) by taking $\tilde{W}$ and $\tilde{\theta}$ as eigenfunctions of the spectral constraint with negative eigenvalue that, in addition, share no Fourier modes with θ and W. The possibility of exploiting orthogonality through the integral in (3.45) grants considerably more freedom

[4]This follows from equation (35) in Wen *et al.* (2015).

compared to (3.44), and this difference is what ultimately prevents one from extending Wen *et al.*'s approach and proving claim 3.2.

It is of course also possible that the claim is incorrect, and that to prove linear instability of ϕ_{ss}, $v_{1,\text{ss}}$, $v_{2,\text{ss}}$, $w_{1,\text{ss}}$ and $w_{2,\text{ss}}$ when one of the spectral constraints is violated one must consider eigenfunctions of (3.42a)–(3.42e) with $\tilde{\phi} \neq 0$ and complex eigenvalues. However, this makes the connection with the spectral constraints more difficult, and it does not seem easy to make progress. A simpler approach is to restrict the attention to particular classes of steady states, such that the uncertainty surrounding the multiplicity of the eigenvalues of (3.19a) and (3.19b) can be circumvented. In this way, one can at least guarantee that the time-marching method described in section 3.4 will not converge to certain types of undesired steady states. The following result is an example of what can be proven.

Theorem 3.3. *Suppose that ϕ_{ss}, $v_{1,\text{ss}}$, $v_{2,\text{ss}}$, $w_{1,\text{ss}}$ and $w_{2,\text{ss}}$ solve the EL equations* (3.34a)–(3.34e) *and satisfy at least one of the following conditions:*

(i) ϕ_{ss} violates the spectral constraint $\mathcal{Q}_1\{v\} \geq 0$ and $v_{2,\text{ss}} = cv_{1,\text{ss}}$ for a constant $c \neq 0$;

(ii) ϕ_{ss} violates the spectral constraint $\mathcal{Q}_2\{w\} \geq 0$ and $w_{2,\text{ss}} = cw_{1,\text{ss}}$ for a constant $c \neq 0$.

Then, the tuple $(\phi_{\text{ss}}, v_{1,\text{ss}}, v_{2,\text{ss}}, w_{1,\text{ss}}, w_{2,\text{ss}})$ is a linearly unstable steady solution of the time-dependent EL equations (3.35a)–(3.35e).

Remark 3.3. The existence of solutions of (3.34a)–(3.34e) that satisfy the conditions of Theorem 3.3 is easily demonstrated. Recall from section 3.3 that one can find $\phi_\star$, $v_\star$ and $w_\star$ that solve (3.23a)–(3.23c) and such that $\phi_\star$ does not satisfy at least one of the spectral constraints. Then, for any $\alpha, \beta \in [0, 1]$, the fields

$$\phi_{\text{ss}} = \phi_\star, \quad v_{1,\text{ss}} = \sqrt{\alpha}\, v_\star, \quad v_{2,\text{ss}} = \sqrt{1-\alpha}\, v_\star, \quad w_{1,\text{ss}} = \sqrt{\beta}\, w_\star, \quad w_{2,\text{ss}} = \sqrt{1-\beta}\, w_\star, \tag{3.46}$$

satisfy the conditions of Theorem 3.3.

Proof. Suppose that condition (i) holds and set $\tilde{\phi} = \tilde{w}_1 = \tilde{w}_2 = 0$ in (3.43a)–(3.43e). With these choices, proving linear instability requires finding $\zeta > 0$ and functions $\tilde{v}_1$, $\tilde{v}_2$ that satisfy

$$\partial\left(v_{1,\text{ss}}\,\tilde{v}_1 + c\, v_{1,\text{ss}}\,\tilde{v}_2\right) = 0, \tag{3.47a}$$

$$\partial^4\tilde{v}_1 + \partial^2\tilde{v}_1 + \left(\frac{1}{3}\partial\phi_{\text{ss}} - \frac{1}{2}\right)\tilde{v}_1 = -\zeta\tilde{v}_1, \tag{3.47b}$$

$$\partial^4\tilde{v}_2 + \partial^2\tilde{v}_2 + \left(\frac{1}{3}\partial\phi_{\text{ss}} - \frac{1}{2}\right)\tilde{v}_2 = -\zeta\tilde{v}_2. \tag{3.47c}$$

This can be done easily: let $\tilde{v}_1 = -c\tilde{v}_2$ and choose $\tilde{v}_2$ to be an eigenfunction of (3.19a) with eigenvalue $\lambda < 0$, which exists because the spectral constraint $\mathcal{Q}_1\{v\} \geq 0$ is violated by assumption. Then, $\zeta = -\lambda$ is a real positive eigenvalue for (3.42a)–(3.42e) and the tuple $(\phi_{\text{ss}}, v_{1,\text{ss}}, v_{2,\text{ss}}, w_{1,\text{ss}}, w_{2,\text{ss}})$ is a linearly unstable steady solution of (3.35a)–(3.35e). A similar argument can be applied when condition (ii) holds. □

3.5 Conclusions

This chapter described an attempt to employ the time-marching method proposed by Wen *et al.* (2013, 2015) to optimise background fields for the one-dimensional KS equation on the periodic domain $[-\ell, \ell]$. The construction of a suitable background field ϕ implies an upper bound on the asymptotic kinetic energy $\mathcal{E}$ of solutions of the KS equation, and the scaling of the best possible upper bound as a function of the domain's half-size ℓ is of interest to confirm that the background method cannot prove the conjecture that $\mathcal{E} \sim \ell$ when $\ell \gg 1$.

It has been shown that time-marching methods can indeed be utilised to solve the variational problem for the optimal ϕ. However, doing so requires careful consideration of how the spectral constraints on the background field are handled when constructing the Lagrangian for the problem. The numerical results presented in section 3.3.2 demonstrate that considering the Lagrangian (3.22), obtained by subtracting the spectral constraints from the objective function as suggested by the examples given by Wen *et al.* (2013, 2015), is not sufficient. Instead, the optimal ϕ could be computed irrespectively of the prescribed initial guess when two copies of each spectral constraint were subtracted from the objective function, yielding the Lagrangian (3.33). The use of this Lagrangian was suggested by an informal comparison between the variational problem for the optimal background field and a certain LMI-constrained problem, motivated by the observation that both spectral constraints and LMIs require the eigenvalues of a linear operator to be non-negative. This comparison revealed that the Lagrangian should be consistent with the multiplicity of the ground-state eigenvalues of this linear operator. Lagrangians of the form (3.22) worked well for every problem studied by Wen *et al.* (2013, 2015), suggesting that either the ground-state eigenvalues are simple, or their corresponding eigenfunctions possess some additional symmetries that make a simplified, "rank-one" solution possible. For the KS equation, instead, the ground-state eigenvalues may be repeated with multiplicity 2 (and, in fact, are so for the optimal background field), making it necessary to consider the modified Lagrangian (3.33).

It has also been demonstrated that studying the convergence properties of the time-marching optimisation method requires careful analysis. One fundamental question that

remains open is whether computations can converge to "spurious" solutions, which satisfy the EL equations characterising stationary points of the Lagrangian, but violate at least one of the spectral constraints. Wen *et al.* (2015) proved that this is impossible for a range of optimal background field problems arising in fluid dynamics. Their proofs work by showing that steady solutions of the time-dependent version of the EL equations are linearly unstable, and hence not attracting, if one spectral constraint is violated. For the KS equation, however, the same argument cannot be run because one cannot exploit orthogonality between spurious steady solutions and the unstable eigenfunctions of the spectral constraint, a key ingredient of the argument by Wen *et al.* (2015). Instead, only a particular class of spurious solutions of the EL equations could be proven not to be attracting. Consequently, it cannot be ruled out that, for certain ICs, the time-marching method returns an incorrect solution even if the correct Lagrangian is used.

Given the difficulties encountered in optimising background fields for the KS equation using time-marching methods, it is natural to wonder how easily these can be applied to solve other variational problems arising from the study of dynamical systems using the background method. One issue is that, in order to construct the correct Lagrangian, one must known *a priori* at least an upper bound on the multiplicity of the ground-states eigenvalues of the linear operator corresponding to each spectral constraint. For the KS equation, this information could be obtained by appealing to symmetry and some existing results on eigenvalue problems (cf. remark 3.2), but it is likely that the same will not be true for more complex systems. Should *a priori* bounds on the multiplicity of ground-state eigenvalues not be available, an iterative procedure may be employed, wherein a guess for the eigenvalue multiplicity is increased until the time-marching optimisation scheme converges to a background field that satisfies all spectral constraints. However, convergence to such a solution cannot actually be guaranteed even when the multiplicity of the ground-state eigenvalues associated with the spectral constraints is known. One reason is that it does not seem currently possible to generalise the argument put forward by Wen *et al.* (2015) and exclude convergence to steady states that violate at least one spectral constraint. Another reason is that, since the time-dependent equations to be solved are nonlinear, ICs may be attracted to a periodic orbit or a chaotic attractor, rather than to a steady solution (Wen *et al.*, 2015). Of course, the lack of theoretical convergence guarantees may not be an issue in practice, as demonstrated by the numerical results of section 3.4.2. Nonetheless, it may be possible to construct a "pathological" optimal background field problem, for which the time-marching method does not reach the correct solution for all but a small set of carefully selected ICs.

For these reasons, while time-marching methods remain attractive because they rely on well established and widely known numerical integration techniques, it is desirable to develop alternative approaches for the optimisation of background fields. One promising strategy, already implemented by Fantuzzi & Wynn (2015) for the KS equation, is to take advantage of the analogy between spectral constraints and LMIs, and construct optimal background fields using semidefinite programming. This line of attack will be pursued in the rest of this thesis. In fact, the next chapter (chapter 4) will go even further and demonstrate how SDPs can be utilised to compute optimal or near-optimal solutions of more general optimisation problems, subject to a particular class of affine and homogeneous integral inequality constraints. This class of constraints encompasses the spectral constraints encountered when the background method is utilised to study long-term or time-averaged properties of many systems, including the KS equation, provided that one seeks background fields parametrised by finitely many degrees of freedom (a mild restriction in practice). The SDP-based methods developed in chapter 4 will be utilised in chapter 5 to compute near-optimal background fields for stress-driven shear flows, while similar techniques, also based on SDPs, will be employed in chapter 6 to optimise bounds on the average convective heat transfer in Bénard–Marangoni convection at infinite Prandtl number.

Chapter 4

Optimisation with affine homogeneous quadratic integral inequalities[†]

When the background method is applied to derive bounds on long-term or time-averaged properties of chaotic systems governed by PDEs, the optimal background field is determined by the solution of a variational problem with spectral constraints. For instance, it was shown in chapter 3 that the background field ϕ yielding the best bound on the asymptotic kinetic energy of the Kuramoto–Sivashinky equation solves

$$\begin{aligned} \inf_{\phi \in H_{\mathrm{p,o}}} \quad & \int_\ell^\ell \left(\frac{8}{3} \left|\phi''\right|^2 - \frac{8}{3} \left|\phi'\right|^2 + |\phi|^2 \right) \mathrm{d}x \\ \text{s.t.} \quad & \int_\ell^\ell \left[\left|v''\right|^2 - \left|v'\right|^2 + \left(\frac{1}{3}\phi' - \frac{1}{2} \right) v^2 \right] \mathrm{d}x \geq 0 \qquad \forall v \in H_{\mathrm{p,o}}, \\ & \int_\ell^\ell \left(\left|w''\right|^2 - \left|w'\right|^2 + \phi' \, w^2 \right) \mathrm{d}x \geq 0 \qquad \forall w \in H_{\mathrm{p,o}}, \end{aligned} \tag{4.1}$$

where $H_{\mathrm{p,o}}$ is the space of odd and periodic square-integrable functions on $[-\ell, \ell]$ with square-integrable periodic derivatives.

Spectral constraints are particular instances of *integral inequality constraints*: the optimisation variables need to be determined such that a certain integral inequality holds for all functions of a certain class. For spectral constraints, this is equivalent to requiring that the eigenvalues of a certain linear, self-adjoint operator defined on this function class are non-negative, but the same need not be true in general. Integral inequality constraints, therefore, include but are not limited to spectral constraints.

[†]Most of the material presented in this chapter has been published in the following works:

Fantuzzi, G. and Wynn, A. (2016). Semidefinite relaxation of a class of quadratic integral inequalities. In: *Proceedings of the 55th IEEE Conference on Decision and Control*, Las Vegas, NV, IEEE. pp. 6192–6197. Available from: doi:10.1109/CDC.2016.7799221.

Fantuzzi, G., Wynn, A., Goulart, P. J. and Pachristodoulou, A. (2017). Optimization with affine homogeneous quadratic integral inequality constraints. *IEEE Transactions on Automatic Control* **62**(12), 6221–6236. Available from: doi:10.1109/TAC.2017.2703927.

Integral inequality constraints are commonly encountered when one is interested in the analysis and/or control of systems governed by PDEs. For example, the stability of an equilibrium of a PDE system with vector-valued state $\boldsymbol{w}(t,\boldsymbol{x})$ in a domain $\Omega \subset \mathbb{R}^n$, or of a control policy designed to stabilise it, can be established by constructing a positive integral Lyapunov functional $\mathcal{V}(t) = \mathcal{V}\{\boldsymbol{w}(t,\cdot)\} = \int_\Omega V[\boldsymbol{w}(t,\boldsymbol{x})]d\boldsymbol{x}$ whose time derivative, also an integral quantity, is non-positive at all times (Straughan, 2004; Valmorbida *et al.*, 2014*a*, 2016). Other input-to-state/output properties such as passivity, reachability, and input-to-state stability can be studied in a similar way by constructing functions of the state variable that satisfy certain integral inequalities (Ahmadi *et al.*, 2014, 2016).

In all these situations, the problem is either to check whether certain integral inequalities hold for all functions in a given set, or to optimise certain variables (for instance a background field) while satisfying some integral inequality constraint. In the simplest case, one is faced with an optimisation problem of the form

$$\begin{aligned} \min_{\boldsymbol{\gamma}\in S} \quad & c(\boldsymbol{\gamma}) \\ \text{s.t.} \quad & \mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\} := \int_\Omega F_{\boldsymbol{\gamma}}(\boldsymbol{x}, \mathcal{D}^{\boldsymbol{k}}\boldsymbol{w})\mathrm{d}^n\boldsymbol{x} \geq 0 \qquad \forall \boldsymbol{w} \in H, \end{aligned} \tag{4.2}$$

where H is a suitable function space, e.g. the space of all $\boldsymbol{k}$-times differentiable functions from $\Omega \subseteq \mathbb{R}^n$ (typically $n = 3$ for physical systems) to $\mathbb{R}^q$ that satisfy a given set of boundary conditions (BCs). The optimisation variable $\boldsymbol{\gamma}$ belongs to a finite- or infinite-dimensional set S,[1] $c : S \to \mathbb{R}$ is a convex cost function, $F_{\boldsymbol{\gamma}}(\cdot,\cdot)$ is a function that depends parametrically on $\boldsymbol{\gamma}$, and $\mathcal{D}^{\boldsymbol{k}}\boldsymbol{w} = [w_1, \partial_{x_1} w_1, \partial_{x_2} w_1, \ldots, \partial^{k_1}_{x_n} w_1, \ldots, \partial^{k_q}_{x_n} w_q]^\mathsf{T}$ lists all partial derivatives of the components of $\boldsymbol{w}$ up to the order specified by the multi-index $\boldsymbol{k} = [k_1, \ldots, k_q]$ (see the end of this section for more details). Problems with additional constraints on $\boldsymbol{\gamma}$ and multiple integral inequalities are common, but the former can always be incorporated into the definition of S, while each integral inequality can be enforced individually. Consequently, there is no loss of generality in considering problems of the form (4.2).

When the dependence on $\boldsymbol{\gamma}$ is affine, S is convex, and strong duality holds, problem (4.2) could in principle be solved using the calculus of variations (for an introduction to the subject, see Courant & Hilbert, 1953; Giaquinta & Hildebrandt, 1996) to solve the min-max problem

$$\min_{\boldsymbol{\gamma}} \max_{\substack{\boldsymbol{w} \\ \lambda \geq 0}} \mathcal{L}\{\boldsymbol{\gamma}, \boldsymbol{w}, \lambda\} := c(\boldsymbol{\gamma}) - \lambda \mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}, \tag{4.3}$$

[1] The set S is infinite-dimensional for problems arising from the application of the background method, such as the one considered in chapter 3: the search for the optimal background field is an optimisation over an infinite-dimensional class of functions.

where $\lambda \geq 0$ is a Lagrange multiplier enforcing the integral inequality. This strategy underlies the time-marching methods described in chapter 3 and, as demonstrated there, requires a very careful treatment of the constraint.

When the optimisation variable $\boldsymbol{\gamma}$ is finite-dimensional, and the integrand $F_{\boldsymbol{\gamma}}(\cdot,\cdot)$ is both linear with respect to $\mathcal{D}^{\boldsymbol{k}}\boldsymbol{w}$ and polynomial in $\boldsymbol{x}$, an alternative is to transform (4.2) into a semidefinite program (SDP) using integration by parts and moment relaxation techniques (Bertsimas & Caramanis, 2006). More recently, it has been suggested that (4.2) can be recast as an SDP even when the integrand is polynomial in $\mathcal{D}^{\boldsymbol{k}}\boldsymbol{w}$ (Papachristodoulou & Peet, 2006; Valmorbida *et al.*, 2014*a*, 2016, 2015): one relates the derivatives of the components of $\boldsymbol{w}$ using integration by parts and algebraic identities, and then requires that the polynomial integrand $F_{\boldsymbol{\gamma}}(\boldsymbol{x}, \mathcal{D}^{\boldsymbol{k}}\boldsymbol{w})$ admits a sum-of-squares (SOS) decomposition over the domain of integration. However, this approach is often impractical because SOS conditions of very high degree are needed to achieve accurate results.

This chapter presents a different SDP-based approach to solving a particular class of problems of type (4.2), which contains a number of non-trivial problems encountered in the study of PDEs. In particular, the methods developed here apply to many variational problems that arise when bounding time-averaged or long-term properties of PDE systems using the background method, after a mild assumption on the form of the background field is introduced. It will be assumed that $\mathcal{F}_{\boldsymbol{\gamma}}$ is a homogeneous quadratic functional over a one-dimensional compact domain. In other words, $\boldsymbol{x} \in \Omega \equiv [a, b] \subset \mathbb{R}$ with a and b finite, and the integrand $F_{\boldsymbol{\gamma}}(\boldsymbol{x}, \mathcal{D}^{\boldsymbol{k}}\boldsymbol{w})$ is a homogeneous quadratic polynomial with respect to $\mathcal{D}^{\boldsymbol{k}}\boldsymbol{w}$. It will also be assumed that the optimisation variable is finite-dimensional. This requirement is necessary to enable computations and is true in many applications. For instance, as described in example 4.1 below, when studying the nonlinear stability of a given fluid flow using the method of energy (Straughan, 2004) one seeks the largest value of a scalar parameter describing the forcing on the flow such that a certain integral inequality holds. Other times, requiring that the optimisation variable is finite-dimensional is only a mild restriction. This is the case when the background method is utilised to study PDE systems: the background field should be optimised over an infinite-dimensional class of functions, but since the optimal choice can be approximated arbitrarily accurately by a polynomial of sufficiently high degree, optimising over the finite-dimensional space of degree-d polynomial background fields suffices in practice when d is large.

The methods described in this chapter rely on Legendre series expansions to formulate SDPs with better scaling properties than the SOS method of Valmorbida *et al.* (2016). The main results of this chapter are:

(i) the formulation of convergent *outer* approximations of the feasible set of (4.2) described by linear matrix inequalities (LMIs), so convergent lower bounds for the optimal cost can be computed with semidefinite programming;

(ii) the derivation of LMI-representable *inner* approximations for the feasible set of (4.2), so upper bounds on its optimal value of can also be obtained by solving SDPs. These complement the aforementioned lower bounds, and enable one to assess their quality.

Legendre polynomials are central to this chapter, so they are briefly reviewed in section 4.1. Subsequently, section 4.2 defines the particular class of optimisation problems considered in this chapter. Outer and inner SDP relaxations are formulated in sections 4.3 and 4.4, respectively. In section 4.5, some assumptions introduced to ease the exposition are removed, and an extension of the inner/outer SDP relaxations to more general problems is presented. Section 4.6 shows how SDP relaxations can be applied to some simple, yet non-trivial problems arising from the analysis of PDEs, demonstrating the advantages of the proposed approaches compared to the SOS method of Valmorbida *et al.* (2016), as well as some limitations. The numerical experiments are carried out using QUINOPT (QUadratic INtegral OPTimisation), an add-on to the MATLAB optimisation toolbox YALMIP (Löfberg, 2004, 2009) developed as part of this work to assist the formulation of the SDP relaxations. Comments on the computational cost of the proposed techniques are given in section 4.7. Finally, section 4.8 offers concluding remarks and discusses possible future developments. To streamline the presentation, some technical results are proven in appendix A.

The following notation will be used throughout this chapter. Recall that $\mathbb{N}^q$ is the set of non-negative multi-indices of the form $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_q]$. Given $\boldsymbol{\alpha} \in \mathbb{N}^q$, the quantity $|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_q$ is known as the length of the multi-index. Moreover, if $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}^q$ with $\alpha_i \leq \beta_i \leq m$ for all $i \in \{1, \ldots, q\}$, the difference $\boldsymbol{\beta} - \boldsymbol{\alpha}$ is defined as $\boldsymbol{\beta} - \boldsymbol{\alpha} = [\beta_1 - \alpha_1, \ldots, \beta_q - \alpha_q] \in \mathbb{N}^q$. Given $\boldsymbol{w} \in C^m([a,b], \mathbb{R}^q)$, its derivatives of order between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ will be listed in the vector

$$\mathcal{D}^{[\boldsymbol{\alpha},\boldsymbol{\beta}]}\boldsymbol{w} := \left[\partial^{\alpha_1} u_1, \ldots, \partial^{\beta_1} u_1, \partial^{\alpha_2} u_2, \ldots, \partial^{\beta_2} u_2, \ldots, \partial^{\beta_q} u_q\right]^{\mathsf{T}} \in \mathbb{R}^{q+|\boldsymbol{\beta}-\boldsymbol{\alpha}|}. \tag{4.4}$$

All boundary values of such derivatives are collected in the vector

$$\mathcal{B}^{[\boldsymbol{\alpha},\boldsymbol{\beta}]}\boldsymbol{w} := \begin{bmatrix} \mathcal{D}^{[\boldsymbol{\alpha},\boldsymbol{\beta}]}\boldsymbol{w}(a) \\ \mathcal{D}^{[\boldsymbol{\alpha},\boldsymbol{\beta}]}\boldsymbol{w}(b) \end{bmatrix} \in \mathbb{R}^{2(q+|\boldsymbol{\beta}-\boldsymbol{\alpha}|)}. \tag{4.5}$$

For simplicity, the notation $\mathcal{D}^{\boldsymbol{\beta}}\boldsymbol{w}$ and $\mathcal{B}^{\boldsymbol{\beta}}\boldsymbol{w}$ will be used instead of $\mathcal{D}^{[\boldsymbol{0},\boldsymbol{\beta}]}\boldsymbol{w}$ and $\mathcal{B}^{[\boldsymbol{0},\boldsymbol{\beta}]}\boldsymbol{w}$.

4.1 Legendre polynomials and Legendre series

The Legendre polynomial of degree n is defined over the interval $[-1, 1]$ as

$$L_n(x) = \frac{1}{n!\,2^n} \frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2 - 1)^n. \tag{4.6}$$

All Legendre polynomials of degree $n \geq 2$ can also be constructed using the relation

$$nL_n(x) = (2n-1)\,x\,L_{n-1}(x) - (n-1)L_{n-2}(x), \tag{4.7}$$

with $L_0(x) = 1$ and $L_1(x) = x$. A straightforward induction argument using (4.7) proves that $L_n(\pm 1) = (\pm 1)^n$ for all $n \geq 0$, and it can be shown that $\|L_n\|_\infty \leq 1$.

The Legendre polynomials satisfy a number of other recurrence relations. A fundamental fact used in this chapter is that (Agarwal & O'Regan, 2009, chapter 7, problem 7.8)

$$(2n+1)L_n(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left[L_{n+1}(x) - L_{n-1}(x)\right], \quad n \geq 1. \tag{4.8}$$

The Legendre polynomials form a complete orthogonal basis for the Lebesgue space $L^2(-1, 1)$ (Zeidler, 1995), and satisfy the orthogonality condition

$$\int_{-1}^{1} L_n\,L_m\,\mathrm{d}x = \frac{2\delta_{mn}}{2n+1}, \tag{4.9}$$

where δ_{mn} is the usual Kronecker delta. This means that any square-integrable function u can be expanded with a convergent series (in the L^2 norm sense)

$$u(x) = \sum_{n=0}^{\infty} \hat{u}_n L_n(x), \qquad \hat{u}_n = \frac{2n+1}{2} \int_{-1}^{1} u\,L_n\,\mathrm{d}x, \tag{4.10}$$

where the values $\hat{u}_n$ are known as Legendre coefficients. From (4.9) it follows that

$$\|u\|_2^2 = \int_{-1}^{1} |u|^2\,\mathrm{d}x = \sum_{n=0}^{\infty} \frac{2|\hat{u}_n|^2}{2n+1}. \tag{4.11}$$

If, in addition, u is continuously differentiable on $[-1, 1]$, then its Legendre series expansion converges uniformly. In fact, u is Lipschitz on $[-1, 1]$ because, by Taylor's theorem, for any $x, y \in [-1, 1]$ there exists $z \in [x, y]$ such that $|u(y) - u(x)| = |\partial u(z)|\,|x - y| \leq C\,|x - y|$, where C is a positive constant whose existence is guaranteed by the continuity of ∂u in $[-1, 1]$. Uniform convergence follows from the Lipschitz condition according to Theorem XI in Jackson (1930).

4.2 A class of optimisation problems

As anticipated in the introduction, this chapter focusses on a particular class of optimisation problems of type (4.2). This class, described below, is sufficiently general to enable the solution of problems arising from applications of the background method to many fluid dynamical systems, which are the main interest of this thesis (cf. chapters 5 and 6).

Let $\boldsymbol{\gamma} \in \mathbb{R}^s$ be a vector of optimisation variables, and consider two integers m, q and two multi-indices $\boldsymbol{k} = [k_1, \ldots, k_q]$, $\boldsymbol{l} = [l_1, \ldots, l_q] \in \mathbb{N}^q$ such that

$$1 \leq k_i \leq m-1, \qquad i = 1, \ldots, q, \tag{4.12a}$$

$$k_i \leq l_i \leq m, \qquad i = 1, \ldots, q. \tag{4.12b}$$

Moreover, let $\boldsymbol{F}_0(x), \ldots, \boldsymbol{F}_s(x) \in \mathbb{S}^{q+|\boldsymbol{k}|}$ be matrices of polynomials of x of degree at most d_F and define

$$\boldsymbol{F}(x;\boldsymbol{\gamma}) := \boldsymbol{F}_0(x) + \sum_{i=1}^{s} \gamma_i \boldsymbol{F}_i(x). \tag{4.13}$$

In other words, $\boldsymbol{F}(x;\boldsymbol{\gamma})$ is a symmetric matrix of polynomials of x of degree at most d_F, the coefficients of which are affine in $\boldsymbol{\gamma}$. The focus of this chapter will be on linear optimisation problems of type (4.2) subject to *affine homogeneous quadratic integral inequalities* with compact domain of integration, meaning problems of the form

$$\begin{aligned} \min_{\boldsymbol{\gamma}} \quad & \boldsymbol{c}^{\mathsf{T}}\boldsymbol{\gamma} \\ \text{s.t.} \quad & \mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\} := \int_{-1}^{1} \left(\mathcal{D}^{\boldsymbol{k}}\boldsymbol{w}\right)^{\mathsf{T}} \boldsymbol{F}(x;\boldsymbol{\gamma}) \mathcal{D}^{\boldsymbol{k}}\boldsymbol{w} \, \mathrm{d}x \geq 0 \qquad \forall \boldsymbol{w} \in H, \end{aligned} \tag{4.14}$$

where $\boldsymbol{c} \in \mathbb{R}^s$ is the cost vector, $\boldsymbol{F}(x;\boldsymbol{\gamma})$ is as in (4.13), and

$$H := \left\{ \boldsymbol{w} \in C^m\left([-1,1], \mathbb{R}^q\right) : \ \boldsymbol{A}\,\mathcal{B}^{\boldsymbol{l}}\boldsymbol{w} = \boldsymbol{0} \right\} \tag{4.15}$$

is the space of m-times continuously differentiable functions satisfying the p homogeneous linear BCs defined by the matrix $\boldsymbol{A} \in \mathbb{R}^{p \times 2(q+|\boldsymbol{l}|)}$. There is no loss of generality in fixing the integration domain for the functional $\mathcal{F}_{\boldsymbol{\gamma}}$ to $[-1,1]$ because any compact interval $[a,b]$ can be mapped to it with a change of integration variable. An affine homogeneous quadratic integral inequality is a convex constraint on $\boldsymbol{\gamma}$, so (4.14) is a convex optimisation problem.

Remark 4.1. For the sake of generality, the space H can be defined by derivatives of higher order than those appearing in $\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}$. This can always be achieved by adding zero columns to $\boldsymbol{A}$. In problems arising from the study of PDEs, this is not uncommon: H encodes the

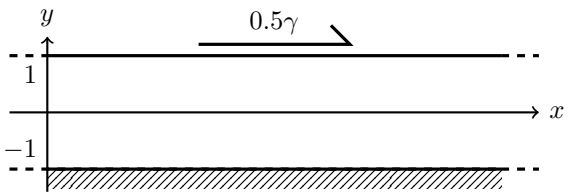


FIGURE 4.1: Sketch of the shear flow considered in example 4.1. The two-dimensional fluid layer extends to infinity along the x direction, is bounded at $y = -1$ by a solid boundary and is driven at the surface ($y = 1$) by a shear stress of non-dimensional magnitude 0.5γ.

BCs of the solution of a PDE, which might involve all derivatives up to the order of the PDE. On the other hand, $\mathcal{F}_\gamma\{\boldsymbol{w}\}$ is typically derived from a weak formulation of the PDE, after integrating some terms by parts.

Assumption 4.1. To ease the exposition, from here onwards the discussion will concentrate on two-dimensional functions $\boldsymbol{w} = [u, v]^\mathsf{T} \in C^m([-1, 1], \mathbb{R}^2)$ and on uniform multi-indices $\boldsymbol{k} = [k, k]$ and $\boldsymbol{l} = [l, l]$, where k and l satisfy (4.12a) and (4.12b). As discussed in section 4.5, however, the results presented in this chapter hold also for the general case.

Example 4.1. Consider a two-dimensional infinite layer of fluid bounded at $y = -1$ by a solid wall and driven at the surface ($y = 1$) by a horizontal shear stress of non-dimensional magnitude 0.5γ, as shown in figure 4.1. The flow is governed by the incompressible Navier–Stokes equations, and admits a steady (time independent) solution in which the flow moves horizontally with velocity $\boldsymbol{w}_0 = (u_0, v_0) = (0.5\gamma y + 0.5\gamma, 0)$ (see for example Tang *et al.*, 2004; Hagstrom & Doering, 2014). This steady flow is stable when the driving stress is small. The critical value $\gamma_{\mathrm{cr}}(\xi)$ at which the steady flow is no longer guaranteed to be stable with respect to a sinusoidal perturbation $\boldsymbol{w}(y)e^{i\xi x+\sigma t}$—where $\boldsymbol{w}(y) = [u(y), v(y)]^\mathsf{T}$ is the amplitude and ξ is the wavenumber—is given by

$$\begin{aligned} \gamma_{\mathrm{cr}}(\xi) := \arg\min \quad & -\gamma \\ \text{s.t.} \quad & \int_{-1}^{1} \left\{ \frac{16}{\xi^2}[(\partial_y^2 u)^2 + (\partial_y^2 v)^2] + 8[(\partial_y u)^2 + (\partial_y v)^2] \right. \\ & \left. \qquad +\xi^2(u^2 + v^2) + \frac{2\gamma}{\xi}(v\partial_y u - u\partial_y v) \right\} \mathrm{d}y \geq 0, \end{aligned} \tag{4.16}$$

where the integral inequality should hold for all functions $u, v \in C^2([-1, 1])$ satisfying the homogeneous BCs

$$\begin{aligned} u(-1) = u(1) = \partial_y u(-1) = \partial_y^2 u(1) = 0, \\ v(-1) = v(1) = \partial_y v(-1) = \partial_y^2 v(1) = 0. \end{aligned} \tag{4.17}$$

The reader is referred to Tang *et al.* (2004) or Hagstrom & Doering (2014) for more details.

The constraint in (4.16) can be rewritten in matrix form as in (4.14) with $\boldsymbol{k} = \boldsymbol{l} = [2, 2]$ and

$$\mathcal{D}^{\boldsymbol{k}}\boldsymbol{w} = \begin{bmatrix} u \\ \partial_y u \\ \partial_y^2 u \\ v \\ \partial_y v \\ \partial_y^2 v \end{bmatrix}, \qquad \boldsymbol{F}(x;\boldsymbol{\gamma}) = \begin{bmatrix} \xi^2 & 0 & 0 & 0 & -\frac{\gamma}{\xi} & 0 \\ 0 & 8 & 0 & \frac{\gamma}{\xi} & 0 & 0 \\ 0 & 0 & \frac{16}{\xi^2} & 0 & 0 & 0 \\ 0 & \frac{\gamma}{\xi} & 0 & \xi^2 & 0 & 0 \\ -\frac{\gamma}{\xi} & 0 & 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{16}{\xi^2} \end{bmatrix}.$$

Note that the matrix $\boldsymbol{F}$ above can be written in the form (4.13) with $s = 1$. The reader can easily verify that the BCs on u and v can also be rewritten in the matrix form $\boldsymbol{A}\,\mathcal{B}^{\boldsymbol{l}}\boldsymbol{w} = \boldsymbol{0}$ with $\boldsymbol{A} \in \mathbb{R}^{8\times 12}$; the details are omitted for brevity. For this problem, it is clear that $\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\} \geq 0$ for $\gamma = 0$, and that definiteness is lost for sufficiently large γ. However, the interaction of the BCs with this behavior makes the problem interesting and non-trivial to solve. Upper and lower bounds for the optimal γ in (4.16) will be computed in section 4.6.1.

4.3 Outer SDP relaxations

The first approach to solve (4.14) is to derive a sequence of *outer* approximations for its feasible set,

$$T := \{\boldsymbol{\gamma} \in \mathbb{R}^s : \ \forall \boldsymbol{w} \in H,\ \mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\} \geq 0\}. \tag{4.18}$$

In other words, one looks for a family of sets $\{T_N^{\text{out}}\}_{N\geq 0}$ such that $T \subset T_N^{\text{out}}$. As will be demonstrated below, such sets can be found by relaxing the integral inequality $\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\} \geq 0$, enforcing it only over a certain subset of the test function space H. Optimising the cost function over T_N^{out} then gives a lower bound for the optimal value of (4.14), and convergence of this lower bound can be guaranteed subject to very mild conditions.

One way to construct an outer approximation set T_N^{out} for the feasible set T of (4.14) is to weaken the integral inequality constraint by enforcing it only for polynomials $\boldsymbol{w}$ of degree N. Precisely, one restricts attention to

$$\boldsymbol{w} = [u, v]^{\mathsf{T}} \in S_N := H \cap (P_N \times P_N) \subset H, \tag{4.19}$$

where P_N is the set of polynomials of degree no larger than N on $[-1, 1]$. The set S_N is non-empty for any degree bound N because H contains the zero polynomial, and it contains

non-zero elements if N is large enough to guarantee sufficient degrees of freedom to satisfy the BCs prescribed on H in (4.15). Finally, $S_N \subset S_{N+1}$ because $P_N \subset P_{N+1}$.

Now, let $\hat{u}_0, \ldots, \hat{u}_N$ and $\hat{v}_0, \ldots, \hat{v}_N$ be the coefficients representing the polynomials u and v in any chosen basis for P_N, and define $\boldsymbol{\varphi}_N := [\hat{u}_0, \ldots, \hat{u}_N, \hat{v}_0, \ldots, \hat{v}_N]^\mathsf{T}$. Since $\mathcal{F}_{\boldsymbol{\gamma}}$ in (4.14) is quadratic and the constraints imposed on H are linear, it is clear that there exist a matrix $\boldsymbol{Q}_N(\boldsymbol{\gamma})$, affine in $\boldsymbol{\gamma}$, such that

$$\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\} = {\boldsymbol{\varphi}_N}^\mathsf{T} \boldsymbol{Q}_N(\boldsymbol{\gamma}) \boldsymbol{\varphi}_N, \tag{4.20}$$

and a matrix $\boldsymbol{A}_N$ such that

$$\boldsymbol{w} \in S_N \Leftrightarrow \boldsymbol{A}_N \boldsymbol{\varphi}_N = \boldsymbol{0}. \tag{4.21}$$

Upon selecting a matrix $\boldsymbol{\Pi}_N$ satisfying $\mathrm{img}(\boldsymbol{\Pi}_N) = \ker(\boldsymbol{A}_N)$, it follows that

$$\begin{aligned} T_N^{\mathrm{out}} &:= \{\boldsymbol{\gamma} \in \mathbb{R}^s : \ \forall \boldsymbol{w} \in S_N, \ \mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\} \geq 0\} \\ &= \left\{\boldsymbol{\gamma} \in \mathbb{R}^s : \ {\boldsymbol{\Pi}_N}^\mathsf{T} \boldsymbol{Q}_N(\boldsymbol{\gamma}) \boldsymbol{\Pi}_N \succeq 0\right\}. \end{aligned} \tag{4.22}$$

Moreover, the inclusion $S_N \subset S_{N+1} \subset H$ implies that the feasible set T of (4.14), defined as in (4.18), satisfies

$$T \subset T_{N+1}^{\mathrm{out}} \subset T_N^{\mathrm{out}} \quad \text{for all } N \in \mathbb{N}. \tag{4.23}$$

Thus, $\{T_N^{\mathrm{out}}\}_{N \geq 1}$ is a sequence of nested outer approximations of the feasible set of (4.14).

As anticipated above, optimising $\boldsymbol{\gamma}$ over each set T_N^{out} yields a sequence of lower bounds on the optimal value of (4.14). In particular, one can prove the following result.

Theorem 4.2. *Let p^* be the optimal value of* (4.14) *and, for each integer N, let p_N^* be the optimal value of the SDP*

$$\begin{aligned} \min_{\boldsymbol{\gamma}} \quad & \boldsymbol{c}^\mathsf{T} \boldsymbol{\gamma} \\ \text{s.t.} \quad & {\boldsymbol{\Pi}_N}^\mathsf{T} \boldsymbol{Q}_N(\boldsymbol{\gamma}) \boldsymbol{\Pi}_N \succeq 0. \end{aligned} \tag{4.24}$$

Then, $\{p_N^\star\}_{N \geq 0}$ is a non-decreasing sequence of lower bounds for $p^\star$. Furthermore, if a minimiser $\gamma^\star$ exists in (4.14)*, then $\lim_{N \to \infty} |p_N^\star - p^\star| = 0$.*

Proof. See appendix A.2. □

Remark 4.2. Clearly, infeasibility of the SDP (4.24) for a certain N provides a *certificate of infeasibility* for (4.14). However, feasibility (resp. unboundedness) of (4.24) for any finite N does *not* prove that (4.14) is indeed feasible (resp. unbounded).

Unfortunately, Theorem 4.2 provides no control on the gap $p^\star - p_N^\star$ *as a function of* N. In other words, an arbitrarily large N might be required for a given level of approximation accuracy. Consequently, to assess the quality of the lower bounds on $p^\star$ obtained with the SDP (4.24) it is fundamental to formulate checkable conditions upon which upper bounds can be placed on $p^\star$. This will be the subject of the next section.

4.4 Inner SDP relaxations

Upper bounds on the optimal value of problem (4.14), that complement the lower bounds from Theorem 4.2, can be found by optimising the cost function over an inner approximation T_N^{in} of the true feasible set. Such an inner approximation can be constructed by replacing the integral inequality $\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\} \geq 0$ with a stronger, but tractable, integral inequality over the space H in (4.15). In particular, one looks for a lower bound $\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\} \geq \mathcal{G}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}$, where $\mathcal{G}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}$ is a functional whose non-negativity over H can be enforced via a set of LMIs. Any $\boldsymbol{\gamma}$ such that $\mathcal{G}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\} \geq 0$ on H is then also feasible for (4.14), and the corresponding cost $\boldsymbol{c}^\mathsf{T}\boldsymbol{\gamma}$ is an upper bound for the optimal value of (4.14). This strategy is somewhat complementary to the approach followed in section 4.3, where the space H was replaced with a finite-dimensional subspace S_N in order to formulate LMIs.

4.4.1 Legendre series expansions

The key to constructing an inner approximation for the feasible set of problem (4.14) is to find a functional $\mathcal{G}_{\boldsymbol{\gamma}} : H \to \mathbb{R}$ such that $\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\} \geq \mathcal{G}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}$ for all $\boldsymbol{w} \in H$. This can be done by expanding the components u and v of $\boldsymbol{w}$ (recall the restriction to the two-dimensional case made in assumption 4.1) and their derivatives using Legendre series such as

$$\partial^\alpha u = \sum_{n=0}^{\infty} \hat{u}_n^\alpha \, L_n(x), \qquad \partial^\beta v = \sum_{n=0}^{\infty} \hat{v}_n^\alpha \, L_n(x), \tag{4.25}$$

where $L_n(x)$ is the degree-n Legendre polynomial and $\hat{u}_n^\alpha$, $\hat{v}_n^\alpha$ are the Legendre coefficients.

Legendre series expansions are useful because the Legendre polynomials are orthogonal on $[-1, 1]$, *i.e.*, $\int_{-1}^{1} L_m \, L_n \, \mathrm{d}x = 0$ if $m \neq n$. This property will be essential to formulate a set of finite-dimensional, numerically tractable conditions enforcing the non-negativity of the functional $\mathcal{F}_{\boldsymbol{\gamma}}$ in (4.14). Firstly, orthogonality enables the exact representation of certain functionals, so conservative estimates need not be introduced. Secondly, it promotes sparsity in the finite-dimensional conditions because many of the terms involved can be shown to vanish. Other polynomial basis functions, such as Chebyshev polynomials, may have more

attractive numerical properties, but do not bring the same benefits because they are only orthogonal with respect to a weight.

To avoid working with infinite series and to facilitate the analysis, it is convenient to decompose the expansions in (4.25) into a finite sum and a remainder function. Precisely, given $i \in \mathbb{N}$, define the remainder functions

$$U_i^\alpha(x) = \sum_{n=i+1}^{\infty} \hat{u}_n^\alpha \, L_n(x), \qquad V_i^\beta(x) = \sum_{n=i+1}^{\infty} \hat{v}_n^\beta \, L_n(x). \tag{4.26}$$

Next, choose $N \in \mathbb{N}$ such that

$$N \geq d_F + k - 1, \tag{4.27}$$

where d_F is the degree of the polynomial matrix $\boldsymbol{F}$ defined in (4.13). Then, for each $\alpha \in \{1, \ldots, k\}$, decompose the Legendre expansion of $\partial^\alpha u$ and $\partial^\beta v$ as

$$\partial^\alpha u = \sum_{n=0}^{N+\alpha} \hat{u}_n^\alpha \, L_n(x) + U_{N+\alpha}^\alpha(x), \qquad \partial^\beta v = \sum_{n=0}^{N+\beta} \hat{v}_n^\beta \, L_n(x) + V_{N+\beta}^\beta(x). \tag{4.28}$$

For notational ease, given integers $0 \leq r \leq s$, the coefficients $\hat{u}_r^\alpha, \ldots, \hat{u}_s^\alpha$ will be recorded in the vector

$$\hat{\boldsymbol{u}}_{[r,s]}^\alpha = \begin{bmatrix} \hat{u}_r^\alpha, & \ldots, & \hat{u}_s^\alpha \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^{s-r+1}, \tag{4.29}$$

and a similar vector $\hat{\boldsymbol{v}}_{[r,s]}^\beta$ will be considered for the coefficients $\hat{v}_r^\beta, \ldots, \hat{v}_s^\beta$. Finally, for technical reasons that will be pointed out in section 4.4.2, it is also useful to introduce "extended" decompositions for the highest-order derivatives, $\partial^k u$ and $\partial^k v$. Specifically, consider

$$\partial^k u = \sum_{n=0}^{M} \hat{u}_n^k \, L_n(x) + U_M^k(x), \qquad \partial^k v = \sum_{n=0}^{M} \hat{v}_n^k \, L_n(x) + V_M^k(x). \tag{4.30}$$

where

$$M := N + 2k + d_F. \tag{4.31}$$

The following result relates the Legendre coefficients of a function u and its derivatives.

Lemma 4.3. *Let $u \in C^m([-1,1])$ and its derivatives up to order $k \leq m-1$ be expanded as in (4.28), and let M be as in (4.31). For any $\alpha \in \{1, \ldots, k\}$ and any two integers r, s with $0 \leq r \leq s \leq M + \alpha - k$, there exist matrices $\boldsymbol{B}_{[r,s]}^\alpha$ and $\boldsymbol{D}_{[r,s]}^\alpha$ such that*

$$\hat{\boldsymbol{u}}_{[r,s]}^\alpha = \boldsymbol{B}_{[r,s]}^\alpha \, \mathcal{D}^{k-1} u(-1) + \boldsymbol{D}_{[r,s]}^\alpha \hat{\boldsymbol{u}}_{[0,M]}^k. \tag{4.32}$$

Furthermore, $\boldsymbol{B}_{[r,s]}^\alpha = \boldsymbol{0}$ if $r \geq k - \alpha$.

Proof. See appendix A.3. □

This lemma simply states that, given the Legendre coefficients $\hat{u}_0^k, \ldots, \hat{u}_M^k$ of $\partial^k u$, the Legendre coefficients of all derivatives of order $\alpha < k$ can be computed uniquely if the vector of boundary values $\mathcal{D}^{k-1}u(-1)$ is specified. These boundary values play the role of integration constants, and should be treated as variables until specific BCs are prescribed. For any integer n it is therefore useful to define the vector of variables

$$\check{\boldsymbol{u}}_n = \left[\left(\mathcal{D}^{k-1}u(-1)\right)^\mathsf{T}, \quad \hat{u}_0^k, \quad \ldots, \quad \hat{u}_n^k\right]^\mathsf{T} \in \mathbb{R}^{k+n+1}. \tag{4.33}$$

The boundary values of u and its derivatives can also be represented using Legendre expansions. This is helpful because the integral inequality in (4.14) is only required to hold for functions that satisfy prescribed BCs. In particular, the following lemma states that if one knows the value of $\partial^\alpha u$ at $x = -1$ for all $\alpha \in \{0, \ldots, k-1\}$ and the first M Legendre coefficients of $\partial^k u$, then for each $\alpha \in \{0, \ldots, k-1\}$ one can compute the boundary value $\partial^\alpha u(1)$. This result, which may seem surprising at first, follows from simple integration using the properties of the Legendre polynomials.

Lemma 4.4. *Let $u \in C^m([-1,1])$ and its derivatives up to order $k \leq m-1$ be expanded as in* (4.28)*, and let $\mathcal{B}^{k-1}u \in \mathbb{R}^{2k}$ be defined according to* (4.5)*. Moreover, let M be as in* (4.31)*, and let $\check{\boldsymbol{u}}_M \in \mathbb{R}^{k+M+1}$ be defined according to* (4.33)*. There exists a matrix $\boldsymbol{G}_M \in \mathbb{R}^{2k\times(k+M+1)}$ such that $\mathcal{B}^{k-1}u = \boldsymbol{G}_M\check{\boldsymbol{u}}_M$.*

Proof. See appendix A.4. □

4.4.2 Legendre expansions of $\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}$

Recalling the definitions of $\mathcal{D}^{\boldsymbol{k}}\boldsymbol{w}$ from (4.4) and of $\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}$ from (4.14), the latter is a sum of terms of the form

$$\int_{-1}^{1} f\,\partial^\alpha u\,\partial^\beta v\,\mathrm{d}x, \tag{4.34a}$$

$$\int_{-1}^{1} f\,\partial^\alpha u\,\partial^\beta u\,\mathrm{d}x, \tag{4.34b}$$

$$\int_{-1}^{1} f\,\partial^\alpha v\,\partial^\beta v\,\mathrm{d}x, \tag{4.34c}$$

where $\alpha, \beta \in \{0, \ldots, k\}$. In each of these expressions, $f = f(x;\boldsymbol{\gamma})$ denotes the appropriate entry of the integrand matrix $\boldsymbol{F}(x;\boldsymbol{\gamma})$ and, consequently, it is a polynomial of degree at most d_F whose coefficients are affine in $\boldsymbol{\gamma}$. For generality, the following discussion focusses on terms such as (4.34a), which involve both components u and v of $\boldsymbol{w}$. Analogous considerations can be made when considering terms of the form (4.34b) or (4.34c).

Each term of the form (4.34a) can be conveniently analysed if $\partial^\alpha u$ and $\partial^\beta v$ are substituted by their decomposed Legendre expansions according to the following strategy:

- if $\alpha \neq k$ or $\beta \neq k$, use (4.28);
- if $\alpha = \beta = k$, use the "extended" decomposition (4.30).

The reasons for this choice will be explained in remark 4.5. In either case, one can rewrite (4.34a) as

$$\int_{-1}^{1} f\, \partial^\alpha u\, \partial^\beta v \,\mathrm{d}x = \mathcal{P}_{uv}^{\alpha\beta} + \mathcal{Q}_{uv}^{\alpha\beta} + \mathcal{R}_{uv}^{\alpha\beta}, \tag{4.35}$$

where

$$\mathcal{P}_{uv}^{\alpha\beta} := \sum_{m=0}^{N_\alpha} \sum_{n=0}^{N_\beta} \hat{u}_m^\alpha \hat{v}_n^\beta \int_{-1}^{1} f\, L_m\, L_n \,\mathrm{d}x, \tag{4.36a}$$

$$\mathcal{Q}_{uv}^{\alpha\beta} := \sum_{n=0}^{N_\alpha} \hat{u}_n^\alpha \int_{-1}^{1} f\, L_n\, V_{N_\beta}^\beta \,\mathrm{d}x + \sum_{n=0}^{N_\beta} \hat{v}_n^\beta \int_{-1}^{1} f\, L_n\, U_{N_\alpha}^\alpha \,\mathrm{d}x, \tag{4.36b}$$

$$\mathcal{R}_{uv}^{\alpha\beta} := \int_{-1}^{1} f\, U_{N_\alpha}^\alpha\, V_{N_\beta}^\beta \,\mathrm{d}x. \tag{4.36c}$$

It should be understood that $N_\alpha = N + \alpha$ and $N_\beta = N + \beta$ if (4.28) is used to expand $\partial^\alpha u$ and $\partial^\beta v$, while $N_\alpha = N_\beta = M = N + 2k + d_F$ if (4.30) is used.

The term $\mathcal{P}_{uv}^{\alpha\beta}$ is finite-dimensional and, for any choice of $\alpha, \beta \in \{0, \dots, k\}$, it can be rewritten as a symmetric quadratic form for the vectors $\hat{\boldsymbol{u}}_{[0,N_\alpha]}^\alpha$ and $\hat{\boldsymbol{v}}_{[0,N_\beta]}^\beta$. Recalling Lemma 4.3 and defining

$$\boldsymbol{\psi}_M := \begin{bmatrix} \check{\boldsymbol{u}}_M \\ \check{\boldsymbol{v}}_M \end{bmatrix} \in \mathbb{R}^{2(k+M+1)}, \tag{4.37}$$

where $\check{\boldsymbol{u}}_M$ and $\check{\boldsymbol{v}}_M$ are as in (4.33), one obtains the following result.

Lemma 4.5. *Let $\mathcal{P}_{uv}^{\alpha\beta}$ be as defined in* (4.36a) *and let $\boldsymbol{\psi}_M$ be defined according to* (4.37)*. There exists a matrix $\boldsymbol{P}_{uv}^{\alpha\beta}(\boldsymbol{\gamma}) \in \mathbb{S}^{2(k+M+1)}$, whose entries depend affinely on $\boldsymbol{\gamma}$, such that $\mathcal{P}_{uv}^{\alpha\beta} = {\boldsymbol{\psi}_M}^\mathsf{T} \boldsymbol{P}_{uv}^{\alpha\beta}(\boldsymbol{\gamma})\, \boldsymbol{\psi}_M$.*

The term $\mathcal{Q}_{uv}^{\alpha\beta}$ is less straightforward to handle, because it couples the first $N_\alpha + 1$ and $N_\beta + 1$ modes of $\partial^\alpha u$ and $\partial^\beta v$, respectively, to the remainder functions $V_{N_\beta}^\beta$ and $U_{N_\alpha}^\alpha$. Considering the extended decomposition (4.30) for the Legendre series of $\partial^k u$ and $\partial^k v$ enables one to write $\mathcal{Q}_{uv}^{\alpha\beta}$ as a finite-dimensional matrix quadratic form for the vector $\boldsymbol{\psi}_M$ if $\alpha \neq k$ or $\beta \neq k$ (details can be found in appendix A.5). If $\alpha = \beta = k$, on the other hand, one cannot do the same unless f in (4.36b) is independent of x (in this case, the orthogonality of the Legendre polynomials and the remainder functions implies that $\mathcal{Q}_{uv}^{kk} = 0$). To decouple the remainder functions from the other terms it is necessary to estimate $\mathcal{Q}_{uv}^{kk}$.

To make these ideas more precise, recall (4.37), (4.33), (4.29) and consider a family of "deflation" matrices $\boldsymbol{\Xi}_n \in \mathbb{R}^{2(M-n+1)\times 2(k+M+1)}$ such that

$$\boldsymbol{\Xi}_n \, \boldsymbol{\psi}_M = \begin{bmatrix} \hat{\boldsymbol{u}}^k_{[n,M]} \\ \hat{\boldsymbol{v}}^k_{[n,M]} \end{bmatrix}, \qquad n = 0, \ldots, M. \tag{4.38}$$

Moreover, given four integers $a \leq b$ and $c \leq d$, let $\boldsymbol{\Phi}^{[c,d]}_{[a,b]}$ be a $(b-a+1)\times(d-c+1)$ matrix whose ij-th element is defined as

$$\left(\boldsymbol{\Phi}^{[c,d]}_{[a,b]}\right)_{i,j} = \int_{-1}^{1} f \, L_{m_i} \, L_{n_j} \, \mathrm{d}x, \tag{4.39}$$

where m_i and n_j are the i-th and j-th elements of the sequences $\{a, \ldots, b\}$ and $\{c, \ldots, d\}$. Note that, strictly speaking, $\boldsymbol{\Phi}^{[c,d]}_{[a,b]}$ depends on f, and its entries are affine on $\boldsymbol{\gamma}$. Such dependencies are not indicated explicitly to avoid complicating the notation further.

Lemma 4.6. *Let $\mathcal{Q}^{\alpha\beta}_{uv}$ be as in* (4.36b) *and let d_F be the degree of $f(x;\boldsymbol{\gamma})$.*

(i) If $\alpha \neq k$ or $\beta \neq k$, there exists a matrix $\boldsymbol{Q}^{\alpha\beta}_{uv}(\boldsymbol{\gamma}) \in \mathbb{S}^{2(k+M+1)}$, whose entries are affine in $\boldsymbol{\gamma}$, such that $\mathcal{Q}^{\alpha\beta}_{uv} = \boldsymbol{\psi}_M{}^{\mathsf{T}} \, \boldsymbol{Q}^{\alpha\beta}_{uv}(\boldsymbol{\gamma}) \, \boldsymbol{\psi}_M$.

(ii) If $\alpha = \beta = k$ and $d_F \geq 1$, let $\overline{M} := M + 1 - d_F$, define $\boldsymbol{\Delta} \in \mathbb{S}^{d_F}$ as

$$\boldsymbol{\Delta} := \operatorname{diag}\left(\frac{2}{2(M+1)+1}, \ldots, \frac{2}{2(M+d_F)+1}\right), \tag{4.40}$$

and define $\boldsymbol{Y}(\boldsymbol{\gamma}) \in \mathbb{R}^{2d_F \times 2d_F}$ as

$$\boldsymbol{Y}(\boldsymbol{\gamma}) := \frac{1}{2}\begin{bmatrix} \boldsymbol{0} & \boldsymbol{\Phi}^{[M+1,M+d_f]}_{[M+1-d_F,M]} \\ \boldsymbol{\Phi}^{[M+1,M+d_f]}_{[M+1-d_F,M]} & \boldsymbol{0} \end{bmatrix}. \tag{4.41}$$

Finally, let $\boldsymbol{Q}^{kk}_{uv} \in \mathbb{S}^{2d_F}$ and a diagonal matrix $\boldsymbol{\Sigma}^{kk}_{uv} \in \mathbb{S}^2$ satisfy the LMI

$$\boldsymbol{\Omega}(\boldsymbol{Q}^{kk}_{uv}, \boldsymbol{\Sigma}^{kk}_{uv}, \boldsymbol{\gamma}) := \begin{bmatrix} \boldsymbol{Q}^{kk}_{uv} & \boldsymbol{Y}(\boldsymbol{\gamma}) \\ \boldsymbol{Y}(\boldsymbol{\gamma})^{\mathsf{T}} & \boldsymbol{\Sigma}^{kk}_{uv} \otimes \boldsymbol{\Delta} \end{bmatrix} \succeq 0, \tag{4.42}$$

where $\otimes$ is the usual Kronecker product. Then, $\mathcal{Q}^{kk}_{uv}$ can be bounded as

$$\mathcal{Q}^{kk}_{uv} \geq -\boldsymbol{\psi}_M{}^{\mathsf{T}} \left(\boldsymbol{\Xi}_{\overline{M}}{}^{\mathsf{T}} \, \boldsymbol{Q}^{kk}_{uv} \, \boldsymbol{\Xi}_{\overline{M}}\right) \boldsymbol{\psi}_M - \int_{-1}^{1} \begin{bmatrix} U^k_M \\ V^k_M \end{bmatrix}^{\mathsf{T}} \boldsymbol{\Sigma}^{kk}_{uv} \begin{bmatrix} U^k_M \\ V^k_M \end{bmatrix} \mathrm{d}x. \tag{4.43}$$

Proof. See appendix A.5. □

Remark 4.3. The case $\alpha = \beta = k$, $d_F = 0$ need not be considered: $\mathcal{Q}_{uv}^{kk} = 0$ if $d_F = 0$ because the Legendre polynomials $\{L_n\}_{n=0}^{M}$ are orthogonal to the remainder functions U_M^k, V_M^k.

Remark 4.4. The LMI (4.42) was chosen such that (4.43), essentially its Schur complement condition, separates the contributions of $\boldsymbol{\psi}_M$, U_M^k and V_M^k. It will be demonstrated in section 4.6.3 that inequality (4.43) is a source of conservativeness. In practical implementations, to make (4.43) as sharp as possible, the matrices $\boldsymbol{Q}_{uv}^{kk}$ and $\boldsymbol{\Sigma}_{uv}^{kk}$ are considered auxiliary variables, to be optimised subject to (4.42).

Remark 4.5. Note that $\mathcal{Q}_{uv}^{\alpha\beta}$ can be represented exactly only by using all Legendre coefficients of $\partial^k u$, $\partial^k v$ up to order M. This is what motivates the use of the extended decomposition (4.30) for these functions. Note also that, instead of using the bound (4.43), one could write $\mathcal{Q}_{uv}^{kk}$ exactly using the vector $\boldsymbol{\psi}_{M+d_F}$. However, doing so is not useful because $\boldsymbol{\psi}_{M+d_F}$ is not decoupled from U_M^k, V_M^k: for instance, the Legendre coefficients $\hat{u}_{M+i}^k$, $1 \leq i \leq d_F$ appear in the definition of U_M^k.

Lemmas 4.5 and 4.6 show that, for any $\alpha, \beta \in \{0, \dots, k\}$, the terms $\mathcal{P}_{uv}^{\alpha\beta}$ and $\mathcal{Q}_{uv}^{\alpha\beta}$ can be either expressed exactly or bounded in terms of $\boldsymbol{\psi}_M$, U_M^k and V_M^k. If $\alpha = \beta = k$, (4.36c) also depends on U_M^k and V_M^k. The following result reveals that $\mathcal{R}_{uv}^{\alpha\beta}$ can be bounded using the same quantities when $\alpha \neq k$ or $\beta \neq k$.

Lemma 4.7. *Suppose $\alpha \neq k$ or $\beta \neq k$, and let $\hat{\boldsymbol{f}}(\boldsymbol{\gamma}) = [\hat{f}_1(\boldsymbol{\gamma}), , \cdots, \hat{f}_{d_F}(\boldsymbol{\gamma})]^{\mathsf{T}}$ be the vector of Legendre coefficients of the polynomial f. There exist a positive semidefinite matrix $\boldsymbol{R}_{uv}^{\alpha\beta} \in \mathbb{S}^{2(M+k+1)}$ with $\|\boldsymbol{R}_{uv}^{\alpha\beta}\|_{\mathrm{F}} \sim N^{\alpha+\beta-2k-1}$ and a positive definite matrix $\boldsymbol{\Sigma}_{uv}^{\alpha\beta} \in \mathbb{S}^2$ with $\|\boldsymbol{\Sigma}_{uv}^{\alpha\beta}\|_{\mathrm{F}} \sim N^{\alpha+\beta-2k}$ such that $\mathcal{R}_{uv}^{\alpha\beta}$ is bounded as*

$$\left|\mathcal{R}_{uv}^{\alpha\beta}\right| \leq \|\hat{\boldsymbol{f}}(\boldsymbol{\gamma})\|_1 \, \boldsymbol{\psi}_M{}^{\mathsf{T}} \boldsymbol{R}_{uv}^{\alpha\beta} \, \boldsymbol{\psi}_M + \|\hat{\boldsymbol{f}}(\boldsymbol{\gamma})\|_1 \int_{-1}^{1} \begin{bmatrix} U_M^k \\ V_M^k \end{bmatrix}^{\mathsf{T}} \boldsymbol{\Sigma}_{uv}^{\alpha\beta} \begin{bmatrix} U_M^k \\ V_M^k \end{bmatrix} \mathrm{d}x. \tag{4.44}$$

Proof. See appendix A.6. □

Remark 4.6. The scaling of the Frobenius norms of $\boldsymbol{R}_{uv}^{\alpha\beta}$ and $\boldsymbol{\Sigma}_{uv}^{\alpha\beta}$ with N reflects the fact that the magnitude of $\mathcal{R}_{uv}^{\alpha\beta}$ diminishes to zero as N is raised. Consequently, the conservativeness of the estimates in Lemma 4.7 can be reduced by simply increasing N.

4.4.3 A lower bound for $\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}$

Lemmas 4.5–4.7 can be combined to find a lower bounding functional $\mathcal{G}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}$ for the integral functional $\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}$ in (4.14). To account for the different cases in Lemma 4.6, terms with $\alpha = \beta = k$ are considered separately from terms with $\alpha \neq k$ or $\beta \neq k$.

Let $\boldsymbol{S}(x;\boldsymbol{\gamma})$ be the symmetric matrix obtained from the rows and columns of the matrix $\boldsymbol{F}(x;\boldsymbol{\gamma})$ in (4.14) corresponding to the entries $\partial^k u$ and $\partial^k v$ of $\mathcal{D}^{\boldsymbol{k}}\boldsymbol{w}$. The contribution of the terms with $\alpha = \beta = k$ to $\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}$ is

$$\int_{-1}^{1} \begin{bmatrix} \partial^k u \\ \partial^k v \end{bmatrix}^{\mathsf{T}} \boldsymbol{S}(x;\boldsymbol{\gamma}) \begin{bmatrix} \partial^k u \\ \partial^k v \end{bmatrix} \mathrm{d}x. \tag{4.45}$$

Assuming for generality that no entry of $\boldsymbol{S}(x;\boldsymbol{\gamma})$ is independent of x, it follows from Lemma 4.5 and part (ii) of Lemma 4.6 that

$$\int_{-1}^{1} \begin{bmatrix} \partial^k u \\ \partial^k v \end{bmatrix}^{\mathsf{T}} \boldsymbol{S}(x;\boldsymbol{\gamma}) \begin{bmatrix} \partial^k u \\ \partial^k v \end{bmatrix} \mathrm{d}x \geq \boldsymbol{\psi}_M{}^{\mathsf{T}} \boldsymbol{\Theta}_1 \boldsymbol{\psi}_M + \int_{-1}^{1} \begin{bmatrix} U_M^k \\ V_M^k \end{bmatrix}^{\mathsf{T}} \boldsymbol{\Theta}_2 \begin{bmatrix} U_M^k \\ V_M^k \end{bmatrix} \mathrm{d}x, \tag{4.46}$$

where

$$\boldsymbol{\Theta}_1 := \boldsymbol{P}_{uu}^{kk} + 2\boldsymbol{P}_{uv}^{kk} + \boldsymbol{P}_{vv}^{kk} - \boldsymbol{\Xi}_{\overline{M}}{}^{\mathsf{T}} \Big(\boldsymbol{Q}_{uu}^{kk} + 2\boldsymbol{Q}_{uv}^{kk} + \boldsymbol{Q}_{vv}^{kk} \Big) \boldsymbol{\Xi}_{\overline{M}}, \tag{4.47a}$$

$$\boldsymbol{\Theta}_2 := \boldsymbol{S} - \boldsymbol{\Sigma}_{uu}^{kk} - 2\boldsymbol{\Sigma}_{uv}^{kk} - \boldsymbol{\Sigma}_{vv}^{kk}. \tag{4.47b}$$

Both matrices $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ depend affinely on $\boldsymbol{\gamma}$ through the matrices $\boldsymbol{P}_{uu}^{kk}$, $\boldsymbol{P}_{uv}^{kk}$, and $\boldsymbol{P}_{vv}^{kk}$. The auxiliary variables $\boldsymbol{Q}_{uu}^{kk}$, $\boldsymbol{\Sigma}_{uu}^{kk}$, $\boldsymbol{Q}_{uv}^{kk}$, $\boldsymbol{\Sigma}_{uv}^{kk}$, $\boldsymbol{Q}_{vv}^{kk}$, and $\boldsymbol{\Sigma}_{vv}^{kk}$ introduced by Lemma 4.6 also appear linearly, and must satisfy suitable LMIs of the form (4.42). For notational convenience, let

$$\mathcal{Y} := \Big\{ \boldsymbol{Q}_{uu}^{kk}, \boldsymbol{\Sigma}_{uu}^{kk}, \boldsymbol{Q}_{uv}^{kk}, \boldsymbol{\Sigma}_{uv}^{kk}, \boldsymbol{Q}_{vv}^{kk}, \boldsymbol{\Sigma}_{vv}^{kk} \Big\} \tag{4.48}$$

be the list of all auxiliary variables, and combine the three LMIs that they must satisfy into the equivalent block-diagonal LMI

$$\overline{\boldsymbol{\Omega}}(\boldsymbol{\gamma}, \mathcal{Y}) := \begin{bmatrix} \boldsymbol{\Omega}(\boldsymbol{Q}_{uu}^{kk}, \boldsymbol{\Sigma}_{uu}^{kk}, \boldsymbol{\gamma}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}(\boldsymbol{Q}_{uv}^{kk}, \boldsymbol{\Sigma}_{uv}^{kk}, \boldsymbol{\gamma}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Omega}(\boldsymbol{Q}_{vv}^{kk}, \boldsymbol{\Sigma}_{vv}^{kk}, \boldsymbol{\gamma}) \end{bmatrix} \succeq 0. \tag{4.49}$$

All terms contributing to $\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}$ with $\alpha \neq k$ or $\beta \neq k$ can instead be lower bounded using Lemmas 4.5–4.7 to obtain expressions such as

$$\int_{-1}^{1} f\, \partial^\alpha u\, \partial^\beta v \,\mathrm{d}x \geq \boldsymbol{\psi}_M{}^{\mathsf{T}} \boldsymbol{\Theta}_{uv}^{\alpha\beta}\, \boldsymbol{\psi}_M - \|\hat{\boldsymbol{f}}(\boldsymbol{\gamma})\|_1 \int_{-1}^{1} \begin{bmatrix} U_M^k \\ V_M^k \end{bmatrix}^{\mathsf{T}} \boldsymbol{\Sigma}_{uv}^{\alpha\beta} \begin{bmatrix} U_M^k \\ V_M^k \end{bmatrix} \mathrm{d}x, \tag{4.50}$$

where

$$\boldsymbol{\Theta}_{uv}^{\alpha\beta} := \boldsymbol{P}_{uv}^{\alpha\beta} + \boldsymbol{Q}_{uv}^{\alpha\beta} - \|\hat{\boldsymbol{f}}(\boldsymbol{\gamma})\|_1 \boldsymbol{R}_{uv}^{\alpha\beta}. \tag{4.51}$$

Contrary to the matrices $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$, each $\boldsymbol{\Theta}_{uv}^{\alpha\beta}$ does not depend affinely on $\boldsymbol{\gamma}$ because the norm $\|\hat{\boldsymbol{f}}(\boldsymbol{\gamma})\|_1$ is a sum of absolute values of affine functions of $\boldsymbol{\gamma}$.

Equations (4.46) and (4.50) imply that there exist a matrix $\boldsymbol{Q}_M = \boldsymbol{Q}_M(\boldsymbol{\gamma}, \mathcal{Y}) \in \mathbb{S}^{2(k+M+1)}$ and a positive definite matrix $\boldsymbol{\Sigma}_M = \boldsymbol{\Sigma}_M(\boldsymbol{\gamma}, \mathcal{Y}) \in \mathbb{S}^2$ such that, for all $\boldsymbol{w}$,

$$\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\} \geq \boldsymbol{\psi}_M{}^{\mathsf{T}} \boldsymbol{Q}_M \boldsymbol{\psi}_M + \int_{-1}^{1} \begin{bmatrix} U_M^k \\ V_M^k \end{bmatrix}^{\mathsf{T}} [\boldsymbol{S}(x;\boldsymbol{\gamma}) - \boldsymbol{\Sigma}_M] \begin{bmatrix} U_M^k \\ V_M^k \end{bmatrix} \mathrm{d}x. \tag{4.52}$$

Note that $\boldsymbol{Q}_M$ and $\boldsymbol{\Sigma}_M$ are affine with respect to the variables listed in $\mathcal{Y}$ but are not affine in $\boldsymbol{\gamma}$ because, as noted above, they depend on absolute values of affine functions of $\boldsymbol{\gamma}$.

Remark 4.7. It is possible to improve on the generic lower bound (4.52) if, for at least one of the terms of the form $\int_{-1}^{1} f(x;\boldsymbol{\gamma}) \left|\partial^\alpha u\right|^2 \mathrm{d}x$ in $\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}$, the function f is non-negative for all values of the optimisation variable $\boldsymbol{\gamma}$. In this case, the term $\mathcal{R}_{uu}^{\alpha\alpha}$ is non-negative and so it can be dropped, rather than estimated using Lemma 4.7.

4.4.4 Projection onto the boundary conditions

The lower bound (4.52) holds for any continuously differentiable function $\boldsymbol{w}$, irrespectively of whether it satisfies the BCs prescribed on the function space H over which the positivity of the functional $\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}$ is of interest. Recalling (4.15), the BCs on H are given by the set of p homogeneous equations

$$\boldsymbol{A}\,\mathcal{B}^{\boldsymbol{l}}\boldsymbol{w} = \boldsymbol{0}. \tag{4.53}$$

To enforce as many BCs as possible in (4.52) and to sharpen the lower bound over the space H, one should rewrite (4.53) using Legendre expansions. Lemma 4.4 enables the expansion of the boundary values of the first $k-1$ derivatives of u and v, so let $\boldsymbol{P}$ be a permutation matrix such that

$$\mathcal{B}^{\boldsymbol{l}}\boldsymbol{w} = \boldsymbol{P} \begin{bmatrix} \mathcal{B}^{\boldsymbol{k-1}}\boldsymbol{w} \\ \mathcal{B}^{[\boldsymbol{k},\boldsymbol{l}]}\boldsymbol{w} \end{bmatrix}. \tag{4.54}$$

Then, (4.53) becomes

$$\boldsymbol{A}\boldsymbol{P} \begin{bmatrix} \mathcal{B}^{\boldsymbol{k-1}}\boldsymbol{w} \\ \mathcal{B}^{[\boldsymbol{k},\boldsymbol{l}]}\boldsymbol{w} \end{bmatrix} = \boldsymbol{0}. \tag{4.55}$$

A straightforward corollary of Lemma 4.4 and of (4.37) is that there exists a matrix $\boldsymbol{J}$ such that $\mathcal{B}^{\boldsymbol{k-1}}\boldsymbol{w} = \boldsymbol{J}\boldsymbol{\psi}_M$, so (4.55) can be rewritten as

$$\boldsymbol{K} \begin{bmatrix} \boldsymbol{\psi}_M \\ \mathcal{B}^{[\boldsymbol{k},\boldsymbol{l}]}\boldsymbol{w} \end{bmatrix} = \boldsymbol{0}, \quad \boldsymbol{K} := \boldsymbol{A}\boldsymbol{P} \begin{bmatrix} \boldsymbol{J} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}. \tag{4.56}$$

Any vector $\boldsymbol{\psi}_M$ satisfying (4.56) can be expressed in the form

$$\boldsymbol{\psi}_M = \boldsymbol{\Pi}_M \boldsymbol{\zeta}, \tag{4.57}$$

for some $\boldsymbol{\zeta} \in \mathbb{R}^p$, where p is the dimension of $\ker(\boldsymbol{K})$ and $\boldsymbol{\Pi}_M$ is a computable projection matrix. Note that $\boldsymbol{\Pi}_M$ may have linearly dependent columns, so the dimension of ζ may be reduced further. This is important for practical efficiency but it makes no difference to the following discussion, so the details are omitted to streamline the presentation.

Upon substituting (4.57) into (4.52), one concludes that if (4.49) holds, then $\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}$ is lower bounded over the space H defined in (4.15) as

$$\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\} \geq \boldsymbol{\zeta}^{\mathsf{T}} \boldsymbol{\Pi}_M^{\mathsf{T}} \boldsymbol{Q}_M \boldsymbol{\Pi}_M \boldsymbol{\zeta} + \int_{-1}^{1} \begin{bmatrix} U_M^k \\ V_M^k \end{bmatrix}^{\mathsf{T}} [\boldsymbol{S}(x;\boldsymbol{\gamma}) - \boldsymbol{\Sigma}_M] \begin{bmatrix} U_M^k \\ V_M^k \end{bmatrix} \mathrm{d}x. \tag{4.58}$$

From (4.56) and (4.57) it is also possible to formulate a set of BCs that constrain the remainder functions U_M^k and V_M^k. It is also known that U_M^k and V_M^k should be orthogonal to all Legendre polynomials of degree less than or equal to M. However, these conditions could not be enforced explicitly in (4.58) to obtain a stronger, but still useful, lower bound on $\mathcal{F}_{\boldsymbol{\gamma}}$. Consequently, U_M^k and V_M^k in (4.58) will be considered arbitrary functions.

4.4.5 Formulating an inner SDP relaxation

The integral inequality in (4.14) is satisfied if the right-hand side of (4.58) is non-negative for all $\boldsymbol{\zeta}$ and all functions U_M^k and V_M^k. Recalling that the bound (4.58) is valid only if (4.49) holds, one arrives at the following statement.

Proposition 4.8. *Let $M = M(N)$ be as in (4.31) for any integer N, and let $\mathcal{Y}$ be as in (4.48). The set $T_N^{\text{in}} \subset \mathbb{R}^s$ of values $\boldsymbol{\gamma} \in \mathbb{R}^s$ for which there exist $\mathcal{Y}$ such that*

$$\overline{\boldsymbol{\Omega}}(\boldsymbol{\gamma};\mathcal{Y}) \succeq 0, \tag{4.59a}$$

$$\boldsymbol{\Pi}_M{}^{\mathsf{T}} \boldsymbol{Q}_M(\boldsymbol{\gamma},\mathcal{Y}) \boldsymbol{\Pi}_M \succeq 0, \tag{4.59b}$$

$$\boldsymbol{S}(x;\boldsymbol{\gamma}) - \boldsymbol{\Sigma}_M(\boldsymbol{\gamma},\mathcal{Y}) \succeq 0, \quad \forall x \in [-1,1], \tag{4.59c}$$

is an inner approximation of the feasible set T of (4.14), meaning $T_N^{\text{in}} \subset T$.

Conditions (4.59b) and (4.59c) are only sufficient, not necessary, to make the right-hand side of (4.58) non-negative: as mentioned at the end of section 4.4.4, they do not take into account the boundary and orthogonality conditions on the remainder functions. However,

they are useful because they can be turned into tractable constraints. For example, (4.59b) is not an LMI because, as already remarked after (4.52), $\boldsymbol{Q}_M(\boldsymbol{\gamma}, \mathcal{Y})$ depends on absolute values of affine functions of $\boldsymbol{\gamma}$ as a consequence of Lemma 4.7. However, (4.59b) can be readily recast as an LMI by replacing each of these absolute values, say $|\hat{f}_n(\boldsymbol{\gamma})|$, with a slack variable t subject to the additional linear constraints $-t \leq \hat{f}_n(\boldsymbol{\gamma}) \leq t$ (Boyd & Vandenberghe, 2004, section 6.1.1). Moreover, (4.59c) is an LMI if the matrix $\boldsymbol{S}(x; \boldsymbol{\gamma})$ is independent of x, which is true in many interesting and non-trivial cases such as example 4.1. When $\boldsymbol{S}(x; \boldsymbol{\gamma})$ does depend on x, instead, (4.59c) is equivalent to the polynomial inequality

$$\boldsymbol{z}^\mathsf{T}[\boldsymbol{S}(x; \boldsymbol{\gamma}) - \boldsymbol{\Sigma}_M(\boldsymbol{\gamma}, \mathcal{Y})]\boldsymbol{z} \geq 0 \quad \forall (x, \boldsymbol{z}) \in [-1, 1] \times \mathbb{R}^2. \tag{4.60}$$

Although checking a polynomial inequality is generally NP-hard (Parrilo, 2003, section 2.1), condition (4.60) can be turned into an LMI through a SOS relaxation (Parrilo, 2003). Using the so-called $\mathcal{S}$-procedure (Tan & Packard, 2006), one introduces a tunable symmetric polynomial matrix $\boldsymbol{T}(x) \in \mathbb{S}^2$ and requires that the multivariate polynomials

$$p_1(x, \boldsymbol{z}) := \boldsymbol{z}^\mathsf{T}[\boldsymbol{S}(x; \boldsymbol{\gamma}) - \boldsymbol{\Sigma}_M(\boldsymbol{\gamma}, \mathcal{Y}) - (1 - x^2)\boldsymbol{T}(x)]\boldsymbol{z}, \tag{4.61a}$$

$$p_2(x, \boldsymbol{z}) := \boldsymbol{z}^\mathsf{T}\boldsymbol{T}(x)\boldsymbol{z}, \tag{4.61b}$$

are sums of squares. It is not difficult to see that these conditions imply (4.60) and, as explained in section 2.3, SOS constraints can be reformulated as LMIs.

Once (4.59a)–(4.59c) are turned into LMIs, an upper bound for the optimal value of (4.14) and, possibly, a feasible point that achieves it can be found by solving an SDP.

Theorem 4.9. *Let $M = M(N)$ be defined as in (4.31) for $N \in \mathbb{N}$, let $\mathcal{Y}$ be as in (4.48), and let $\boldsymbol{T}(x) \in \mathbb{S}^2$ be a tunable polynomial matrix. Let $p^\star$ be the optimal value of (4.14), and let $p_N^\star$ be the optimal value of the SDP*

$$\begin{aligned} \min_{\boldsymbol{\gamma}, \mathcal{Y}, \boldsymbol{T}(x)} \quad & \boldsymbol{c}^\mathsf{T}\boldsymbol{\gamma}, \\ s.t. \quad & \overline{\boldsymbol{\Omega}}(\boldsymbol{\gamma}; \mathcal{Y}) \succeq 0, \\ & {\boldsymbol{\Pi}_M}^\mathsf{T} \boldsymbol{Q}_M(\boldsymbol{\gamma}, \mathcal{Y})\, \boldsymbol{\Pi}_M \succeq 0, \\ & \boldsymbol{z}^\mathsf{T}\left[\boldsymbol{S}(x; \boldsymbol{\gamma}) - \boldsymbol{\Sigma}_M(\boldsymbol{\gamma}, \mathcal{Y}) - (1 - x^2)\boldsymbol{T}(x)\right]\boldsymbol{z} \text{ is SOS}, \\ & \boldsymbol{z}^\mathsf{T}\boldsymbol{T}(x)\boldsymbol{z} \text{ is SOS}. \end{aligned} \tag{4.62}$$

Then, $p^\star \leq p_N^\star$ and any feasible point for (4.62) is also feasible for (4.14). In particular, if the point $\boldsymbol{\gamma}_N^\star$ is optimal for (4.62), then it is feasible for (4.14) with objective value $p_N^\star$.

Remark 4.8. Contrary to the case of outer SDP relaxations described in section 4.3, one cannot generally prove that the optimal value of (4.62) converges monotonically to that of the original problem as N is increased. In fact, without further assumptions on the functional $\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}$ in (4.14), it is possible that (4.62) is infeasible for any N (meaning that $p_N^\star = +\infty$) even if (4.14) is feasible (so $p^\star < +\infty$). To confirm this, note that since the matrix $\boldsymbol{\Sigma}_M$ is positive definite, condition (4.59c) and its corresponding SOS relaxation are feasible only if $\boldsymbol{S}(x;\boldsymbol{\gamma})$ can be made sufficiently positive definite for all $x \in [-1,1]$. This is not always possible, and one simple example is the integral inequality

$$\int_{-1}^{1} \left[x^2(\partial u)^2 + (\partial v)^2 - \gamma uv\right] \mathrm{d}x \geq 0, \tag{4.63}$$

where u and v are subject to the Dirichlet BCs $u(-1) = u(1) = v(-1) = v(1) = 0$. This inequality is clearly feasible for $\gamma = 0$. Yet, (4.62) is infeasible for any N because $\boldsymbol{S}(x;\boldsymbol{\gamma}) = \left[\begin{smallmatrix} x^2 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ is not positive definite at $x = 0$. In fact, for this example *any* approach requiring estimates of tail terms of series expansions will necessarily be ineffective. Conversely, with the SOS method of Valmorbida *et al.* (2016) it was established that (4.63) is feasible for $|\gamma| \leq 2.2$. With the exception of such cases, however, the proposed inner SDP relaxations work well in practice, and this will be demonstrated by means of examples in section 4.6. It may even be possible to identify classes of integral inequalities for which the SDP (4.62) is not only provably feasible, but such that the sequence of upper bounds $\{p_N^\star\}_{N\in\mathbb{N}}$ converges to the optimal value of (4.14). This task, however, is left to future research.

4.5 Extensions

4.5.1 Inequalities with explicit dependence on boundary values

Integral inequalities arising from the study of PDEs are often derived from a weak formulation of the PDE, after integrating some terms by parts. Occasionally, the BCs are such that the boundary terms from such integrations by parts do not vanish. The results described in the previous sections should therefore be extended to quadratic homogeneous functionals that depend explicitly on boundary values. Such functionals take the general form

$$\begin{aligned}\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\} := \int_{-1}^{1} \Big[&\left(\mathcal{B}^{\boldsymbol{l}}\boldsymbol{w}\right)^{\mathsf{T}} \boldsymbol{F}_{\mathrm{bnd}}(x;\boldsymbol{\gamma})\, \mathcal{B}^{\boldsymbol{l}}\boldsymbol{w} \\ &+ \left(\mathcal{B}^{\boldsymbol{l}}\boldsymbol{w}\right)^{\mathsf{T}} \boldsymbol{F}_{\mathrm{mix}}(x;\boldsymbol{\gamma})\, \mathcal{D}^{\boldsymbol{k}}\boldsymbol{w} \\ &+ \left(\mathcal{D}^{\boldsymbol{k}}\boldsymbol{w}\right)^{\mathsf{T}} \boldsymbol{F}_{\mathrm{int}}(x;\boldsymbol{\gamma})\, \mathcal{D}^{\boldsymbol{k}}\boldsymbol{w} \Big] \mathrm{d}x, \end{aligned} \tag{4.64}$$

where $\boldsymbol{F}_{\text{int}}$, $\boldsymbol{F}_{\text{mix}}$ and $\boldsymbol{F}_{\text{bnd}}$ are matrices of polynomials of degree at most d_F of the form (4.13). Note that this functional reduces to that in (4.14) when $\boldsymbol{F}_{\text{mix}} = \boldsymbol{0}$, $\boldsymbol{F}_{\text{bnd}} = \boldsymbol{0}$ and $\boldsymbol{F}_{\text{int}} = \boldsymbol{F}$.

The extension of Theorem 4.2 is immediate, because the boundary values of a polynomial can be easily expressed in terms of the coefficients of the polynomial.

To extend Proposition 4.8 and Theorem 4.9, recall the definition of the permutation matrix $\boldsymbol{P}$ in (4.54). Upon integrating the known matrix $\boldsymbol{P}^{\mathsf{T}}\boldsymbol{F}_{\text{bnd}}(x;\boldsymbol{\gamma})\boldsymbol{P}$, it follows from (4.37) and Lemma 4.4 that there exists a symmetric matrix $\boldsymbol{Q}_M^{\text{bnd}}(\boldsymbol{\gamma})$ such that

$$\int_{-1}^{1} \left(\mathcal{B}^{\boldsymbol{l}}\boldsymbol{w}\right)^{\mathsf{T}} \boldsymbol{F}_{\text{bnd}}(x;\boldsymbol{\gamma})\, \mathcal{B}^{\boldsymbol{l}}\boldsymbol{w}\, \mathrm{d}x = \begin{bmatrix} \boldsymbol{\psi}_M \\ \mathcal{B}^{[\boldsymbol{k},\boldsymbol{l}]}\boldsymbol{w} \end{bmatrix}^{\mathsf{T}} \boldsymbol{Q}_M^{\text{bnd}}(\boldsymbol{\gamma}) \begin{bmatrix} \boldsymbol{\psi}_M \\ \mathcal{B}^{[\boldsymbol{k},\boldsymbol{l}]}\boldsymbol{w} \end{bmatrix}. \tag{4.65}$$

Moreover, let $\boldsymbol{g}(x;\boldsymbol{\gamma})$ be the column of the matrix $\boldsymbol{P}^{\mathsf{T}}\boldsymbol{F}_{\text{mix}}(x;\boldsymbol{\gamma})$ corresponding to the entry $\partial^\alpha u$ of $\mathcal{D}^{\boldsymbol{k}}\boldsymbol{w}$. Each element $g_i(x;\boldsymbol{\gamma})$ is a polynomial of degree at most d_F, written in the Legendre basis with coefficients $\hat{g}_{i,0}(\boldsymbol{\gamma})$, …, $\hat{g}_{i,d_F}(\boldsymbol{\gamma})$. Recalling from (4.27) that the Legendre expansion of $\partial^\alpha u$ has been decomposed using $N \geq d_F + k - 1$, one obtains

$$\begin{aligned} \int_{-1}^{1} g_i(x;\boldsymbol{\gamma})\partial^\alpha u\, \mathrm{d}x &= \sum_{m=0}^{d_F}\sum_{n=0}^{\infty} \hat{g}_{i,m}(\boldsymbol{\gamma})\hat{u}_n^\alpha \int_{-1}^{1} L_m\, L_n\, \mathrm{d}x \\ &= \left[2\hat{g}_{i,0}(\boldsymbol{\gamma}),\quad \frac{2\hat{g}_{i,1}(\boldsymbol{\gamma})}{3},\quad \ldots,\quad \frac{2\hat{g}_{i,d_F}(\boldsymbol{\gamma})}{2d_F+1}\right] \hat{\boldsymbol{u}}^\alpha_{[0,d_F]}. \end{aligned} \tag{4.66}$$

With the help of Lemma 4.3, (4.37) and Lemma 4.4 it is then possible to find a matrix $\boldsymbol{Q}_M^{\text{mix}}(\boldsymbol{\gamma})$ that satisfies

$$\begin{aligned} \int_{-1}^{1} \left(\mathcal{B}^{\boldsymbol{l}}\boldsymbol{w}\right)^{\mathsf{T}} \boldsymbol{F}_{\text{mix}}(x;\boldsymbol{\gamma})\mathcal{D}^{\boldsymbol{k}}\boldsymbol{w}\, \mathrm{d}x &= \begin{bmatrix} \mathcal{B}^{\boldsymbol{k-1}}\boldsymbol{w} \\ \mathcal{B}^{[\boldsymbol{k},\boldsymbol{l}]}\boldsymbol{w} \end{bmatrix}^{\mathsf{T}} \int_{-1}^{1} \boldsymbol{P}^{\mathsf{T}}\boldsymbol{F}_{\text{mix}}(x;\boldsymbol{\gamma})\mathcal{D}^{\boldsymbol{k}}\boldsymbol{w}\, \mathrm{d}x \\ &= \begin{bmatrix} \boldsymbol{\psi}_M \\ \mathcal{B}^{[\boldsymbol{k},\boldsymbol{l}]}\boldsymbol{w} \end{bmatrix}^{\mathsf{T}} \boldsymbol{Q}_M^{\text{mix}}(\boldsymbol{\gamma})\, \boldsymbol{\psi}_M. \end{aligned} \tag{4.67}$$

Note that (4.65) and (4.67) are exact formulae, and no approximation is made. Upon combining these results with (4.52), one can construct a symmetric matrix $\boldsymbol{Q}_M^{\text{tot}} = \boldsymbol{Q}_M^{\text{tot}}(\boldsymbol{\gamma}, \mathcal{Y})$ such that

$$\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\} \geq \begin{bmatrix} \boldsymbol{\psi}_M \\ \mathcal{B}^{[\boldsymbol{k},\boldsymbol{l}]}\boldsymbol{w} \end{bmatrix}^{\mathsf{T}} \boldsymbol{Q}_M^{\text{tot}} \begin{bmatrix} \boldsymbol{\psi}_M \\ \mathcal{B}^{[\boldsymbol{k},\boldsymbol{l}]}\boldsymbol{w} \end{bmatrix} + \int_{-1}^{1} \begin{bmatrix} U_M^k \\ V_M^k \end{bmatrix}^{\mathsf{T}} [\boldsymbol{S}(x;\boldsymbol{\gamma}) - \boldsymbol{\Sigma}_M] \begin{bmatrix} U_M^k \\ V_M^k \end{bmatrix} \mathrm{d}x. \tag{4.68}$$

Finally, the BCs can be enforced by using (4.56) to write

$$\begin{bmatrix} \boldsymbol{\psi}_M \\ \mathcal{B}^{[\boldsymbol{k},\boldsymbol{l}]}\boldsymbol{w} \end{bmatrix} = \boldsymbol{\Lambda}\boldsymbol{\zeta} \tag{4.69}$$

for some $\boldsymbol{\zeta} \in \mathbb{R}^p$ where, as before, p is the dimension of $\ker(\boldsymbol{K})$ and the matrix $\boldsymbol{\Lambda}$ satisfies $\operatorname{img}(\boldsymbol{\Lambda}) = \ker(\boldsymbol{K})$. Thus, Proposition 4.8 and Theorem 4.9 hold also for problems with integral inequalities of the form (4.64) if one replaces (4.59b) and the corresponding constraint in (4.62) with $\boldsymbol{\Lambda}^\mathsf{T}\boldsymbol{Q}_M^{\mathrm{tot}}(\boldsymbol{\gamma},\mathcal{Y})\boldsymbol{\Lambda} \succeq 0$.

4.5.2 Higher-dimensional function spaces, generic multi-index derivatives

Theorems 4.2 and 4.9 were derived under the simplifying assumption 4.1, which restricted attention to $\boldsymbol{w} \in C^m([-1,1],\mathbb{R}^2)$ and to the particular multi-indices $\boldsymbol{k} = [k,k]$, $\boldsymbol{l} = [l,l]$. All results discussed so far, including the extensions presented in section 4.5.1, hold also when one lets $\boldsymbol{w} \in C^m([-1,1],\mathbb{R}^q)$ with $q \geq 1$ and when $\boldsymbol{k}, \boldsymbol{l} \in \mathbb{N}^q$ are generic multi-indices, as long as they satisfy (4.12a) and (4.12b). In particular, if the q-dimensional multi-indices $\boldsymbol{k}$ and $\boldsymbol{l}$ are uniform, meaning $\boldsymbol{k} = [k,\ldots,k]$ and $\boldsymbol{l} = [l,\ldots,l]$, then the proofs of all results extend *verbatim* upon identifying the functions u, v used throughout sections 4.3 and 4.4 with any two components w_i, w_j of $\boldsymbol{w}$. The extension to non-uniform multi-indices $\boldsymbol{k}, \boldsymbol{l} \in \mathbb{N}^q$ requires only minor and largely uninteresting modifications, so the details are left to the interested reader.

4.6 Computational experiments with QUINOPT

This section demonstrates the efficacy of the SDP-based solution techniques developed in this chapter on some simple but non-trivial examples. The relevant SDPs were formulated using QUINOPT (QUadratic INtegral OPTimisation), an open-source add-on for the MATLAB optimisation toolbox YALMIP (Löfberg, 2004, 2009). QUINOPT implements the methods presented in sections 4.3 and 4.4, including the extensions discussed in section 4.5. The outer SDP relaxations of section 4.3 are implemented using Legendre polynomials because their orthogonality promotes sparsity of the SDP data, thereby improving efficiency. The source code, the scripts used to implement the problems considered in the following subsections, and an online documentation with additional examples are available from

https://github.com/aeroimperial-optimization/QUINOPT (source code),
http://quinopt.readthedocs.io/ (documentation).

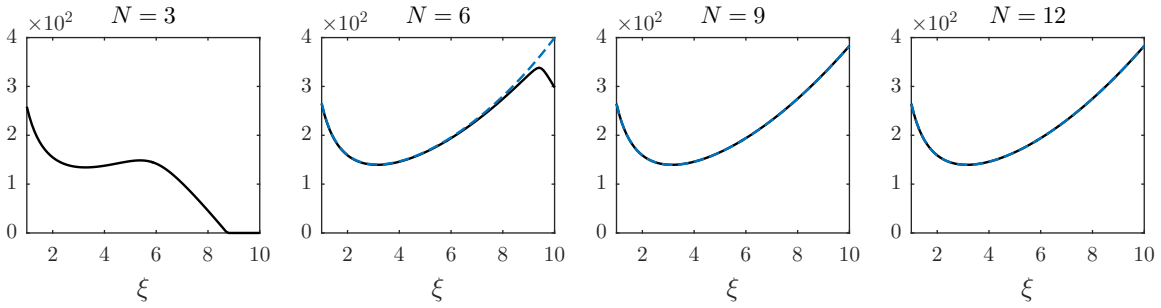


FIGURE 4.2: Lower bounds (——) and upper bounds (- - -) on the optimal value of (4.16) as a function of ξ for different values of N. The upper bound for $N = 3$ is infinite and so it is not plotted. The bounds are indistinguishable for $N = 9$ and $N = 12$.

Computations were carried out on a PC with a 3.40 GHz Intel® Core™ i7-4770 CPU and 16 GB of RAM, using MOSEK (Andersen *et al.*, 2009) or SDPT3 (Toh *et al.*, 1999; Tütüncü *et al.*, 2003) to solve the SDPs.

4.6.1 Stability of a stress-driven shear flow

Consider the system described in example 4.1, which is a model of flows driven by a shear stress of magnitude 0.5γ. The optimisation problem (4.16) determines the largest value of γ such that sinusoidal perturbations from the basic steady flow decay. QUINOPT was used to formulate and solve SDPs (4.24) and (4.62) corresponding, respectively, to outer and innner approximations of (4.16). Since in (4.16) one minimises the *negative* of γ, the optimal values of these SDPs are, respectively, an upper bound and a lower bound for γ_{cr}, the largest stress for which the basic flow is provably stable.

Figure 4.2 shows the bounds on γ_{cr} as a function of the wavenumber ξ, a parameter in (4.16). These were computed for four different values of the Legendre series truncation parameter N. No upper bounds are plotted for $N = 3$ because, in this case, only the zero polynomial satisfies the BCs in (4.17), so (4.24) reduces to an unconstrained minimisation problem yielding an infinite upper bound. Detailed numerical results, wall time, and the number of primal and dual variables in the SDP relaxations[2] (denoted n and m, respectively) are reported in table 4.1 for two values of the wavenumber, $\xi = 3$ and $\xi = 9$. For comparison, table 4.2 gives lower bounds on γ_{cr} computed with the inner SOS relaxation method of Valmorbida *et al.* (2016) using polynomials of degree d, as well as the number of primal and dual variables in the corresponding SDPs returned by YALMIP's SOS module (Löfberg, 2009) and the wall time required to solve them.

[2]The number of primal variables, n, refers to the size of the positive semidefinite matrix variable once the SDP is written in the standard primal form (2.13), cf. section 2.4. The number of dual variables, m, is the size of the optimisation variable in the dual SDP (2.19).

TABLE 4.1: Parameters of the outer and inner SDP relaxations of problem (4.16) formulated with QUINOPT, as a function of the Legendre truncation parameter N and the wavenumber ξ. Tabulated values are: upper and lower bounds on γ_{cr} (LB and UB), wall time (t, in seconds), number of primal variables (n), and number of dual variables (m).

	QUINOPT, outer					QUINOPT, inner				
ξ	N	n	m	UB	t	N	n	m	LB	t
3	3	0	1	$+\infty$	0.03	3	202	2	134.8594	0.09
3	6	36	1	140.4087	0.04	6	406	2	139.7656	0.10
3	9	144	1	139.7701	0.06	9	683	2	139.7700	0.08
3	12	324	1	139.7700	0.05	12	1030	2	139.7700	0.10
9	3	0	1	$+\infty$	0.03	3	202	2	0.0000	0.08
9	6	36	1	335.1022	0.04	6	406	2	323.5764	0.08
9	9	144	1	325.6764	0.05	9	683	2	325.6449	0.09
9	12	324	1	325.6455	0.05	12	1030	2	325.6453	0.10

TABLE 4.2: Parameters of the inner SDP relaxations of problem (4.16) formulated with the SOS method of Valmorbida *et al.* (2016) using polynomials of degree d. Values, tabulated for different d, are: lower bounds on γ_{cr} (LB), wall time (t, in seconds), number of primal variables (n), and number of dual variables (m).

$\xi = 3$					$\xi = 9$				
d	n	m	LB	t	d	n	m	LB	t
4	805	230	79.4435	0.19	4	805	230	285.9021	0.18
8	2349	454	119.8619	0.28	8	2349	454	314.1146	0.27
16	7789	902	130.5796	0.68	16	7789	902	321.2403	0.65
32	28077	1798	134.4737	3.16	32	28077	1798	323.1421	2.98

The results show that, at least within the tested range of ξ, the upper and lower bounds converge to each other at relatively small values of N (values in table 4.1 agree to at least three decimal places for $N = 12$). This means that the techniques described in this chapter can bound the optimal solution γ_{cr} of (4.16) both from above and from below extremely accurately and efficiently. In addition, in this example the inner SDP relaxations converge to the full problem (4.16) despite the lack of a proof of this fact in general (cf. remark 4.8). Finally, the present techniques significantly outperform the SOS method of Valmorbida *et al.* (2016), both in terms of computational cost and of quality of bounds.

4.6.2 Stability of a system of coupled PDEs

Let $\boldsymbol{w} = [u(t,x),\, v(t,x)]^{\mathsf{T}}$ and consider the system of PDEs

$$\partial_t \boldsymbol{w} = \gamma \partial_x^2 \boldsymbol{w} + \boldsymbol{A}\boldsymbol{w}, \qquad \boldsymbol{A} = \begin{bmatrix} 1 & 1.5 \\ 5 & 0.2 \end{bmatrix}, \tag{4.70}$$

over the domain $[0,1]$, subject to the BCs $u(0) = u(1) = v(0) = v(1) = 0$. This system was also studied by Valmorbida *et al.* (2014*b*, section V-D). The stabilising effect of the diffusive term $\gamma\partial_x^2\boldsymbol{w}$ decreases with γ, until the equilibrium solution $[u,v]^\mathsf{T} = [0,0]^\mathsf{T}$ becomes unstable. It can be shown that the amplitude of infinitesimal sinusoidal perturbations to the zero solution grows exponentially in time if $\gamma < \gamma_{\mathrm{cr}} \approx 0.3412$. Since the system is linear, it is stable with respect to finite-amplitude perturbations for all $\gamma \geq \gamma_{\mathrm{cr}}$.

Following Valmorbida *et al.* (2014*b*), the stability of the system with respect to arbitrary perturbations is investigated by means of Lyapunov functionals of the form

$$\mathcal{V}(t) = \frac{1}{2}\int_0^1 \boldsymbol{w}^\mathsf{T}\boldsymbol{P}(x)\boldsymbol{w}\,\mathrm{d}x, \tag{4.71}$$

where $\boldsymbol{P}(x)$ is a tunable polynomial matrix of given degree d_P to be chosen such that, for some $c > 0$,

$$\mathcal{V}(t) \geq c\,\|\boldsymbol{w}\|_2^2\,, \tag{4.72a}$$

$$-\frac{\mathrm{d}\mathcal{V}}{\mathrm{d}t} \geq 0. \tag{4.72b}$$

Note that since $\boldsymbol{P}(x)$ can always be rescaled by c without changing the sign of the inequalities, one may fix $c = 1$ without any loss of generality.

After using (4.70) to compute $\frac{\mathrm{d}\mathcal{V}}{\mathrm{d}t}$, it is relatively easy to see that the critical value of γ at which (4.71) stops being a valid Lyapunov function for a given degree d_P is given by

$$\begin{aligned} \min_{\gamma,\boldsymbol{P}(x)} \quad & \gamma \\ \text{s.t.} \quad & \int_0^1 \boldsymbol{w}^\mathsf{T}\left[\boldsymbol{P}(x) - \boldsymbol{I}\right]\boldsymbol{w}\,\mathrm{d}x \geq 0, \\ & \int_0^1 \boldsymbol{w}^\mathsf{T}\boldsymbol{P}(x)(-\gamma\partial_x^2\boldsymbol{w} - \boldsymbol{A}\boldsymbol{w})\,\mathrm{d}x \geq 0. \end{aligned} \tag{4.73}$$

Note that although the system state $\boldsymbol{w}$ is a function of time, the integral inequalities above are imposed *pointwise in time*. Therefore, the time dependence can be formally dropped, and (4.73) is in the form (4.14) with two integral inequalities.

Since the optimisation variables are γ and the coefficients of the entries of $\boldsymbol{P}(x)$, the problem is not jointly convex in γ and $\boldsymbol{P}$, so one cannot minimise γ directly. The problem can be readily resolved by fixing a trial value for γ and checking whether a feasible $\boldsymbol{P}(x)$ of degree d_P exists. The optimal γ for (4.73), which must be finite because the system is linearly unstable when $\gamma = 0$, is the value at which a feasible $\boldsymbol{P}(x)$ ceases to exist, and can be determined with a simple bisection procedure.

TABLE 4.3: Upper bounds (UB) and lower bounds (LB) on the optimal solution of (4.73) obtained with Lyapunov functionals of the form (4.70) for different values d_P, and for the case $\boldsymbol{P}(x) = \boldsymbol{I}$. Also reported are: the average wall time (t_{UB} and t_{LB}, in seconds) required to solve a single feasibility problem in the bisection procedure to compute the bounds; the wall time required to minimise γ in the case $\boldsymbol{P}(x) = \boldsymbol{I}$; upper bounds on the optimal solution of (4.73) computed by Valmorbida *et al.* (2014*b*).

d_P	UB by Valmorbida *et al.* (2014*a*)	UB	t_{UB}	LB	t_{LB}	
$\boldsymbol{P}(x) = \boldsymbol{I}$	5	0.3925	0.26		0.3925	0.09
0	3.3333	0.3412	0.14		0.3412	0.12
2	0.5882	0.3412	1.32		0.3412	0.99
4	0.4347	0.3412	1.57		0.3412	1.07
6	0.4166	0.3412	1.82		0.3412	1.18

To formulate the SDPs (4.24) and (4.62), the domain of integration for the constraints in (4.73) must be rescaled to $[-1, 1]$. Moreover, in light of remark 4.8, the second integral inequality should be integrated by parts to prevent the inner SDP relaxation (4.62) from being infeasible. Both tasks (rescaling and integration by parts) are performed automatically by QUINOPT. Note that after rescaling and integration by parts the second integral inequality in (4.73) depends explicitly on the unspecified boundary values $\partial_x u(\pm 1)$ and $\partial_x v(\pm 1)$, making the extensions discussed in section 4.5.1 necessary.

Table 4.3 reports upper and lower bounds for the optimal solution of (4.73) as a function of the degree d_P of $\boldsymbol{P}(x)$, obtained by applying the bisection procedure described above to the SDPs (4.62) and (4.24) respectively. Also tabulated are results for the particular choice $\boldsymbol{P}(x) = \boldsymbol{I}$, corresponding to the classical approach of taking the energy of the system as the candidate Lyapunov function. In this case, a direct minimisation over γ could be performed. Finally, the table reports the average wall time taken by QUINOPT to set up and solve each feasibility problem in the bisection procedure used to optimise γ, and to minimise γ for the particular case $\boldsymbol{P}(x) = \boldsymbol{I}$. In all computations, the Legendre series truncation parameter was set to $N = 10$, while the degree of the matrix $\boldsymbol{T}(x)$ in (4.62) was set to 6. No significant change is observed when either of these parameters is increased.

The numerical results demonstrate that stability can be established up to the known critical value $\gamma_{\text{cr}} \approx 0.3412$ for all choices of d_P, but not when one chooses $\boldsymbol{P}(x) = \boldsymbol{I}$. This drastically improves the conservative results, also reported in table 4.3, obtained by Valmorbida *et al.* (2014*b*) for the same problem using their SOS method (the original results are for a parameter $R = \gamma^{-1}$ and have been adapted). One concludes that the SDP relaxations proposed in this chapter accurately approximate (4.73). This observation is particularly significant for the inner SDP (4.62), which relies on typically conservative estimates and for which no convergence result could be proven.

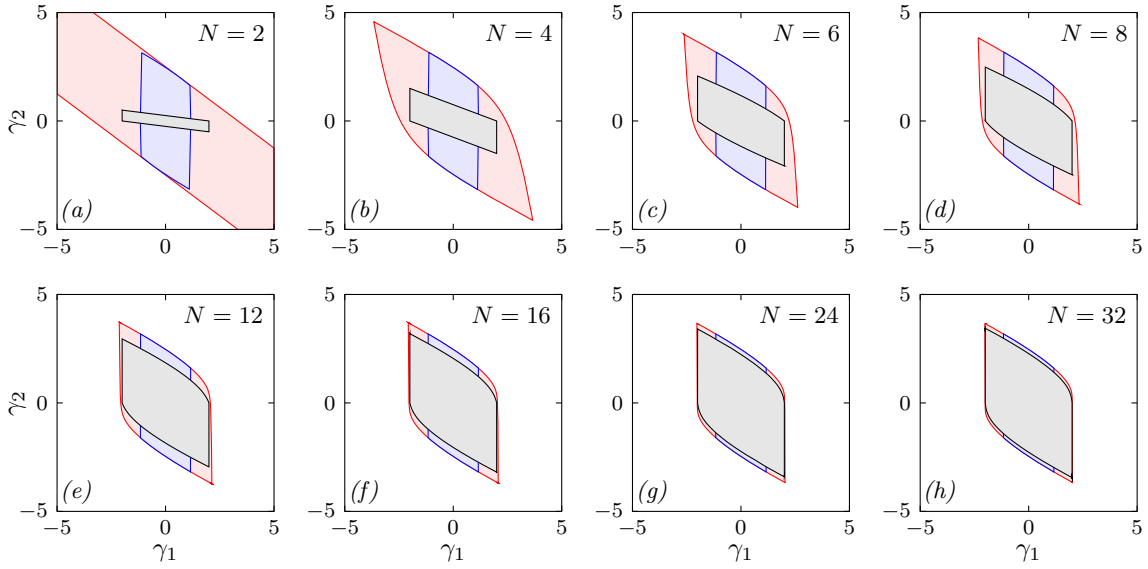


FIGURE 4.3: Inner and outer approximations of the feasible set of (4.74), computed with MOSEK (Andersen *et al.*, 2009). Sets plotted in each panel are: T_N^{in}, blue boundary with blue shading; T_N^{out}, red boundary with red shading; T_N^{sos}, black boundary with gray shading. Panels *(a)*–*(h)* are for $N = 2$, 4, 6, 8, 12, 16, 24, and 32, respectively.

4.6.3 Feasible set approximation

As a final example, consider the problem of computing the entire unknown feasible set of the integral inequality

$$\int_{-1}^{1} \left[(\partial u)^2 + (\partial v)^2 + \gamma_1 x^2 \partial u\, \partial v + 2\gamma_2 uv \right] \mathrm{d}x \geq 0, \tag{4.74}$$

where u and v are subject to the Dirichlet BCs $u(-1) = 0$, $u(1) = 0$, $v(-1) = 0$, $v(1) = 0$. This inequality does not arise from a particular PDE, but was constructed to illustrate some subtle properties of the proposed SDP relaxations and highlight the main sources of conservativeness.

Outer and inner approximation sets T_N^{out} and T_N^{in} can be found using (4.24) and (4.62), respectively. The boundaries of T_N^{out} and T_N^{in} were approximated by optimising the function $\gamma_1 \sin\theta + \gamma_2 \cos\theta$ for 300 equispaced values of $\theta \in [0, 2\pi]$. When solving (4.62), the degree of the tunable polynomial matrix $\boldsymbol{T}(x)$ was fixed to the smallest of $N-2$ and 6; the results do not improve when this value is increased. Inner approximation sets can also be computed with the SOS method of Valmorbida *et al.* (2016). The set obtained when this method is applied with polynomials of degree N will be denoted T_N^{sos} in the following.

Figure 4.3 illustrates the approximations T_N^{out}, T_N^{in}, and T_N^{sos} obtained for for six values of N starting with $N = 2$, the minimum value satisfying (4.27). Computations were repeated with two SDP solvers, MOSEK (Andersen *et al.*, 2009) and SDPT3 (Toh *et al.*,

TABLE 4.4: Wall time (in seconds) for the computation of the sets T_N^{out}, T_N^{in}, and T_N^{sos} as a function of N. Tabulated values are for two SDP solvers: MOSEK (Andersen *et al.*, 2009) and SDPT3 (Toh *et al.*, 1999; Tütüncü *et al.*, 2003).

	MOSEK			SDPT3		
N	T_N^{out}	T_N^{in}	T_N^{sos}	T_N^{out}	T_N^{in}	T_N^{sos}
2	0.55	1.41	1.09	10.5	25.4	21.5
4	0.73	2.19	2.69	11.6	30.9	45.1
6	0.82	2.29	5.90	12.1	35.5	86.0
8	1.22	2.75	12.8	14.5	42.9	167
12	1.88	3.37	56.6	16.1	51.4	194
16	2.91	4.11	218	21.1	58.2	600
24	6.65	5.79	1656	25.5	59.0	10500
32	11.1	8.52	5106	35.5	71.5	117000

1999; Tütüncü *et al.*, 2003), and the required wall time is reported in table 4.4. Note that MOSEK is implemented in C, while SDPT3 is implemented in MATLAB. Surprisingly, MOSEK computes T_N^{in} more efficiently than T_N^{out} at large N, despite the latter being nominally cheaper. This is not the case for SDPT3, and the reason for these observations could not be determined with certainty. However, it was observed that the computation of T_N^{out} requires solving SDPs with denser data matrices, which could cause MOSEK to run slower.

Evidently, for high-degree relaxations the SOS approach developed by Valmorbida *et al.* (2016) is much more computationally expensive than the methods proposed here. On the other hand, while T_N^{sos} seems to converge to T_N^{out} as N increases, the inner approximation sets T_N^{in} do not: for this example, the conservativeness introduced by the estimates in Lemmas 4.6 and 4.7 cannot be reduced by raising N.

Nevertheless, there are parts where the boundaries of T_N^{out} and T_N^{in} almost coincide even for N as low as 4. In fact, the figures reveal that the inner approximation sets T_N^{in} are only over-constrained in the γ_1 direction. This happens because γ_2 appears only in the term $\int_{-1}^{1} 2\gamma_2 uv \,\mathrm{d}x$, to which the estimates of Lemma 4.7 are applied when formulating the inner SDP relaxation. According to the decay rates stated in the Lemma, these estimates become negligible at large N. On the contrary, γ_1 appears in the term $\int_{-1}^{1} \gamma_1 x^2 \partial u \, \partial v \,\mathrm{d}x$, to which the estimates in part (ii) of Lemma 4.6 must be applied. Despite efforts to tune the auxiliary matrices in (4.43), the magnitude of such estimates does not decay compared to other terms in the SDP as N is increased, limiting the range of feasible values of γ_1.

4.7 Comments on computational cost

It may be checked that when $\boldsymbol{w}(x) \in \mathbb{R}^q$ is subject to p independent boundary conditions, the degree of the polynomials in the matrix $\boldsymbol{F}(x;\boldsymbol{\gamma})$ is at most d_F, and $\boldsymbol{\gamma} \in \mathbb{R}^s$, then the

N-th outer relaxation (4.24) is an SDP with an LMI of linear dimension $q(N+1)-p$ and with s decision variables. The inner SDP relaxation (4.62), instead, has:

(i) an LMI of dimension $2|\boldsymbol{l}| + q(N + |\boldsymbol{k}|_\infty + d_F + 2) - p$, where $|\boldsymbol{k}|_\infty := \max_{i\in\{1,\ldots,q\}} k_i$;

(ii) a $q \times q$ matrix SOS constraint of degree $\deg \boldsymbol{S}(x;\boldsymbol{\gamma})$, where $\boldsymbol{S}(x;\boldsymbol{\gamma})$ is defined as in section 4.4.3;

(iii) at most $q(q+1)/2$ auxiliary LMIs of size $4\,d_F$ from Lemma 4.6;

(iv) at most $(d_F+1)(2\,q+|\boldsymbol{k}|+1)|\boldsymbol{k}|$ linear inequalities to lift the absolute values introduced by Lemma 4.7; and

(v) at most $s + q(q+1)(2\,d_F^2 + d_F + 2)/2 + (d_F+1)(2\,q+|\boldsymbol{k}|+1)|\boldsymbol{k}|/2$ decision variables.

At the time of writing, general-purpose solvers are practical only for small to medium-size SDPs. Therefore, even though the numerical examples of section 4.6 demonstrate that the techniques developed in this chapter are cheaper than the SOS method of Valmorbida *et al.* (2016), one might expect that they can be implemented only when q, d_F, s and $|\boldsymbol{k}|_\infty$ are sufficiently small.

This is true to a certain extent, but the current poor scalability of software for SDPs may not be too severe an issue for many problems of practical interest. One reason is that the number of constraints in the inner SDP relaxation can be considerably smaller than the worst-case count presented above. In fact, Lemma 4.6 introduces auxiliary LMIs and variables only for the entries in the upper-triangular part of $\boldsymbol{S}(x;\boldsymbol{\gamma})$ that depend on x (the restriction to the upper-triangular part follows from symmetry considerations). For example, only one auxiliary LMI is needed for inequality (4.74). In addition, the size of the auxiliary LMI associated with the entry $\boldsymbol{S}_{ij}$ can be reduced to $4 \times \deg \boldsymbol{S}_{ij}$, yielding considerable savings if $\deg \boldsymbol{S}(x;\boldsymbol{\gamma}) \ll d_F$. In the extreme case $\deg \boldsymbol{S}(x;\boldsymbol{\gamma}) = 0$, meaning that the matrix $\boldsymbol{S}(x;\boldsymbol{\gamma})$ is independent of x, there are no auxiliary variables and LMIs from Lemma 4.6. In this case, moreover, the $q \times q$ matrix SOS constraint becomes a $q \times q$ LMI, which is cheaper to implement. This situation is common when energy-Lyapunov-function methods are applied to turbulent fluid flows (see for instance Constantin & Doering, 1995*b*; Doering & Constantin, 1994, 1996), so the techniques developed in this chapter are particularly suited to tackle problems in this field. This has already been demonstrated by the results of section 4.6.1, and further evidence will be given in chapter 5.

Another reason why only medium-size SDP relaxations are needed in many applications is that a moderate Legendre truncation parameter N often suffices to obtain accurate bounds on the objective function of problem (4.14). Approximately speaking, to obtain a good bound one should choose N such that the minimiser $\boldsymbol{w}^\star$ of $\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}$ at the optimal point

$\boldsymbol{\gamma} = \boldsymbol{\gamma}^\star$ is approximated sufficiently well by a polynomial of degree N (here it is assumed for simplicity that the minimiser $\boldsymbol{w}^\star$ exists and is unique, but the argument can be extended to cases in which multiple or no minimisers exist). This can often be done with moderate N because $\boldsymbol{w}^\star$ is typically a "well-behaved" function: the highest-order derivatives of highly oscillatory test functions $\boldsymbol{w}$ tend to give a large contribution to $\mathcal{F}_{\boldsymbol{\gamma}^\star}\{\boldsymbol{w}\}$, making highly-oscillatory minimisers unlikely.

Finally, computational efficiency can be improved by applying decomposition techniques based on chordal sparsity (Fukuda *et al.* 2000; Nakata *et al.* 2003; Kim *et al.* 2011; see also section 2.6). The development of solvers for large scale SDPs, which take advantage of chordal sparsity as well as other "computationally friendly" structures, is also receiving increasing attention, and new tools are being developed that should facilitate solving problems at larger scales. Recent examples include the solvers SCS (O'Donoghue *et al.*, 2016) and CDCS (Zheng *et al.*, 2017*a*,*b*).

4.8 Concluding remarks

This chapter presented SDP-based methods to optimise a linear cost function subject to a class of quadratic integral inequality constraints, characterised by one-dimensional compact integration domains, homogeneous integrands, and affine dependence on the optimisation variable. This class includes many problems that arise when studying the stability of systems governed by PDEs, as well as when bounding time-averaged or long-term quantities of interest using the background method. In fact, the proposed solution techniques extend the approach taken by Fantuzzi & Wynn (2015) to bound the asymptotic energy of the Kuramoto–Sivashinsky equation. It has been shown that, given an optimisation problem subject to integral inequality constraints with the properties summarised above, LMI-representable inner and outer approximations of its feasible set can be derived using Legendre series expansions and functional estimates. As a result, upper and lower bounds on the optimal value can be computed efficiently using semidefinite programming. In particular, the lower bounds obtained using outer approximations form a non-decreasing sequence converging to the exact optimal cost value, provided that this is attained. Unfortunately, the same is not true in general for the inner approximations, as confirmed by a simple counterexample (cf. remark 4.8).

Although the steps leading to both inner and outer SDP relaxations are technical, they are amenable to numerical implementation. To aid the formulation and solution of optimisation problems with integral inequality constraints in practice, the methods developed in

this chapter have been implemented in the MATLAB package QUINOPT, an open-source add-on for the optimisation toolbox YALMIP. The capabilities of the software have been demonstrated on non-trivial problems that arise when studying the stability of autonomous systems of PDEs. In particular, the numerical results presented in section 4.6 are clear evidence that the proposed methodologies can work extremely well in practice even though the formulation of numerically tractable constraints relies on typically conservative estimates. Future research should try to formalise these observations and, in particular, determine conditions under which the feasibility and/or convergence of the inner SDP relaxations of section 4.4 can be ensured. The results of section 4.6.3 suggest that doing so successfully will require more stringent assumptions on the properties of the integral inequality than assumed throughout this chapter.

While the class of integral inequalities considered in this chapter includes many non-trivial problems, in particular some interesting problems arising in fluid dynamics, the analysis of systems governed by PDEs often leads to more general types of integral inequality constraints. To enable the study of such systems, the present work should be extended to *(i)* integral inequalities with explicit time dependence that arise from non-autonomous PDEs, and *(ii)* inequalities over two- or higher-dimensional domains. Explicit polynomial time dependence could be dealt with by relaxing the inner/outer LMI constraints, now time-dependent, into matrix SOS conditions, although the (current) poor scalability of SOS optimisation makes this strategy unlikely implementable. Multi-dimensional compact "box" domains could be analysed by introducing Legendre expansions in each coordinate direction and adapting the ideas presented in this chapter, while for more general domains—including the non-compact case—other basis functions could be used. Doing so seems fairly straightforward in the case of outer approximations, but the estimates required to derive inner approximations may become intractable. In addition, unless sparsity and/or problem structure are exploited, multi-dimensional inequalities are likely to be constrained by the current computational limitations: with n spatial dimensions and q dependent variables ($\boldsymbol{w} \in \mathbb{R}^q$), the LMI size for a simple outer approximations using polynomials of degree N is of the order of magnitude of qN^n.

Finally, the methods presented in this chapter should be extended to more general integral inequalities than the homogeneous quadratic type. Complete (*i.e.*, inhomogeneous) quadratic integral inequalities over spaces described by homogeneous BCs can be analysed with ideas similar to those discussed in section 4.5.1, and can already be handled by QUINOPT. The details are not reported in this thesis for brevity. Inhomogeneous BCs need not be studied, because they can always be "lifted" upon shifting the test functions in

an integral inequality by a suitable polynomial. Extensions to higher-than-quadratic integral inequalities are also essential if complex nonlinear systems of PDEs of interest in physics and engineering are to be studied successfully using recent techniques based on dissipation inequalities (Ahmadi *et al.*, 2016).

Rather than pursuing these extensions, however, the rest of this thesis will focus on applying and further developing the methods presented in this chapter to study two particular fluid dynamical systems. In chapter 5, QUINOPT will be utilised to compute optimal upper bounds on the energy dissipated by two- and three-dimensional shear flows driven by a surface stress, very similar to the flow considered in example 4.1. It will demonstrated that the optimal bounds can be computed accurately, but not across as wide a range of system parameters as required to infer their asymptotic behaviour. In chapter 6, instead, SDPs will be used to bound the average vertical heat transfer in Bénard–Marangoni convection at infinite Prandtl number. These computation will rely on a slightly different outer approximation method than described in section 4.3, which yields LMIs with chordal sparsity. The LMI decomposition methods described in section 2.6 can therefore be employed, enabling the analysis of the flow for a much larger range of system parameters.

Chapter 5

Bounds on energy dissipation in stress-driven shear flows[†]

As discussed in chapter 1, a fundamental problem in many applications is to determine the extent to which unsteady or turbulent flows enhance the dissipation of energy, the transport of heat, or the mixing of a passive tracer compared to steady ones. Often, direct quantitative analysis is challenging due to the lack of closed-form solutions to the Navier–Stokes equations and due to the large computational cost of high-resolution numerical simulations. In some cases, progress can be made using indirect methods that return upper or lower bounds for the quantity of interest (dissipation, heat transport, mixing) instead of its exact value.

In the context of incompressible parallel shear flows the key quantity of interest is the average bulk energy dissipation ε (equivalently, the non-dimensional dissipation coefficient C_ε), which in a statistically steady state must equal the average power required to drive the flow. Direct evaluation of ε for all possible flow states is not feasible, but an upper bound on it can be derived using the background method. The analysis rests on the decomposition of the flow velocity into the sum of a steady background velocity $\boldsymbol{\phi} = \phi(z)\boldsymbol{e}_1$ aligned with the streamwise direction, which absorbs any inhomogeneous boundary conditions (BCs) but is otherwise arbitrary, plus an incompressible perturbation $\tilde{\boldsymbol{u}}$. Using an argument similar to energy stability analysis, the dissipation coefficient C_ε can be bounded from above as a function of ϕ alone, provided that this satisfies all prescribed BCs and makes a certain integral quadratic form $\mathcal{Q}\{\tilde{\boldsymbol{u}}\}$ positive semidefinite for all possible perturbations. The latter condition is a spectral constraint—meaning that it requires the eigenvalues of a ϕ-dependent, self-adjoint, linear operator associated with the quadratic form $\mathcal{Q}\{\tilde{\boldsymbol{u}}\}$ to be non-negative—and amounts to an energy stability condition on ϕ as if it solved the governing equations.

[†]Results similar to those reported in this chapter, computed using analysis similar to that presented in chapter 4 and with an *ad-hoc* numerical implementation, have been published in:
Fantuzzi, G. and Wynn, A. (2016). Optimal bounds with semidefinite programming: An application to stress driven shear flows. *Physical Review E* **93**(4), 043308. Available from: doi:10.1103/PhysRevE.93.043308.

The classical plane Couette configuration, in which the flow of a horizontal layer of fluid is driven by an imposed surface velocity, has been analysed extensively using the background method, analytically (see, for example, Doering & Constantin, 1992, 1994; Marchioro, 1994; Nicodemus *et al.*, 1997*a*,*b*, 1998) as well as numerically (Doering & Hyman, 1997; Nicodemus *et al.*, 1997*b*, 1998; Plasting & Kerswell, 2003). Flows driven by a shear stress imposed at the surface, instead, have been studied much less despite being relevant both as a paradigm for shear-driven phenomena and as a model of geophysical flows forced by winds (Hagstrom & Doering, 2014). For such stress-driven flows, Tang *et al.* (2004) found that $C_\varepsilon \leq Gr\,(7.531\,Gr^{0.5} - 20.3)^{-2}$, where the non-dimensional Grashoff number Gr measures the strength of the imposed shear stress. This bound, valid for $Gr \gtrsim 500$, is a fit to numerical results obtained when the background method is applied to a slightly different flow, wherein the applied stress is modelled using a body force localised near the surface. A bounding problem that incorporates the fixed-shear boundary condition was not formulated until the recent work by Hagstrom & Doering (2014), who used piecewise-linear background fields to prove $C_\varepsilon \leq 1/16$ for $Gr \geq 16$ in two spatial dimensions, and $C_\varepsilon \leq 1/(2\sqrt{2})$ uniformly in Gr for three-dimensional flows.

The aim of this chapter is to compute the best bounds on C_ε available within Hagstrom & Doering's background method analysis, which are of interest for two reasons. First, in order to confirm that stress-driven flows can indeed be approximated well by flows driven by a suitably chosen body force, a quantitative comparison of the two configurations is needed. Hagstrom & Doering (2014) have already demonstrated that the approximation is valid in terms of the energy stability properties of the laminar flow, which is the usual Couette profile varying linearly with depth (cf. section 5.1 below), and one would like to confirm that the same is true for the energy dissipation. Second, numerical optimisation of background fields for flows driven by shear—meaning that the flow has either Neumann or mixed Neumann–Robin inhomogeneous BCs—has so far only been attempted by Wittenberg & Gao (2010), who optimised piecewise-linear background temperature fields for Rayleigh-Bénard convection between imperfectly conducting plates. Fully optimal background fields for flows subject to imposed boundary fluxes, however, have never been computed.

The solution of upper-bounding problems for shear-driven flows has traditionally been seen as challenging because the expression for the bound depends on unknown boundary values of the background field (Tang *et al.*, 2004; Wittenberg, 2010; Wittenberg & Gao, 2010; Hagstrom & Doering, 2010, 2014). Similarly, the test functions for which the spectral constraint must hold have prescribed boundary derivatives, instead of fixed boundary values. Classical approaches, based on the solution of the Euler–Lagrange (EL) equations

for the optimal background field, are therefore often complicated by the need to enforce so-called *natural boundary conditions* (see, for instance, Courant & Hilbert, 1953, chapter IV, section 5.1), which are not prescribed initially, but arise when insisting that the variational derivatives of the problem's Lagrangian must vanish. This difficulty will be bypassed here because, instead of considering the EL equations, optimal bounds will be computed through the solution of SDPs, obtained upon replacing the spectral constraint with a set of LMIs derived with the Legendre series expansions methods developed in chapter 4.

The rest of this chapter is organised as follows. Section 5.1 introduces the equations used to model stress-driven shear flows. These will be presented in three spatial dimensions, but a two-dimensional model of the flow obtained by removing the horizontal direction normal to the imposed shear will also be considered. The background method analysis by Hagstrom & Doering (2014), which yields a variational problem for an upper bound on the dissipation coefficient C_ε, is reviewed in section 5.2. Section 5.3 demonstrates how bounds on C_ε for the two- and three-dimensional flow models can be optimised via semidefinite programming. The LMIs used to enforce the spectral constraint on the background field can be formulated using the analysis presented in chapter 4 and can be constructed in practice with QUINOPT, so their derivation will only be outlined. Numerical results are presented and commented on in section 5.4, while further discussion and conclusions are offered in section 5.5.

5.1 Equations of motion

Consider an incompressible fluid of kinematic viscosity ν and density ρ, confined to the three-dimensional domain $\Omega_3 \equiv [0, \Gamma_x h] \times [0, \Gamma_y h] \times [0, h]$, where h is the dimensional height of the layer and Γ_x, Γ_y are the domain's aspect ratios in the horizontal directions. The dimensional position vector is $\boldsymbol{x}_\star = x_\star \boldsymbol{e}_1 + y_\star \boldsymbol{e}_2 + z_\star \boldsymbol{e}_3$, where $\boldsymbol{e}_i$ is the unit vector in the i-th coordinate direction and the suffix $\star$ indicates dimensional quantities. The fluid's velocity and pressure satisfy no-slip conditions at the bottom boundary ($z_\star = 0$) and are periodic in the horizontal directions. The flow is driven by a shear stress τ applied at the top boundary ($z_\star = 1$) along the $x_\star$ direction. This configuration is illustrated in figure 5.1.

Following Tang *et al.* (2004), the problem is made non-dimensional using h as the length scale and h^2/ν as the time scale. Then, the relevant non-dimensional Navier–Stokes equations take the form

$$\partial_t \boldsymbol{u} + (\boldsymbol{u} \cdot \nabla)\boldsymbol{u} + \nabla p = \nabla^2 \boldsymbol{u}, \tag{5.1a}$$

$$\nabla \cdot \boldsymbol{u} = 0. \tag{5.1b}$$

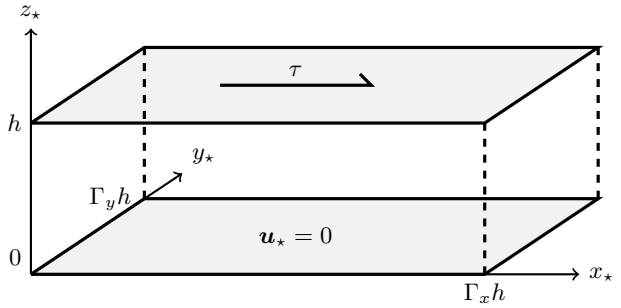


FIGURE 5.1: Three-dimensional model of a shear flow, driven by a shear stress τ at $z_\star = h$. The fluid's velocity and pressure are periodic in the $x_\star$ and $y_\star$ directions with period $\Gamma_x h$ and $\Gamma_y h$, respectively, and satisfy no-slip condition at the bottom plate.

The non-dimensional velocity $\boldsymbol{u} = u\boldsymbol{e}_1 + v\boldsymbol{e}_2 + w\boldsymbol{e}_3$ has period Γ_x and Γ_y in the horizontal x and y directions, respectively, and satisfies the vertical BCs

$$\boldsymbol{u}|_{z=0} = 0, \qquad \partial_z u|_{z=1} = Gr, \qquad \partial_z v|_{z=1} = 0, \qquad w|_{z=1} = 0. \tag{5.2}$$

In these expressions, the Grashoff number $Gr := \tau h^2/(\rho\nu^2)$ is a non-dimensional measure of the strength of the imposed shear stress, and serves as the governing parameter of the flow.

When subject to horizontal periodicity and the vertical BCs (5.2), equations (5.1a)–(5.1b) admit the laminar flow solution $\boldsymbol{u}_\ell = Gr\, z\, \boldsymbol{e}_1$, $p =$ constant. Note that $\boldsymbol{u}_\ell$ is the usual Couette profile, which varies linearly with depth. Energy stability analysis demonstrates that the laminar flow is globally asymptotically stable, meaning that the energy of perturbations of arbitrary initial amplitude decays to zero as time tends to infinity, when $Gr \lessapprox 51.7300$ (Tang *et al.*, 2004; Hagstrom & Doering, 2014). This value was computed for a horizontally infinite layer, so it is a (sharp) lower bound on the critical Grashoff number Gr_E for energy stability of a fluid layer with finite horizontal periods Γ_x and Γ_y, because in this case perturbations are characterised only by a finite number of Fourier modes. For instance, under the assumption that the critical modes are streamwise invariant (hence, independent of Γ_x) one finds $Gr_E \approx 57.1989$ for $\Gamma_y = 2$ and $Gr_E \approx 51.7305$ for $\Gamma_y = 3$. More details on energy stability analysis for finite periodic layers are given in appendix B.

The occurrence of unsteady and possibly turbulent flows cannot be excluded when $Gr \geq Gr_E$.[1] Irrespective of the steady or unsteady nature of the flow, its intensity can be quantified by the average bulk energy dissipation rate per unit mass,

$$\varepsilon := \langle \nu \, \|\nabla_\star \boldsymbol{u}_\star\|^2 \rangle = \frac{\nu^3}{h^4} \langle \|\nabla \boldsymbol{u}\|^2 \rangle, \tag{5.3}$$

[1]Linear stability analysis was not carried out, but the close similarity with the classical plane Couette flow suggests that the laminar flow is likely to be linearly stable at all Grashoff numbers.

where $\nabla_\star$ is the dimensional gradient and angle brackets denote an average over volume and infinite time (cf. section 1.2). The non-dimensional quantity associated with ε is the dissipation coefficient

$$C_\varepsilon := \frac{\varepsilon h}{\overline{u_\star}(h)^3} = \frac{Gr}{\overline{u}(1)^2}, \tag{5.4}$$

where overlines denote horizontal and infinite-time averages (cf. section 1.2). The second equality in (5.4) follows after space-time averaging the dot product of (5.1a) with $\boldsymbol{u}$ and integrating by parts with the help of incompressibility and the BCs to prove the identity

$$\langle \|\nabla \boldsymbol{u}\|^2 \rangle = Gr\, \overline{u}(1). \tag{5.5}$$

Finally, the three-dimensional model described above can be rendered two-dimensional simply by removing any variables, equations, and BCs associated with the y direction. The laminar flow $\boldsymbol{u}_\ell = Gr\, z\, \boldsymbol{e}_1$ remains a steady solution, and is globally stable for $Gr \lessapprox 139.5396$ (Hagstrom & Doering, 2014). Again, this value is a (sharp) lower bound on the critical Grashoff number Gr_E for energy stability in finite periodic domains. For example, it is shown in appendix B that $Gr_E \approx 139.5399$ for $\Gamma_x = 2$ and $Gr_E \approx 148.6624$ for $\Gamma_x = 3$. The definitions of ε and C_ε are unchanged, provided that volume averaging is understood over the two-dimensional domain $\Omega_2 \equiv [0, \Gamma_x] \times [0, 1]$.

5.2 Bounds on the dissipation coefficient

The dissipation coefficient C_ε can be bounded from above at any Grashoff number using the background method (Hagstrom & Doering, 2014). The analysis, both in two and three dimensions, begins by letting the velocity of the fluid be decomposed as $\boldsymbol{u} = \phi(z)\boldsymbol{e}_1 + \tilde{\boldsymbol{u}}$, where the background field $\phi(z)$ is subject to the inhomogeneous BCs

$$\phi(0) = 0, \qquad \phi'(1) = Gr. \tag{5.6}$$

The perturbation field $\tilde{\boldsymbol{u}}$ is periodic in the horizontal directions, satisfies

$$\tilde{\boldsymbol{u}}|_{z=0} = 0, \qquad \partial_z \tilde{u}|_{z=1} = 0, \qquad \partial_z \tilde{v}|_{z=1} = 0, \qquad \tilde{w}|_{z=1} = 0, \tag{5.7}$$

and is governed by the Navier–Stokes equations in perturbation form,

$$\partial_t \tilde{u} + (\tilde{\boldsymbol{u}} \cdot \nabla)\tilde{\boldsymbol{u}} + \nabla p = \nabla^2 \tilde{\boldsymbol{u}} + \partial_z^2 \phi\, \boldsymbol{e}_1 - \phi\, \partial_x \tilde{\boldsymbol{u}} - \partial_z \phi\, \tilde{w}\, \boldsymbol{e}_1, \tag{5.8a}$$

$$\nabla \cdot \tilde{\boldsymbol{u}} = 0. \tag{5.8b}$$

Space-time averaging the dot product of $\tilde{\boldsymbol{u}}$ and (5.8a), followed by suitable integration by parts using the incompressibility condition (5.8b) and the homogeneous BCs in (5.7), shows that (cf. equation (37) in Hagstrom & Doering, 2014)

$$\langle \|\nabla \tilde{\boldsymbol{u}}\|^2 + \phi' \, \partial_z \tilde{u} + \phi' \, \tilde{u} \, \tilde{w} \rangle - Gr \, \overline{u}(1) + Gr \, \phi(1) = 0. \tag{5.9}$$

Moreover, substituting the background decomposition $\boldsymbol{u} = \phi(z)\boldsymbol{e}_1 + \tilde{\boldsymbol{u}}$ into (5.5) yields

$$Gr \, \overline{u}(1) = \langle \|\nabla \tilde{\boldsymbol{u}}\|^2 + 2 \, \phi' \, \partial_z \tilde{u} \rangle + \int_0^1 \left| \phi'(z) \right|^2 \mathrm{d}z. \tag{5.10}$$

Subtracting $2\times$(5.9) from the right-hand side of (5.10) to eliminate the term $\phi' \, \partial_z \tilde{u}$ and noticing that $\phi(1) = \int_0^1 \phi'(z) \, \mathrm{d}z$ by virtue of the BC $\phi(0) = 0$ gives, after some rearrangement,

$$\overline{u}(1) = Gr^{-1} \langle \|\nabla \tilde{\boldsymbol{u}}\|^2 + 2 \, \phi' \, \tilde{u} \, \tilde{w} \rangle + \int_0^1 2 \, \phi'(z) - Gr^{-1} \left| \phi'(z) \right|^2 \mathrm{d}z. \tag{5.11}$$

At this stage, consider the space of smooth perturbations

$$H := \left\{ \boldsymbol{u} \in C^\infty(\Omega_d, \mathbb{R}^d) \, : \, \boldsymbol{u} \text{ is horizontally periodic and satisfies (5.8b), (5.7)} \right\}, \tag{5.12}$$

where $d = 2$ for the two-dimensional flow model and $d = 3$ for the three-dimensional one, and suppose that ϕ is chosen such that

$$\mathcal{Q}\{\boldsymbol{u}\} := \langle \|\nabla \boldsymbol{u}\|^2 + 2 \, \phi' \, u \, w \rangle \geq 0 \quad \forall \boldsymbol{u} \in H, \tag{5.13a}$$

$$\mathcal{B}\{\phi\} := \int_0^1 Gr^{-1} \left| \phi'(z) \right|^2 - 2 \, \phi'(z) \, \mathrm{d}z < 0. \tag{5.13b}$$

Then, $\mathcal{Q}\{\tilde{\boldsymbol{u}}\} \geq 0$ for any perturbation[2] and one can bound $\overline{u}(1) \geq -\mathcal{B}\{\phi\} > 0$. Using (5.4), one concludes that the dissipation coefficient is bounded from above according to

$$C_\varepsilon \leq \frac{Gr}{|\mathcal{B}\{\phi\}|^2}. \tag{5.14}$$

Condition (5.13a) is a spectral constraint on ϕ, and is needed to ensure that the term involving the unknown perturbation $\tilde{\boldsymbol{u}}$ in (5.11) can be dropped. Any background field that satisfies the spectral constraint subject to the BCs (5.6) will be called *feasible* and, if in addition $\mathcal{Q}\{\boldsymbol{u}\}$ is strictly positive for all non-zero $\boldsymbol{u} \in H$, then ϕ will be called *strictly*

[2]That perturbations governed by (5.8a)–(5.8b) are smooth at all times if the initial condition is smooth is unproven. Formally, one should consider weak solutions in a suitable Sobolev space, and use a density argument to prove that the non-negativity of $\mathcal{Q}\{\boldsymbol{u}\}$ for the smooth functions in H suffices to conclude non-negativity for all weak solutions. This is not done here under the assumption that the velocity field of any physically relevant incompressible flow is a smooth function of space.

feasible. Note that, although strictly speaking $\mathcal{Q}\{\boldsymbol{u}\}$ is defined as an average over both space and infinite time, any time dependence can be ignored when testing its non-negativity on H, and $\langle\cdot\rangle$ can be interpreted as an average over the volume only. Thus, the spectral constraint (5.13a) is a multi-dimensional integral inequality constraint on ϕ.

Condition (5.13b), instead, is simply needed to guarantee that the lower bound on $\overline{u}(1)$ is positive, otherwise (5.14) cannot be deduced from (5.4). For the purposes of optimising the background field, however, the negativity of $\mathcal{B}\{\phi\}$ need not be imposed because there exist feasible background fields—such as those constructed by Hagstrom & Doering (2014), or smooth approximations thereof—for which the condition holds. The optimal ϕ is therefore guaranteed to make $\mathcal{B}\{\phi\}$ negative definite.

Finally, note that both the bound (5.14) and the spectral constraint depend only on the derivative of the background field. This means that, for any background field such that ϕ' satisfies the spectral constraint, the BC $\phi(0) = 0$ can always be enforced by adding a suitable constant to ϕ without affecting the value of $\mathcal{B}\{\phi\}$, and so it can be dropped. Consequently, the best bound on C_ε available to the background method analysis described above is found upon solving the variational problem

$$\begin{aligned} \min_{\phi} \quad & \mathcal{B}\{\phi\} = \int_0^1 Gr^{-1} \left|\phi'(z)\right|^2 - 2\,\phi'(z)\,\mathrm{d}z \\ \text{s.t.} \quad & \mathcal{Q}\{\boldsymbol{u}\} = \langle \|\nabla \boldsymbol{u}\|^2 + 2\,\phi'\,u\,w\rangle \geq 0 \quad \forall \boldsymbol{u} \in H, \\ & \phi'(1) = Gr. \end{aligned} \tag{5.15}$$

This problem takes the same form for both the two- and the three-dimensional flow models, provided that the average in the definition of the quadratic form $\mathcal{Q}\{\boldsymbol{u}\}$ and the definition test function space H are understood in the appropriate dimension. The analysis needed to replace the spectral constraint with a set of sufficient finite-dimensional conditions, however, differs slightly depending on the model's dimension.

Remark 5.1. It is almost immediate to check that $\mathcal{B}\{\phi\} = Gr^{-1}\,\|\phi' - Gr\|_2^2 - Gr$. Hence, to minimise the objective in (5.15) one would like to choose $\phi'(z) = Gr$ for all $z \in [0,1]$, and hence $\phi(z) = Gr\,z$. On the other hand, in order to be able to control the sign-indefinite term in $\mathcal{Q}\{\boldsymbol{u}\}$ and satisfy the spectral constraint, one requires that $\phi' \approx 0$ across the domain except near the boundaries, where w is small since it must vanish at the walls. This means that, at least at large Gr, one expects the optimal background field to be characterised by two boundary layers, in which $\phi'(z) \approx Gr$. In particular, one expects the boundary condition $\phi'(1) = Gr$ to be satisfied even when it is not imposed explicitly in (5.15).

Remark 5.2. The condition for energy stability of the laminar flow at a given Grashoff number Gr_0 is (Hagstrom & Doering, 2014; see also appendix B)

$$\langle \|\nabla \boldsymbol{u}\|^2 + Gr_0\, u\, w \rangle \geq 0 \quad \forall \boldsymbol{u} \in H. \tag{5.16}$$

At Grashoff number Gr ($\neq Gr_0$), choosing $\phi(z) = Grz$ means that the spectral constraint (5.13a) reduces to (5.16) with $Gr_0 = 2Gr$. In other words, the spectral constraint for $\phi(z) = Grz$ requires that the laminar flow profile is energy stable at *twice* the imposed Grashoff number. Consequently, the laminar flow is a feasible background field for $Gr \leq Gr_E/2$, where Gr_E is the critical Grashoff number for energy stability. Moreover, since the laminar flow minimises $\mathcal{B}\{\phi\}$ (cf. remark 5.1), it must be the optimal background field for all $Gr \leq Gr_E/2$, while for $Gr > Gr_E/2$ the optimal bounds on C_ε must be larger than the laminar dissipation value $1/Gr$. This observation is useful to verify that bounds computed with a finite-dimensional approximation of (5.15) converge to the true optimal solution as the number of degrees of freedom included in the finite-dimensional problem increases.

5.3 Optimal bounds via semidefinite programming

For both the two- and the three-dimensional flow models, the background field problem (5.15) can be implemented numerically as an SDP if three obstacles are overcome. The first one is that the optimisation variable in (5.15) is infinite-dimensional because it is a function, but in computations one can only consider finitely many decision variables. The second hurdle is that SDPs minimise a linear objective function, but the objective in (5.15) is quadratic even after restricting attention to background field described by a finite-dimensional parametrisation. Finally, one must find a way of enforcing the spectral constraint via computationally tractable conditions.

5.3.1 Parametrisation of the background field

The first step to reduce (5.15) to any computationally tractable problem is to restrict the attention to background fields ϕ that admit a finite-dimensional parametrisation. The optimal ϕ must be at least continuous and one expects it to be smooth, but it is unlikely that it has a finite-dimensional representation in any of the traditional expansion bases for the space of smooth functions. Nonetheless, an extension of the Weierstrass theorem (Peet & Bliman, 2007) guarantees that the optimal ϕ can be approximated arbitrarily accurately by a degree-P polynomial that satisfies the BCs in (5.6), provided that P is sufficiently

large. Therefore, solving (5.15) over degree-P polynomial background fields—which clearly admit a finite-dimensional representation—constitutes only a mild restriction so long as the value P required to obtain a good approximation is not too large. In fact, the (sub)optimal objective values computed in this way form a non-increasing sequence that converges to the true optimal value of (5.15) as P is raised.

Since the background field enters (5.15) only via its derivative, it is convenient to introduce a polynomial ansatz for ϕ' directly, and recover ϕ by integration using the BC $\phi(0) = 0$ when needed. In order to simplify the following analysis, it is also convenient to work in the Legendre polynomial basis and write

$$\phi'(z) = \sum_{n=0}^{P-1} \hat{\phi}_n \, L_n(2z-1), \tag{5.17}$$

where $L_n(\xi)$, $\xi \in [-1,1]$, is the Legendre polynomial of degree n (cf. section 4.1).

With this ansatz, the vector $\hat{\boldsymbol{\phi}} := [\hat{\phi}_0, \ldots, \hat{\phi}_{P-1}]^\mathsf{T} \in \mathbb{R}^P$ becomes the optimisation variable in (5.15), and it must be chosen to enforce the spectral constraint and the BC $\phi'(1) = Gr$. The former will be discussed in sections 5.3.3 and 5.3.4 for the two- and three-dimensional flows, respectively. The latter, instead, can be rewritten in terms of the vector $\hat{\boldsymbol{\phi}}$ after recalling from section 4.1 that Legendre polynomials take the boundary value $L_n(1) = 1$ for all $n \geq 0$. It is then not difficult to see that, for background fields defined through (5.17),

$$\phi'(1) = Gr \quad \Leftrightarrow \quad \mathbf{1}^\mathsf{T} \hat{\boldsymbol{\phi}} = Gr. \tag{5.18}$$

5.3.2 Formulation of a linear objective

When the polynomial ansatz (5.17) is introduced in the optimisation problem (5.15), the orthogonality condition (4.9) for the Legendre polynomials can be used to express the objective function as

$$\mathcal{B}\{\phi\} = Gr^{-1} \, \hat{\boldsymbol{\phi}}^\mathsf{T} \boldsymbol{B} \hat{\boldsymbol{\phi}} - 2 \, \hat{\phi}_0, \tag{5.19}$$

where

$$\boldsymbol{B} := \operatorname{diag}\left(2, \frac{2}{3}, \ldots, \frac{2}{2n+1}, \ldots, \frac{2}{2(P-1)+1}\right). \tag{5.20}$$

Since algorithms for semidefinite programming minimise a linear function, the quadratic objective $\mathcal{B}\{\phi\}$ must be replaced with a linear one that bounds it from above. This can be done without introducing any conservativeness by introducing a so-called *slack variable* s such that

$$\hat{\boldsymbol{\phi}}^\mathsf{T} \boldsymbol{B} \hat{\boldsymbol{\phi}} \leq s. \tag{5.21}$$

One can then minimise the linear function $s - 2\,\hat{\phi}_0$, which is a sharp upper bound for the original objective $\mathcal{B}\{\phi\}$ by virtue of (5.21). Note that the quadratic constraint (5.21) is convex because $\boldsymbol{B}$ is a positive definite matrix and, as explained in section 2.3, one has

$$\hat{\boldsymbol{\phi}}^{\mathsf{T}}\boldsymbol{B}\hat{\boldsymbol{\phi}} \le s \quad \Leftrightarrow \quad \boldsymbol{S}(\hat{\boldsymbol{\phi}}, s) := \begin{bmatrix} \boldsymbol{B}^{-1} & \hat{\boldsymbol{\phi}} \\ \hat{\boldsymbol{\phi}}^{\mathsf{T}} & s \end{bmatrix} \succeq 0. \tag{5.22}$$

Consequently, when attention is restricted to degree-P polynomial background fields defined through (5.17), problem (5.15) is equivalent to

$$\begin{aligned} \min_{\hat{\boldsymbol{\phi}},\, s} \quad & s - 2\,\hat{\phi}_0 \\ \text{s.t.} \quad & \mathcal{Q}\{\boldsymbol{u}\} \ge 0 \quad \forall \boldsymbol{u} \in H, \\ & \boldsymbol{S}(\hat{\boldsymbol{\phi}}, s) \succeq 0, \\ & \mathbf{1}^{\mathsf{T}}\hat{\boldsymbol{\phi}} = Gr. \end{aligned} \tag{5.23}$$

All that is left to do to make this optimisation problem an SDP is to replace the spectral constraint with a finite number of LMIs or, more generally, of LMI-representable constraints. The analysis for the two-dimensional flow model differs slightly from that needed in three dimensions, so the two cases will be treated separately.

5.3.3 Analysis of the spectral constraint: two-dimensional flows

In two dimensions, each function $\boldsymbol{u} \in H$ can be expanded using the Fourier series

$$\boldsymbol{u}(x, z) = \sum_{m \in \mathbb{Z}} \boldsymbol{U}_m(z)\, \mathrm{e}^{i\alpha_m x}, \tag{5.24}$$

where $\alpha_m := 2\pi m/\Gamma_x$ is the m-th horizontal wavenumber and each z-dependent Fourier amplitude $\boldsymbol{U}_m(z) := U_m(z)\boldsymbol{e}_1 + W_m(z)\boldsymbol{e}_3$ is a complex-valued vector field. The requirement that the Fourier modes combine into the real-valued function $\boldsymbol{u}$ implies that $\boldsymbol{U}_{-m} = \boldsymbol{U}_m^*$ (both here and in the following, the superscript * denotes complex conjugation).

Using expansion (5.24), the incompressibility condition (5.8b) requires

$$i\alpha_m U_m(z) + W_m'(z) = 0, \qquad m \in \mathbb{Z}, \tag{5.25}$$

while the vertical BCs in (5.7) become

$$U_m(0) = W_m(0) = U_m'(1) = W_m(1) = 0, \qquad m \in \mathbb{Z}. \tag{5.26}$$

Since $\alpha_0 = 0$, (5.25) and (5.26) imply that $W_0(z) = 0$, whereas for each $m \neq 0$ one can express U_m as a function of W_m. After substituting the Fourier series expansion (5.24) into $\mathcal{Q}\{\boldsymbol{u}\}$, one can combine these observations with the identities $\boldsymbol{U}_{-m} = \boldsymbol{U}_m^*$ and $\alpha_{-m} = -\alpha_m$ to show that

$$\mathcal{Q}\{\boldsymbol{u}\} = \Gamma_x \int_0^1 \left|U_0'(z)\right|^2 \,\mathrm{d}z + 2\Gamma_x \sum_{m=1}^{+\infty} \mathcal{Q}_m\{W_m\}, \tag{5.27}$$

where

$$\mathcal{Q}_m\{W_m\} := \int_0^1 \frac{1}{\alpha_m^2}\left|W_m''(z)\right|^2 + 2\left|W_m'(z)\right|^2 + \alpha_m^2\left|W_m(z)\right|^2 - \frac{2}{\alpha_m}\phi'(z)\,\mathrm{Im}\left[W_m'(z)W_m^*(z)\right]\,\mathrm{d}z. \tag{5.28}$$

Note now that among all $\boldsymbol{u} \in H$ are those defined by a single Fourier mode, meaning that $\boldsymbol{U}_m = \boldsymbol{0}$ for all but one value m. Then, it follows from (5.25), (5.26) and (5.27) that $\mathcal{Q}\{\boldsymbol{u}\}$ is non-negative on the function space H if and only if, for each positive integer m, the quadratic form $\mathcal{Q}_m\{W_m\}$ is positive semidefinite for all functions W_m that satisfy the BCs

$$W_m(0) = W_m(1) = W_m'(0) = W_m''(1) = 0. \tag{5.29}$$

Such functions will be called *admissible*, and each of the conditions that $\mathcal{Q}_m\{W_m\} \geq 0$ for all admissible W_m will be referred to as a *Fourier-transformed spectral constraint*.

For any candidate background field, only a finite number of Fourier-transformed spectral constraints need be considered because $\mathcal{Q}_m\{W_m\}$ is always non-negative when m is sufficiently large. Indeed, since the Legendre polynomials satisfy $\|L_n\|_\infty = 1$ (cf. section 4.1) one has

$$\left\|\phi'\right\|_\infty \leq \sum_{n=0}^{P-1}\left|\hat{\phi}_n\right| \|L_n\|_\infty = \|\hat{\boldsymbol{\phi}}\|_1. \tag{5.30}$$

Combining this estimate with the Cauchy-Schwarz inequality and the elementary inequality $ab \leq a^2/(\sqrt{2}\alpha_m) + \alpha_m b^2/(2\sqrt{2})$ yields

$$\begin{aligned}
\mathcal{Q}_m\{W_m\} &\geq 2\left\|W_m'\right\|_2^2 + \alpha_m^2\left\|W_m\right\|_2^2 - \frac{2}{\alpha_m}\left\|\phi'\right\|_\infty \left\|W_m'\right\|_2 \left\|W_m\right\|_2, \\
&\geq 2\left\|W_m'\right\|_2^2 + \alpha_m^2\left\|W_m\right\|_2^2 - \frac{2}{\alpha_m}\|\hat{\boldsymbol{\phi}}\|_1 \left\|W_m'\right\|_2 \left\|W_m\right\|_2, \\
&\geq \left(1 - \frac{\|\hat{\boldsymbol{\phi}}\|_1}{\sqrt{2}\alpha_m^2}\right)\left(2\left\|W_m'\right\|_2^2 + \alpha_m^2\left\|W_m\right\|_2^2\right),
\end{aligned} \tag{5.31}$$

so $\mathcal{Q}_m\{W_m\}$ is non-negative if $\alpha_m^2 \geq \|\hat{\boldsymbol{\phi}}\|_1/\sqrt{2}$. Consequently, for a given background field ϕ, the Fourier-transformed spectral constraints are guaranteed to hold for m larger than the

critical value

$$m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}}) := \left\lfloor \frac{\Gamma_x}{\pi} \sqrt{\frac{\|\hat{\boldsymbol{\phi}}\|_1}{4\sqrt{2}}} \right\rfloor, \tag{5.32}$$

where $\lfloor \cdot \rfloor$ denotes the integer part of a number as usual. Thus, problem (5.23) for the optimal degree-P polynomial background field is equivalent to

$$\begin{aligned} \min_{\hat{\boldsymbol{\phi}},\, s} \quad & s - 2\,\hat{\phi}_0 \\ \text{s.t.} \quad & \mathcal{Q}_m\{W_m\} \geq 0 \quad \forall \text{ admissible } W_m, \quad m = 1,\, 2,\, \ldots,\, m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}}), \\ & \boldsymbol{S}(\hat{\boldsymbol{\phi}}, s) \succeq 0, \\ & \mathbf{1}^{\mathsf{T}} \hat{\boldsymbol{\phi}} = Gr. \end{aligned} \tag{5.33}$$

At this stage, note that each Fourier-transformed spectral constraint is an affine homogeneous integral inequality of the type studied in chapter 4, so the feasible set of (5.33) can be approximated using LMIs to obtain an SDP. The inner approximations described in section 4.4 are particularly useful because, according to Theorem 4.9, they enable one to find upper bounds on the optimal value of (5.33). This, in turn, is no smaller than the optimal value of the variational problem (5.15) for the fully optimal (*i.e.*, not necessarily polynomial) background field. Consequently, modulo numerical roundoff errors, by solving inner SDP approximations of (5.33) one can estimate rigorously from above the best bound on C_ε available within the background method analysis of section 5.2. The outer approximations described in section 4.3, instead, are not as useful because they yield lower bounds for the optimal value of (5.33), which cannot be related to the optimal value of (5.15). For this reason, outer SDP approximations of (5.33) will not be considered.

To formulate inner approximations of (5.33), for each Fourier-transformed spectral constraint one rescales the integration domain in (5.28) to $[-1, 1]$ by changing variables to $\xi = 2z - 1$. Then, following the Legendre expansion strategy outlined in sections 4.4.1 and 4.4.2, one chooses an integer $N \geq P$, so (4.27) is satisfied, and writes

$$W_m = \sum_{n=0}^{N} a_n L_n(\xi) + A(\xi), \qquad A(z) := \sum_{n \geq N+1} a_n L_n(\xi), \tag{5.34a}$$

$$W'_m = \sum_{n=0}^{N+1} b_n L_n(\xi) + B(\xi), \qquad B(z) := \sum_{n \geq N+2} b_n L_n(\xi), \tag{5.34b}$$

$$W''_m = \sum_{n=0}^{N+P+3} c_n L_n(\xi) + C(\xi), \qquad C(z) := \sum_{n \geq N+P+4} c_n L_n(\xi). \tag{5.34c}$$

The limits for the finite sums in these expression are chosen such that the complex-valued coefficients $\{a_n\}_{n=0}^{N}$ and $\{b_n\}_{n=0}^{N+1}$ can be expressed in terms of $\{c_n\}_{n=0}^{N+P+3}$ using the BCs in (5.29) and Lemma 4.3,[3] and such that part (i) of Lemma 4.6 can be applied when expanding the last term in (5.28). Moreover, the BC $W_m'(0) = 0$ can be used to rewrite c_0 in terms of the remaining coefficients $\{c_n\}_{n=1}^{N+P+3}$. Hence, using (5.34a)–(5.34c) the quadratic form $\mathcal{Q}_m\{W_m\}$ in (5.28) can be expanded as a quadratic form of the vector

$$\boldsymbol{\omega} := \Big[\mathrm{Re}(c_1),\, \ldots,\, \mathrm{Re}(c_{N+P+3}),\, \mathrm{Im}(c_1),\, \ldots,\, \mathrm{Im}(c_{N+P+3})\Big]^{\mathsf{T}} \in \mathbb{R}^{2(N+P+3)}, \tag{5.35}$$

plus some "tail" terms that depend on the remainder functions $A(\xi)$, $B(\xi)$, and $C(\xi)$.

To make these ideas more precise, substitute expansions (5.34a)–(5.34c) into (5.28) after changing variables to $\xi = 2z - 1$. The orthogonality relation (4.9) for the Legendre polynomials enables one to write $\mathcal{Q}_m\{W_m\} = (\mathcal{P} - \mathcal{Q} + \mathcal{R})/2$, with

$$\mathcal{P} := \frac{16}{\alpha_m^2} \sum_{n=1}^{N+P+3} \frac{2\,|c_n|^2}{2n+1} + 8 \sum_{n=0}^{N+1} \frac{2\,|b_n|^2}{2n+1} + \alpha_m^2 \sum_{n=0}^{N} \frac{2\,|a_n|^2}{2n+1}, \tag{5.36a}$$

$$\mathcal{Q} := \frac{8}{\alpha_m} \int_{-1}^{1} \frac{\mathrm{d}\phi}{\mathrm{d}\xi}\, \mathrm{Im}\left[A^*(\xi) \sum_{n=0}^{N+1} b_n L_n(\xi) + B(\xi) \sum_{n=0}^{N} a_n^* L_n(\xi)\right] \mathrm{d}\xi, \tag{5.36b}$$

$$\mathcal{R} := \int_{-1}^{1} \frac{16}{\alpha_m^2}\, |C(\xi)|^2 + 8\,|B(\xi)|^2 + \alpha_m^2\,|A(\xi)|^2 - \frac{8}{\alpha_m}\, \frac{\mathrm{d}\phi}{\mathrm{d}\xi}\, \mathrm{Im}\left[B(\xi) A^*(\xi)\right] \mathrm{d}\xi. \tag{5.36c}$$

Using the BCs in (5.29) and Lemma 4.3 to write the coefficients $\{a_n\}_{n=0}^{N}$ and $\{b_n\}_{n=0}^{N+1}$ in terms of $\{c_n\}_{n=1}^{N+P+3}$, it is clear that $\mathcal{P}$ is a quadratic form of the vector $\boldsymbol{\omega}$ defined in (5.35). Similarly, the proof of part (i) of Lemma 4.6 can be adapted to show that, by virtue of the choice of limits for the finite sums in (5.34a)–(5.34c), the term $\mathcal{Q}$ is also a finite-dimensional quadratic form for $\boldsymbol{\omega}$, which depends affinely on the Legendre coefficients of $\frac{\mathrm{d}\phi}{\mathrm{d}\xi}$. Therefore, one can construct a real-valued symmetric matrix $\boldsymbol{Q}_m(\hat{\boldsymbol{\phi}}) \in \mathbb{S}^{2(N+P+3)}$, affinely dependent on $\hat{\boldsymbol{\phi}}$, such that

$$\mathcal{P} - \mathcal{Q} = \boldsymbol{\omega}^{\mathsf{T}} \boldsymbol{Q}_m(\hat{\boldsymbol{\phi}})\, \boldsymbol{\omega}. \tag{5.37}$$

Finally, after dropping the second and third terms from $\mathcal{R}$ and estimating the last term using Lemma 4.7—which amounts to an application of the L^1-L^∞ Hölder inequality, the Cauchy–Schwarz inequality, and Poincaré-type estimates to relate the norms $\|A\|_2$ and $\|B\|_2$ to $\|C\|_2$—one can construct a positive definite matrix $\boldsymbol{R} \in \mathbb{S}^{2(N+P+3)}$ and a positive constant

[3] All lemmas in chapter 4 are proven for real-valued Legendre expansion, but hold also in the complex-valued case because they can be applied independently to the real and imaginary parts.

κ, both independent of m, such that

$$\mathcal{R} \geq \int_{-1}^{1} \frac{16}{\alpha_m^2} \left(1 - \alpha_m \|\hat{\boldsymbol{\phi}}\|_1 \kappa\right) |C(\xi)|^2 \, \mathrm{d}\xi - \frac{1}{\alpha_m} \|\hat{\boldsymbol{\phi}}\|_1 \boldsymbol{\omega}^\mathsf{T} \boldsymbol{R} \boldsymbol{\omega}. \tag{5.38}$$

The quadratic form $\mathcal{Q}_m\{W_m\} = (\mathcal{P} - \mathcal{Q} + \mathcal{R})/2$ can then be bounded from below as

$$\mathcal{Q}_m\{W_m\} \geq \frac{1}{2} \boldsymbol{\omega}^\mathsf{T} \left[\boldsymbol{Q}_m(\hat{\boldsymbol{\phi}}) - \alpha_m^{-1} \|\hat{\boldsymbol{\phi}}\|_1 \boldsymbol{R}\right] \boldsymbol{\omega} + \frac{1}{2} \int_{-1}^{1} \frac{16}{\alpha_m^2} \left(1 - \alpha_m \|\hat{\boldsymbol{\phi}}\|_1 \kappa\right) |C(\xi)|^2 \, \mathrm{d}\xi, \tag{5.39}$$

and the m-th Fourier-transformed spectral constraint is satisfied if

$$\boldsymbol{Q}_m(\hat{\boldsymbol{\phi}}) - \alpha_m^{-1} \|\hat{\boldsymbol{\phi}}\|_1 \boldsymbol{R} \succeq 0, \tag{5.40a}$$

$$1 - \alpha_m \|\hat{\boldsymbol{\phi}}\|_1 \kappa \geq 0. \tag{5.40b}$$

Conditions (5.40a) and (5.40b) can be recast as LMIs upon introducing a vector $\boldsymbol{t} \in \mathbb{R}^P$ of slack variables such that $-\boldsymbol{t} \leq \hat{\boldsymbol{\phi}} \leq \boldsymbol{t}$, so $\|\hat{\boldsymbol{\phi}}\|_1 \leq \mathbf{1}^\mathsf{T} \boldsymbol{t}$. Therefore, upper bounds on the optimal value of (5.23) can be computed by solving the SDP

$$\begin{aligned}
\min_{\hat{\boldsymbol{\phi}},\, s,\, \boldsymbol{t}} \quad & s - 2\,\hat{\phi}_0 \\
\text{s.t.} \quad & \boldsymbol{Q}_m(\hat{\boldsymbol{\phi}}) - \alpha_m^{-1} (\mathbf{1}^\mathsf{T} \boldsymbol{t}) \boldsymbol{R} \succeq 0, \qquad m = 1, 2, \ldots, m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}}), \\
& 1 - \alpha_m (\mathbf{1}^\mathsf{T} \boldsymbol{t}) \kappa \geq 0, \qquad m = 1, 2, \ldots, m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}}), \\
& \boldsymbol{t} - \hat{\boldsymbol{\phi}} \geq \mathbf{0}, \\
& \boldsymbol{t} + \hat{\boldsymbol{\phi}} \geq \mathbf{0}, \\
& \boldsymbol{S}(\hat{\boldsymbol{\phi}}, s) \succeq 0, \\
& \mathbf{1}^\mathsf{T} \hat{\boldsymbol{\phi}} = Gr.
\end{aligned} \tag{5.41}$$

The only difficulty preventing a direct implementation of (5.41) is that the number of constraints is not known *a priori* because it depends on the optimisation variable. This issue, however, is easily resolved in practice by employing an iterative procedure: fix an integer m_0, compute a candidate optimal $\hat{\boldsymbol{\phi}}$ by solving (5.41) considering only constraints with $m \leq m_0$, check *a posteriori* that either $m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}}) \leq m_0$ or $\mathcal{Q}_m\{W_m\} \geq 0$ for all admissible W_m and all $m \leq m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}})$, and repeat the process with larger m_0 if any of these checks fail.

Remark 5.3. The role of $m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}})$ is that of an upper bound on the largest critical Fourier mode, meaning the largest m for which the constraints in the SDP (5.41) are active. Of course, if the critical modes were known *a priori*, one could solve the SDP by considering

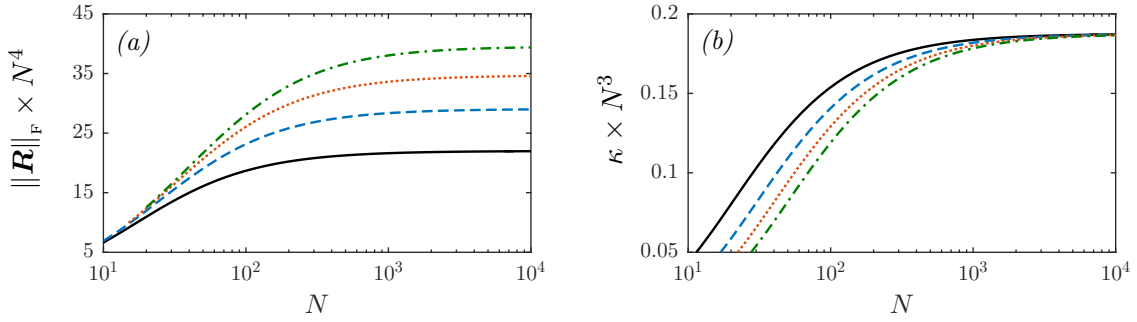


FIGURE 5.2: Compensated plots of *(a)* $\|\boldsymbol{R}\|_{\mathrm{F}}$ as function of N, and *(b)* κ as function of N. Results are plotted for $P = 5$ (——), $P = 10$ (- - -), $P = 15$ (·····), and $P = 20$ (-·-·-).

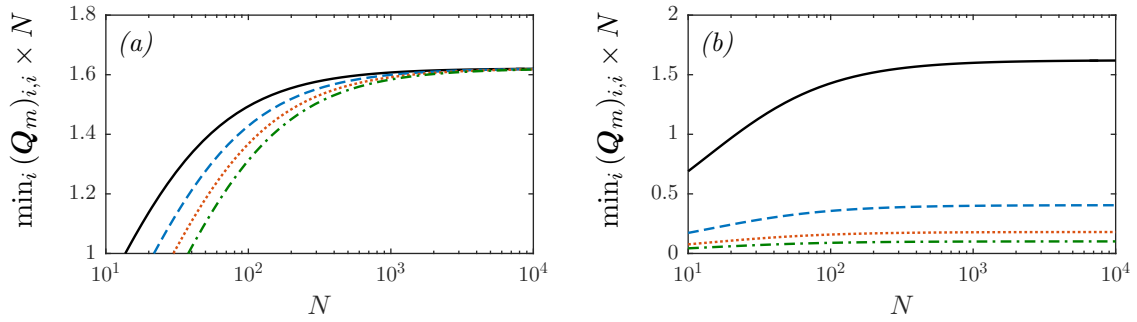


FIGURE 5.3: Compensated plots of the minimum diagonal element of the matrix $\boldsymbol{Q}_m(\hat{\boldsymbol{\phi}})$ as function of N. *(a)* Results for $m = 1$, $\Gamma_x = 2$, and $P = 5$ (——), $P = 10$ (- - -), $P = 15$ (·····), and $P = 20$ (-·-·-). *(b)* Results for $P = 10$, $\Gamma_x = 2$, and $m = 1$ (——), $m = 2$ (- - -), $m = 3$ (·····), and $m = 4$ (-·-·-).

only the corresponding constraints and still obtain the correct solution. It must be stressed, however, that the number of constraints in (5.41) is unknown not due to the lack of knowledge of the exact critical modes, but because $m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}})$ depends on $\hat{\boldsymbol{\phi}}$. If the largest critical m could be bounded independently of the decision variables, say as a function of the Grashoff number alone, then the number of constraints in the SDP would be well defined and the iterative procedure described above would not be required.

Remark 5.4. It is not obvious that the SDP (5.41) is feasible because the diagonal matrix $\boldsymbol{R}$ and the constant κ appearing in its constraints are positive definite. When both P (the degree of the polynomial background field) and N (the number of Legendre modes used to expand each Fourier-transformed spectral constraint) are large, however, a strong argument in support of feasibility can be made based on two observations. First, both $\boldsymbol{R}$ and κ arise from the application of Lemma 4.7 to the ϕ-dependent term in (5.36c) and, as illustrated in figure 5.2, one has $\|\boldsymbol{R}\|_{\mathrm{F}} \sim N^{-4}$, $\kappa \sim N^{-3}$. In contrast, as demonstrated in figure 5.3, the smallest diagonal entry of $\boldsymbol{Q}_m(\hat{\boldsymbol{\phi}})$ decays as N^{-1} (it can be verified that diagonal entries are independent of $\hat{\boldsymbol{\phi}}$). Hence, the adverse contribution of $\boldsymbol{R}$ and κ becomes vanishingly small compared to the remaining terms as N is increased. Second, it seems reasonable to

assume that there exists a strictly feasible polynomial background field ϕ_* of sufficiently high degree, for which the spectral constraint (5.13a) can be perturbed slightly without violating it. Given the fast decay of $\|\boldsymbol{R}\|_{\text{F}}$ and κ with N one expects ϕ_* to be a feasible solution of the SDP (5.41), provided that this is formulated using sufficiently large P and N. However, note also that the smallest N needed to make the condition $1-\alpha_m(\mathbf{1}^{\mathsf{T}}\boldsymbol{t})\kappa \geq 0$ feasible must grow as m, and hence the wavenumber α_m, is increased. Similarly, larger N is required to make the LMI (5.40a) feasible at large m. This has repercussion on the cost of checking that the optimal solution of (5.41) satisfies the Fourier-transformed spectral constraint for all $m \leq m_{\text{cr}}(\hat{\boldsymbol{\phi}})$, because the definition of $m_{\text{cr}}(\hat{\boldsymbol{\phi}})$ and the constraint $\mathbf{1}^{\mathsf{T}}\hat{\boldsymbol{\phi}} = Gr$ imply that

$$m_{\text{cr}}(\hat{\boldsymbol{\phi}}) = \left\lfloor \frac{\Gamma_x}{\pi}\sqrt{\frac{\|\hat{\boldsymbol{\phi}}\|_1}{4\sqrt{2}}} \right\rfloor \geq \left\lfloor \frac{\Gamma_x}{\pi}\sqrt{\frac{\mathbf{1}^{\mathsf{T}}\hat{\boldsymbol{\phi}}}{4\sqrt{2}}} \right\rfloor \sim \sqrt{Gr}. \tag{5.42}$$

Thus, when Gr is high, verifying that a given background field is feasible by testing the finite-dimensional conditions (5.40a) and (5.40b) can become computationally expensive.

Remark 5.5. Problem (5.23) for the optimal degree-P polynomial background field is very closely related to that for the energy stability of the laminar flow. Results obtained for inner SDP approximations of the latter (cf. section 4.6.1 and appendix B) suggest that not only should the SDP (5.41) be feasible, but also that its optimal solution should converge to that of (5.23) as N, the number of Legendre modes used to expand each Fourier-transformed spectral constraint, is raised.[4] As discussed in section 5.3.1, moreover, bounds on C_ε obtained with degree-P polynomial background fields converge to the fully optimal bounds as P tends to infinity. Consequently, one expects that bounds on C_ε computed via the SDP (5.41) for large P and N—which are strictly speaking suboptimal due to the restrictions and estimates made to derive the SDP—can be considered fully optimal in practice.

5.3.4 Analysis of the spectral constraint: three-dimensional flows

The spectral constraint for three-dimensional flows can be analysed with steps similar to those used in two dimensions. Following Tang *et al.* (2004), it will be assumed that the critical test functions are streamwise invariant, *i.e.*, constant along the x direction. This is reasonable because comparing the spectral constraint to the energy stability condition (5.16) for the laminar flow reveals that the former is an energy stability condition on the background field *as if it were* a solution of the governing equations, and critical modes for stability tend to be longitudinal rolls (Hagstrom & Doering, 2014). In addition, Nicodemus *et al.* (1997*a*)

[4] One could confirm this expectation by comparing the optimal value of (5.41) with that obtained using an outer approximation of (5.33), but this is not done here.

demonstrated that streamwise-varying modes do not affect the optimal background field for plane Couette flow, and one expects the same to be true for the stress-driven shear flows studied here given the similarity between the two flow configurations.

The assumption of streamwise invariance is not essential, but simplifies the analysis and, most importantly, reduces the cost of computations because it enables one to expand test functions $\boldsymbol{u} \in H$ for the spectral constraint using a Fourier series in the y direction only,

$$\boldsymbol{u}(x,y,z) = \sum_{m\in\mathbb{Z}} \boldsymbol{U}_m(z)\, \mathrm{e}^{i\beta_m y}. \tag{5.43}$$

Here, $\beta_m := 2\pi m/\Gamma_y$ is the wavenumber and the complex-valued Fourier amplitudes $\boldsymbol{U}_m(z) := U_m(z)\boldsymbol{e}_1 + V_m(z)\boldsymbol{e}_2 + W_m(z)\boldsymbol{e}_3$ satisfy $\boldsymbol{U}_{-m} = \boldsymbol{U}_m^*$, so $\boldsymbol{u}$ is real-valued. Incompressibility requires that

$$i\beta_m V_m(z) + W_m'(z) = 0, \qquad m \in \mathbb{Z}, \tag{5.44}$$

while the BCs (5.7) become

$$U_m(0) = V_m(0) = W_m(0) = U_m'(1) = V_m'(1) = W_m(1) = 0, \qquad m \in \mathbb{Z}. \tag{5.45}$$

Conditions (5.44) and (5.45) can be used to deduce that $W_0(z) = 0$, while $V_m = i\beta_m^{-1} W_m'$ for $m \neq 0$. Using the identities $\boldsymbol{U}_{-m} = \boldsymbol{U}_m^*$ and $\beta_{-m} = -\beta_m$, the quadratic form $\mathcal{Q}\{\boldsymbol{u}\}$ in (5.13a) can then be expanded as

$$\mathcal{Q}\{\boldsymbol{u}\} = \Gamma_x\,\Gamma_y \int_0^1 \left|U_0'(z)\right|^2 + \left|V_0'(z)\right|^2 \mathrm{d}z + \Gamma_x\,\Gamma_y \sum_{m\geq 1} \mathcal{Q}_m\{U_m, W_m\}, \tag{5.46}$$

where

$$\mathcal{Q}_m\{U_m, W_m\} := \int_0^1 \left|U_m'(z)\right|^2 + \beta_m^2 \left|U_m(z)\right|^2 + \frac{1}{\beta_m^2}\left|W_m(z)''\right|^2 + 2\left|W_m(z)'\right|^2 + \beta_m^2 \left|W_m(z)\right|^2 + 2\,\phi'(z)\,\mathrm{Re}[U_m(z)W_m^*(z)]\,\mathrm{d}z. \tag{5.47}$$

Noticing that among streamwise-invariant test functions $\boldsymbol{u} \in H$ are those with only one non-zero Fourier amplitude, and using (5.44) to rewrite the BCs in (5.45) in terms of U_m and W_m alone, one concludes that the spectral constraint holds if and only if for each integer m the quadratic form $\mathcal{Q}_m\{U_m, W_m\}$ is positive semidefinite for all complex-valued functions U_m, W_m that safisty the BCs

$$U_m(0) = W_m'(0) = W_m(0) = U_m'(1) = W_m''(1) = W_m(1) = 0. \tag{5.48}$$

In fact, it suffices to consider real-valued U_m and W_m because their real and imaginary parts contribute two identical and independent terms to $\mathcal{Q}_m\{U_m, W_m\}$. As in section 5.3.3, real-valued functions that satisfy (5.48) will be called *admissible*, while for each m the requirement that $\mathcal{Q}_m\{U_m, W_m\} \geq 0$ for all admissible U_m and W_m will be called a *Fourier-transformed spectral constraint*.

Combining straightforward functional estimates with (5.30) shows that, for any polynomial background field ϕ defined through (5.17),

$$\mathcal{Q}_m\{U_m, W_m\} \geq \beta_m^2 \, \|U_m\|_2^2 + 2\|\hat{\boldsymbol{\phi}}\|_1 \, \|U_m\|_2 \, \|W_m\|_2 + \beta_m^2 \, \|W_m\|_2^2 \,, \tag{5.49}$$

so the Fourier-transformed spectral constraint is guaranteed to hold when m is larger than the critical value

$$m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}}) := \left\lfloor \frac{\Gamma_y}{2\pi} \sqrt{\|\hat{\boldsymbol{\phi}}\|_1} \right\rfloor . \tag{5.50}$$

Consequently, one can replace the original spectral constraint in (5.23) with a finite number of Fourier-transformed spectral constraints, and optimise the background field by solving

$$\begin{aligned} \min_{\hat{\boldsymbol{\phi}},\, s} \quad & s - 2\,\hat{\phi}_0 \\ \text{s.t.} \quad & \mathcal{Q}_m\{U_m, W_m\} \geq 0 \quad \forall \text{ admissible } U_m,\, W_m, \quad m = 1,\, 2,\, \ldots,\, m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}}), \\ & \boldsymbol{S}(\hat{\boldsymbol{\phi}}, s) \succeq 0, \\ & \mathbf{1}^{\mathsf{T}} \hat{\boldsymbol{\phi}} = Gr. \end{aligned} \tag{5.51}$$

Each Fourier-transformed spectral constraint in this problem is an affine homogeneous integral inequality, and can be enforced using the inner approximation methods from section 4.4. For each m, the functions U_m and W_m can be expanded using Legendre series with only N terms considered explicitly. Following an argument similar to that outlined in section 5.3.3 for the two-dimensional flow, one can construct an affine matrix $\mathcal{Q}_m(\hat{\boldsymbol{\phi}}) \in \mathbb{S}^{2(N+P+3)}$, a positive definite diagonal matrix $\boldsymbol{R} \in \mathbb{S}^{2(N+P+3)}$, and two positive constants κ, ζ such that $\mathcal{Q}_m\{U_m, W_m\} \geq 0$ for all admissible test functions if

$$\boldsymbol{Q}_m(\hat{\boldsymbol{\phi}}) - \|\hat{\boldsymbol{\phi}}\|_1 \boldsymbol{R} \succeq 0, \tag{5.52a}$$

$$1 - \|\hat{\boldsymbol{\phi}}\|_1 \kappa \geq 0, \tag{5.52b}$$

$$1 - \beta_m^2 \, \|\hat{\boldsymbol{\phi}}\|_1 \zeta \geq 0. \tag{5.52c}$$

Note that $\boldsymbol{R}$, κ, and ζ are independent of m and have properties similar to those of the corresponding quantities for the two-dimensional flow discussed in remark 5.4. Conditions

(5.52a)–(5.52c) can be recast as LMIs by introducing a slack variable vector $\boldsymbol{t} \in \mathbb{R}^P$ subject to $-\boldsymbol{t} \leq \hat{\boldsymbol{\phi}} \leq \boldsymbol{t}$ and replacing $\|\hat{\boldsymbol{\phi}}\|_1$ with $\mathbf{1}^\mathsf{T}\boldsymbol{t}$. One can therefore optimise the background field ϕ by solving the SDP

$$\begin{aligned}
\min_{\hat{\boldsymbol{\phi}},\, s,\, \boldsymbol{t}} \quad & s - 2\,\hat{\phi}_0 \\
\text{s.t.} \quad & \boldsymbol{Q}_m(\hat{\boldsymbol{\phi}}) - (\mathbf{1}^\mathsf{T}\boldsymbol{t})\boldsymbol{R} \succeq 0, \qquad m = 1,\, 2,\, \ldots,\, m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}}), \\
& 1 - \beta_m^2\,(\mathbf{1}^\mathsf{T}\boldsymbol{t})\zeta \geq 0, \qquad m = 1,\, 2,\, \ldots,\, m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}}), \\
& 1 - (\mathbf{1}^\mathsf{T}\boldsymbol{t})\kappa \geq 0, \\
& \boldsymbol{t} - \hat{\boldsymbol{\phi}} \geq \mathbf{0}, \\
& \boldsymbol{t} + \hat{\boldsymbol{\phi}} \geq \mathbf{0}, \\
& \boldsymbol{S}(\hat{\boldsymbol{\phi}}, s) \succeq 0, \\
& \mathbf{1}^\mathsf{T}\hat{\boldsymbol{\phi}} = Gr,
\end{aligned} \tag{5.53}$$

using the same iterative procedure described at the end of section 5.3.3. The comments on feasibility and convergence made in remarks 5.4 and 5.5 apply *mutatis mutandis*.

5.4 Results

Optimal background fields and the corresponding bounds on the dissipation coefficient C_ε for both two- and three-dimensional flows were computed for a selection of Grashoff numbers by solving the SDPs (5.41) and (5.53). These were set up in MATLAB using QUINOPT and solved with SDPT3 (Toh *et al.*, 1999; Tütüncü *et al.*, 2003) on a PC with a 3.40 GHz Intel® Core™ i7-4770 CPU and 16 GB of RAM.

At each Gr, near-optimal bounds on C_ε were computed by solving the SDPs with a fixed initial guess m_0 for the number of constraints, and increasing N (the number of Legendre modes used to expand each Fourier-transformed spectral constraint) and P (the degree of the background field) in an alternate fashion until the optimal value decreased by less than 1%. If $m_0 \leq m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}})$ for the candidate solution vector $\hat{\boldsymbol{\phi}}$, the feasibility of all Fourier-transformed spectral constraints with $m \leq m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}})$ was subsequently tested via conditions (5.40a) and (5.40b) for the two-dimensional problem, and via conditions (5.52a)–(5.52c) for the three-dimensional one. For reasons discussed in remark 5.4, if feasibility could not be established for some value $m > m_0$, the verification steps were repeated up to five times after re-formulating the finite-dimensional conditions using larger N, denoted N_{checks} in the following. Every time, N_{checks} was increased by 50. If after five attempts the candidate solution could still not be verified, the optimisation was repeated after increasing m_0 by 5.

TABLE 5.1: Parameters used to set up and solve SDP (5.41) for a selection of Grashoff numbers. Also reported are the wall time (in seconds) and peak memory (in MB) required to set up the SDP, solve it with SDPT3, and post-process the solution. Tabulated values of $m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}})$ are for the optimal background field, obtained using the tabulated values of m_0 after one iteration of the procedure needed to determine the number of constraints in (5.41).

Γ_x	Gr	m_0	P	N	Memory (MB)	Time (s)	$m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}})$	N_{checks}
2	10^3	5	20	100	789	26	11	100
2	10^4	10	30	125	813	128	37	125
2	10^5	15	35	150	1192	562	115	350
3	10^3	5	20	100	854	26	17	100
3	10^4	10	30	125	941	192	56	125
3	10^5	25	35	150	1634	1098	173	350

TABLE 5.2: Parameters used to set up and solve SDP (5.53) for a selection of Grashoff numbers. Also reported are the wall time (in seconds) and peak memory (in MB) required to set up the SDP, solve it with SDPT3, and post-process the solution. Tabulated values of $m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}})$ are for the optimal background field, obtained using the tabulated values of m_0 after one iteration of the procedure needed to determine the number of constraints in (5.53).

Γ_y	Gr	m_0	P	N	Memory (MB)	Time (s)	$m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}})$	N_{checks}
2	10^2	5	35	100	704	48	4	100
2	10^3	10	40	125	983	270	13	125
2	10^4	10	45	150	1148	537	43	150
3	10^2	5	35	100	752	43	6	100
3	10^3	10	40	125	944	252	20	125
3	10^4	10	45	150	1063	500	64	150

Details of the parameters used in the computations for a selection of Grashoff numbers are given in table 5.1 for the two-dimensional flow, and in table 5.2 for the three-dimensional flow. For the tabulated values m_0, identification of the correct number of constraints in the SDPs required only one iteration. Also reported are the wall time and peak memory (as measured by the operating system) required to set up the SDPs, solve them using SDPT3, and post-process the solution. The reason for the disparity between N and P is that large N is needed to ensure feasibility (cf. remark 5.4), whereas the optimal ϕ is well resolved with modest P. In addition, computing near-optimal bounds for the three-dimensional flow at a given Gr required higher P than in two dimensions in order to resolve finer structures in the optimal background field.

5.4.1 Two-dimensional flows

The SDP (5.41) was successfully solved for $10 \leq Gr \leq 10^5$ and for two values of the horizontal period, $\Gamma_x = 2$ and $\Gamma_x = 3$. Figure 5.4 shows some of the optimal background

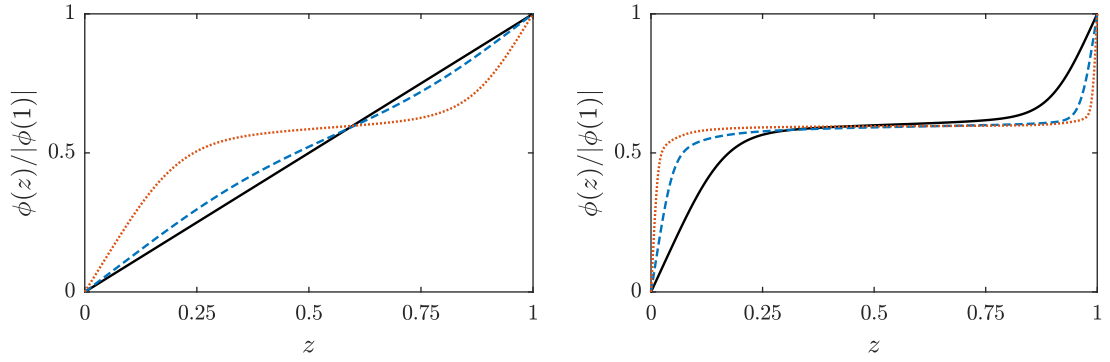


FIGURE 5.4: Numerically optimal background fields for the two-dimensional flow, computed with (5.41) for $\Gamma_x = 2$. Profiles in panel *(a)* are for $Gr = 10$ (——), $Gr = 100$ (- - -), and $Gr = 500$ (·····). Profiles in panel *(b)* are for $Gr = 10^3$ (——), $Gr = 10^4$ (- - -), and $Gr = 10^5$ (·····). All profiles are normalised by their absolute value at $z = 1$ to ease the comparison.

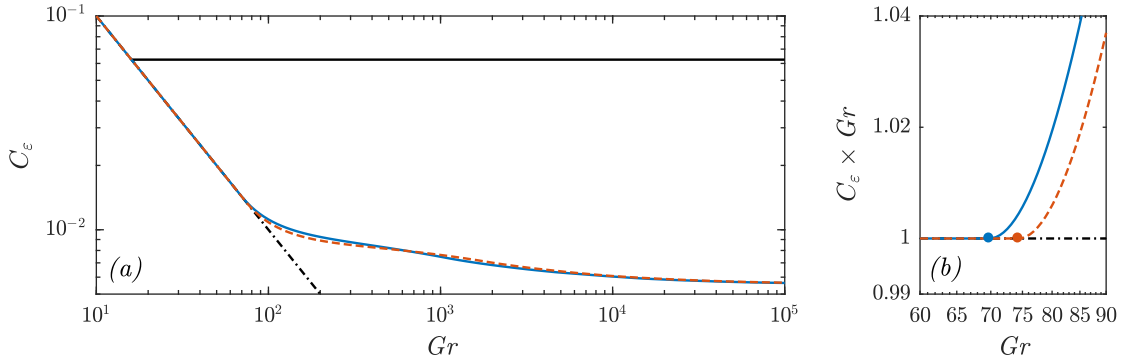


FIGURE 5.5: *(a)* Numerically optimal upper bounds on C_ε for the two-dimensional flow model, computed with (5.41) for $\Gamma_x = 2$ (——) and $\Gamma_x = 3$ (- - -). Results are compared to the analytical bound $C_\varepsilon \leq 1/16$ (——) and the laminar dissipation value, $1/Gr$ (-·-·-). *(b)* The near-optimal bounds depart from the laminar dissipation value $C_\varepsilon = 1/Gr$ at $Gr = 0.5\,Gr_E \approx 69.77$ for $\Gamma_x = 2$ (●), and at $Gr = 0.5\,Gr_E \approx 74.33$ for $\Gamma_x = 3$ (●). The vertical axis in panel *(b)* is rescaled by Gr to ease the visualisation.

fields computed for $\Gamma_x = 2$, obtained by substituting the optimal solution $\hat{\boldsymbol{\phi}}$ of (5.41) into (5.17) and integrating the resulting polynomial ϕ' subject to the BC $\phi(0) = 0$ (cf. section 5.3.1). To ease the visual comparison, the profiles are normalised by the magnitude of their boundary value $\phi(1)$. Similar profiles were obtained for $\Gamma_x = 3$ and are not shown for brevity. The optimal background fields coincide with the laminar flow profile $\phi = Gr\,z$ for $Gr \leq 0.5\,Gr_E$ (cf. remark 5.2) and, as Gr is raised, two monotonic boundary layers develop. Optimal background fields for the classical plane Couette flow behave similarly (Nicodemus *et al.*, 1997*b*), but in the shear-driven case the boundary layers are asymmetric due to the different nature of the BCs at $z = 0$ (Dirichlet) and at $z = 1$ (Neumann).

Upper bounds on C_ε were computed at each Gr by substituting the numerically optimal background field into (5.14), and are plotted in figure 5.5*(a)*. The results are compared to the laminar dissipation coefficient $1/Gr$, which is a sharp lower bound on C_ε (Tang *et al.*, 2004),

and to the analytical upper bound $C_\varepsilon \leq 1/16 = 0.0625$ proven by Hagstrom & Doering (2014). Unfortunately, the limited range of Gr for which (5.41) could be solved reliably does not permit a confident estimation of the asymptotic behaviour of the optimal bound (the computational issues at large Gr will be discussed further in section 5.5). Nonetheless, it seems that the optimal bounds on C_ε approach a constant value independently of Γ_x when $Gr \to \infty$. What is evident, on the other hand, is the quantitative improvement compared to the analytical bound, which is more than one order of magnitude larger than the numerical one at $Gr = 10^5$.

Although energy stability analysis indicates that the laminar Couette flow is stable up to the critical Grashoff number $Gr_E \approx 139.54$ for $\Gamma_x = 2$ and $Gr_E \approx 148.66$ for $\Gamma_x = 3$ (cf. table B.1 in appendix B), figure 5.5*(b)* shows that the numerical bounds deviate from the laminar value $C_\varepsilon = Gr^{-1}$ when $Gr > 0.5 Gr_E$. As discussed in remark 5.2, this is to be expected for the full bounding problem (5.15), and confirms the expectation that background fields and bounds on C_ε computed with the SDP (5.41) approximate well the fully optimal ones when the parameters P and N are sufficiently large. Consequently, while strictly speaking the bounds plotted in figure 5.5*(a)* are only upper estimate for the best possible bounds on C_ε available to the background method analysis of section 5.2, in practice they can be considered fully optimal.

Finally, for both values of Γ_x tested, the optimal bounds oscillate slightly for Grashoff numbers in the range $10^2 \lesssim Gr \lesssim 10^4$. The most likely reasons for this behaviour are that computations are carried out in a periodic layer, so the Fourier-transformed spectral constraints need be enforced only at discrete wavenumbers, and the occurrence of bifurcations in the critical Fourier modes, meaning that the constraints in (5.41) become active at more values of m. Critical Fourier modes can be clearly identified by considering the quantity

$$\begin{aligned} \lambda_0(m) := \max_{\lambda} \quad & \lambda \\ \text{s.t.} \quad & \mathcal{Q}_m\{W_m\} \geq \lambda \, \|W_m\|_2^2 \quad \forall \text{ admissible } W_m, \end{aligned} \tag{5.54}$$

which corresponds to the smallest eigenvalue of the self-adjoint linear operator associated with the quadratic form $\mathcal{Q}_m\{W_m\}$. The spectral constraint implies that $\lambda_0(m)$ must be non-negative for all m, and critical modes are characterised by $\lambda_0(m) = 0$. As an example, values of $\lambda_0(m)$ corresponding to the optimal background field at $Gr = 10^4$ were computed by solving (5.54) with QUINOPT for each $m \leq m_{\text{cr}}(\hat{\boldsymbol{\phi}})$ and are plotted in figure 5.6. At this Grashoff number, three bifurcations have occurred and there are four critical modes, $m = 1$, 2, 4, and 5.

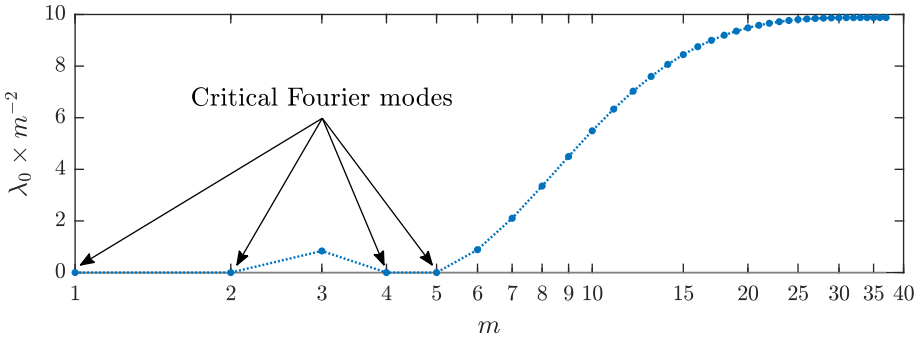


FIGURE 5.6: Values $\lambda_0(m)$ corresponding to the optimal background field for the two-dimensional flow at $Gr = 10^4$, computed by solving (5.54) with QUINOPT. Critical Fourier modes satisfy $\lambda_0(m) = 0$.

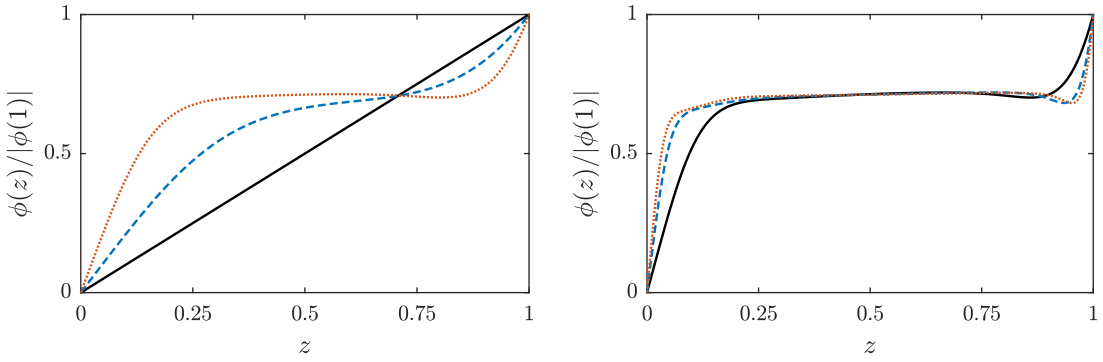


FIGURE 5.7: Numerically optimal background fields for the three-dimensional flow, computed with (5.53) for $\Gamma_y = 3$. Profiles in panel *(a)* are for $Gr = 10$ (——), $Gr = 100$ (- - -), and $Gr = 500$ (·····). Profiles in panel *(b)* are for $Gr = 1000$ (——), $Gr = 5000$ (- - -), and $Gr = 10\,000$ (·····). All profiles are normalised by the magnitude of their value at $z = 1$ to ease the comparison.

5.4.2 Three-dimensional flows

For the three-dimensional flow model, optimal background fields and bounds on C_ε were computed by solving the SDP (5.53) for $10 \leq Gr \leq 10^4$ and two values of the horizontal period in the y direction, $\Gamma_y = 2$ and $\Gamma_y = 3$. Note that the SDP is independent of the period in the x direction, Γ_x, because it was derived under the assumption that the critical modes for the spectral constraint are streamwise invariant.

A selection of optimal background fields computed with $\Gamma_y = 3$ is shown in figure 5.7, and very similar results were obtained for $\Gamma_y = 2$. Interestingly, as Gr is raised the boundary layer near the top boundary ($z = 1$) overshoots the approximately constant value in the bulk of the domain, making the optimal ϕ non-monotonic. The boundary layer near $z = 0$, instead, develops two approximately linear sub-layers characterised by different slope: steeper near the boundary, flatter towards the edge. The formation of such sub-layers seems to be related to the occurrence of bifurcations in the number of critical Fourier modes in the SDP. As in

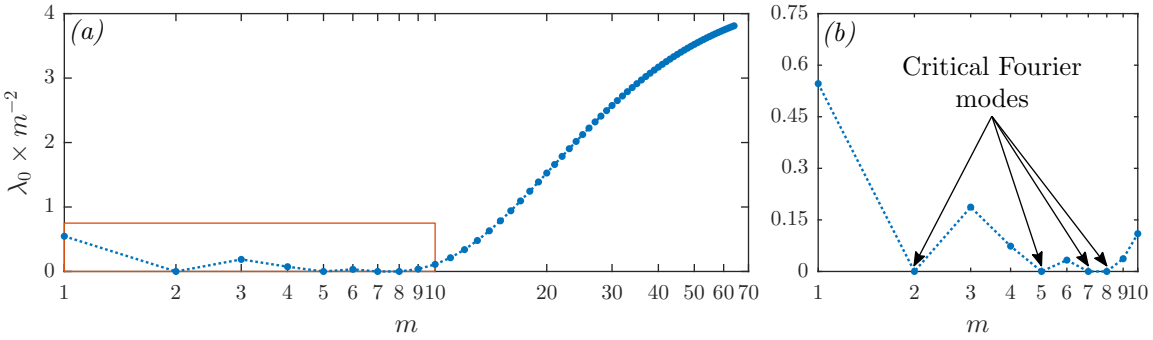


FIGURE 5.8: *(a)* Values $\lambda_0(m)$ corresponding to the optimal background field for the three-dimensional flow at $Gr = 10^4$, computed by solving (5.55) with QUINOPT. *(b)* Detailed view of the region marked by a red rectangle in panel *(a)*, showing the critical Fourier modes.

the two-dimensional case, these are characterised by $\lambda_0(m) = 0$, where

$$\begin{aligned} \lambda_0(m) := \max_{\lambda} \quad & \lambda \\ \text{s.t.} \quad & \mathcal{Q}_m\{U_m, W_m\} \geq \lambda \left(\|U_m\|_2^2 + \|W_m\|_2^2 \right) \quad \forall \text{ admissible } U_m, W_m. \end{aligned} \tag{5.55}$$

Figure 5.8 demonstrates that three bifurcations have occurred at $Gr = 10^4$, the critical Fourier modes being $m = 2$, 5, 7, and 8 (note that these are different from the critical modes for the two-dimensional flow at the same Grashoff number, cf. figure 5.6). Near-optimal background fields for the classical plane Couette flow exhibit similar sub-layers (Nicodemus *et al.*, 1997*b*), suggesting that the qualitative structure of the optimal background field near the bottom boundary is unaffected when the boundary condition at $z = 1$ is changed from fixed velocity to fixed shear.

The optimal bounds on C_ε are plotted in figure 5.9*(a)*, along with the laminar dissipation value $C_\varepsilon = 1/Gr$, the approximate numerical bound $C_\varepsilon \leq Gr\,(7.531\,Gr^{0.5} - 20.3)^{-2}$ found by Tang *et al.* (2004) for $Gr \gtrsim 500$, and the analytical bound $C_\varepsilon \leq 1/(2\sqrt{2}) \approx 0.3536$ proven by Hagstrom & Doering (2014). The bounds deviate from the laminar dissipation coefficient at the expected value $Gr = 0.5\,Gr_E$ (with $Gr_E \approx 57.20$ for $\Gamma_y = 2$ and $Gr_E \approx 51.73$ for $\Gamma_y = 3$, cf. table B.2 in appendix B), confirming that the optimal solution of the SDP has converged to that of the infinite-dimensional variational problem (5.15). The quantitative improvement compared to the analytical bound is evident, the optimal results being more than 10 times smaller at large Gr. In addition, although the range of Gr for which the SDP (5.53) could be solved accurately does not reach the asymptotic regime, for both values of Γ_y considered here the optimal bounds are in very good agreement with the fit to the approximate numerical bounds by Tang *et al.* (2004). These were computed by replacing the applied shear with a body force localised in a narrow region near the boundary, and

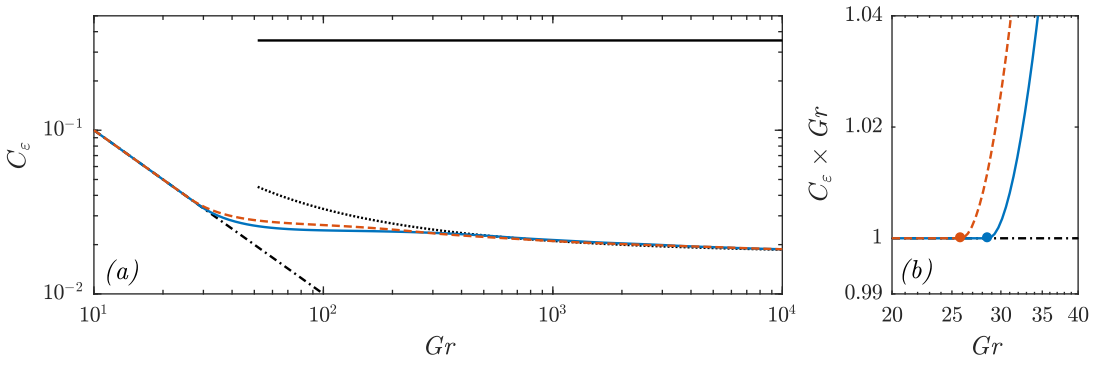


FIGURE 5.9: *(a)* Numerically optimal upper bounds on C_ε for $\Gamma_y = 2$ (——) and $\Gamma_y = 3$ (- - -). Also shown are: the approximate numerical bound $C_\varepsilon \leq Gr(7.531\,Gr^{0.5} - 20.3)^{-2}$ from Tang *et al.* (2004), valid for $Gr \gtrsim 500$ (·····); the laminar dissipation coefficient $C_\varepsilon = 1/Gr$ (-·-·-); the analytical bound $C_\varepsilon \leq 1/(2\sqrt{2})$ proven by Hagstrom & Doering (2014) (——). *(b)* The numerically optimal bounds depart from the laminar dissipation value at $Gr = 0.5\,Gr_E \approx 28.60$ for $\Gamma_x = 2$ (●), and at $Gr = 0.5\,Gr_E \approx 25.87$ for $\Gamma_x = 3$ (●). The vertical axis in panel *(b)* is rescaled by Gr to ease the visualisation.

Hagstrom & Doering (2014) demonstrated that this approximation does not change the critical Gr for energy stability of the laminar flow. The present results indicate that the energy dissipation rate is also preserved, at least as far as upper bounds on the dissipation coefficient are concerned. Given that the background method used to formulate the bounding problem for C_ε can be seen as a generalisation of energy stability theory, this observation is perhaps not surprising.

5.5 Further discussion and conclusions

The optimal bounds on C_ε appear to approach a constant as Gr grows to infinity for both two- and three-dimensional flows, although the range of Grashoff numbers that could be studied does not stretch sufficiently far into the predicted asymptotic regime to confidently estimate the value of these constants. Barring unexpected changes in the behaviour of the optimal bounds at larger Grashoff numbers and ignoring logarithmic corrections—which would be difficult to identify even if the data spanned a much larger range of Gr—this observation has two implications. First, the analytical bounds proven by Hagstrom & Doering (2014) using piecewise-linear background fields capture the asymptotic scaling of the optimal bounds available within their bounding framework. This may not come as a surprise given that the same is true for velocity-driven plane Couette flow (see, for example, Plasting & Kerswell, 2003), and that this flow is very closely related to the stress-driven shear flows studied here. Second, constant C_ε means that the dissipation becomes independent of the fluid's viscosity in the limit of infinite Gr, in accordance with Kolmogorov's theory

of turbulence (Pope, 2000, chapter 6). However, it must be stressed that real flows whose dissipation coefficient equals the bounds computed in this chapter are unlikely to exist (Tang *et al.*, 2004). Confirming this expectation, of course, requires a comparison with measurements from experiments or numerical simulations, but unfortunately no such data seem to be available in the literature.

It should be noted that the bounding problem formulated in section 5.2 can be improved if, instead of subtracting $2\times$(5.9) from the right-hand side of (5.10) to obtain (5.11), one subtracts an unspecified multiple $a\times$(5.9). Optimisation over a, known in the literature as a *balance parameter*, was first suggested by Nicodemus *et al.* (1997*a*) and can of course only improve the bounds computed here with $a = 2$. On the other hand, the variational problem for the optimal bounds on C_ε obtained after the addition of the balance parameter is not jointly convex in ϕ and a, so it cannot be solved directly using SDPs. This difficulty can in principle be overcome by changing variables to obtain a convex problem (Chernyshenko, 2017). However, given previous experience with other flows (Nicodemus *et al.*, 1998; Plasting & Kerswell, 2003; Wen *et al.*, 2013, 2015), it is expected that any improvements obtained by tuning a in the asymptotic regime will be only *quantitative*, not *qualitative*, and therefore will not offer additional insight as far as the scaling of the optimal bounds with Gr is concerned.

Another issue that deserves further discussion is the inability to optimise bounds at very large Grashoff numbers. This is a current drawback of the SDP-based methods utilised in this chapter, as one would like to be able to extract accurate scaling laws for large values of Gr. One limiting factor is that, even though the memory and wall times required to solve the largest SDP considered here are modest, computations at large Gr become less accurate and prone to bad numerical conditioning. Carefully rescaling the SDP data is expected to enable the extension of the present numerical results to larger Grashoff numbers, but unfortunately there is no general rescaling rule to improve the conditioning of SDPs. Simply rescaling the background field by various powers of Gr was not helpful, and devising a suitable rescaling strategy remains a task for future research.

A second obstacle preventing computations at large Gr is that validating the solution returned by the SDP solver can become a burden. Focussing on the two-dimensional problem for definiteness and recalling equation (5.42), this is due to the cost of performing a number of eigenvalue decompositions that grows at least as fast as $\sqrt{Gr}$ to check that the LMI (5.40a) holds for all Fourier modes up to the critical value $m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}})$. For example, at $Gr = 10^5$ and with $P = 35$, $N_{\mathrm{checks}} = 350$, the eigenvalue decomposition of a 776×776 matrix had to be computed 115 times for $\Gamma_x = 2$, and 173 times for $\Gamma_x = 3$ (cf. table 5.1). In these cases, the validation step was the most time-consuming of the entire computation, and the

situation can only be expected to worsen as Gr is raised. It may be possible to reduce wall time if LMIs are tested by attempting a Cholesky decomposition, rather than computing eigenvalues, but this was not tested. In addition, one could try to improve the bound $m_{\mathrm{cr}}(\hat{\boldsymbol{\phi}})$ on the largest critical wavenumber, which was derived here using only elementary estimates.

A final challenge comes from the fact that memory and time requirements of the algorithms implemented in general-purpose SDP solvers such as SDPT3 become prohibitive for problems with very large LMIs (Fukuda *et al.*, 2000; Wen *et al.*, 2010). This was not an issue here, as computations were constrained by poor solver convergence, but, once numerical conditioning is improved, SDPs with increasingly large LMIs will have to be solved as Gr is raised in order to approximate the variational problem for the optimal bounds accurately. This is especially true because the number of Fourier modes to be included in the SDP increases due to bifurcations and, as discussed in remark 5.4, large LMIs are needed at high wavenumbers. Therefore, limitations in the computing power available are likely to prevent calculations at Gr well within the asymptotic regime (say, $Gr \sim 10^{10}$).

Fortunately, addressing the challenges posed by large SDPs is the subject of a very active field of research, and a growing number of options are becoming available to try and combat the increase in computational cost at high Gr. One possibility is to utilise first-order algorithms to solve a large SDP either directly (Wen *et al.*, 2010; O'Donoghue *et al.*, 2016), or by considering a low-rank reformulation of its Lagrangian (Burer & Monteiro, 2003, 2005; Burer & Choi, 2006). Such algorithms have a low memory footprint and are effective if solutions of moderate accuracy are required, but convergence to high-accuracy is often slow. Another strategy is to exploit sparsity to decompose large LMIs into a set of smaller constraints, which can be handled more efficiently by existing software packages (Fukuda *et al.*, 2000; Nakata *et al.*, 2003; Kim *et al.*, 2011; Sun *et al.*, 2014; Pakazad *et al.*, 2017; see also section 2.6). This approach seems promising because, as illustrated in figures 5.10*(a,c)*, the LMIs corresponding to the Fourier-transformed spectral constraints in (5.41) and (5.53) are sparse by virtue of the orthogonality of the Legendre polynomials used in their formulation.[5] On the other hand, the sparsity patterns are not chordal, and performing so-called *chordal extensions* based on a symbolic Cholesky decomposition with approximate-minimum-degree reordering (Fukuda *et al.*, 2000) leads to consideration of the denser matrices shown in figures 5.10*(b,d)*. A decomposition of the chordal-extended LMIs was attempted with the MATLAB package SparseCoLO (Fujisawa *et al.*, 2009), but resulted

[5]It should be noticed that if polynomial expansions are used to replace each Fourier-transformed spectral constraint with an LMI, then expansions in the Legendre basis are the best choice as far as sparsity is concerned. In fact, the entries of each LMI are obtained by integrating products of polynomials in the basis, and if these are not orthogonal with respect to the usual Lebesgue inner product, one generally obtains a dense LMI.

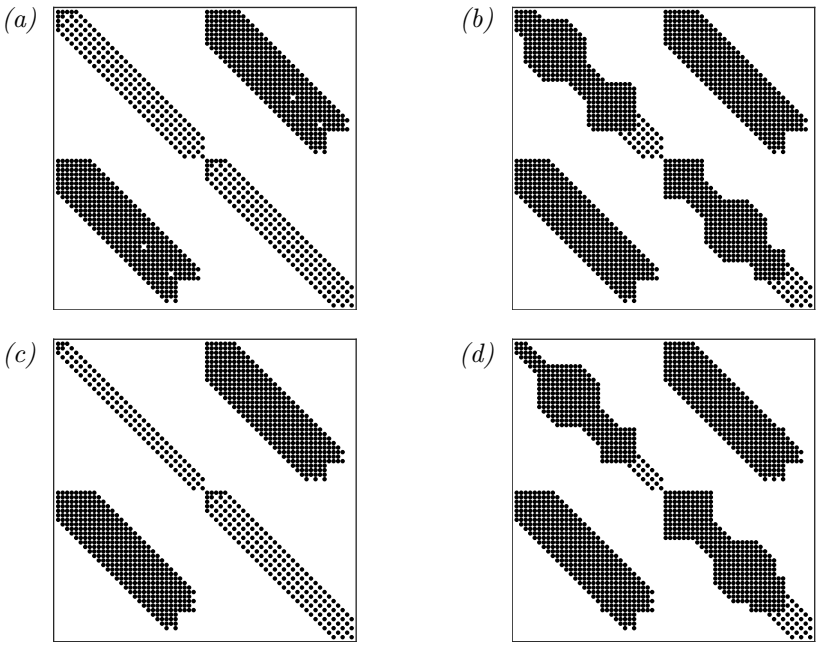


FIGURE 5.10: *(a)* Sparsity pattern of the LMIs corresponding to the Fourier-transformed spectral constraints in (5.41). *(b)* Chordal extension of the matrix sparsity pattern shown in panel *(a)*. *(c)* Sparsity pattern of the LMIs corresponding to the Fourier-transformed spectral constraints in (5.53). *(d)* Chordal extension of the matrix sparsity pattern shown in panel *(c)*. The sparsity patterns in panels *(a)* and *(c)* were obtained with QUINOPT for $P = 5$ and $N = 25$, and their chordal extensions were obtained with a symbolic Cholesky decomposition with approximate-minimum-degree reordering.

in minor performance improvements. These observations motivate the search for alternative ways to replace spectral constraints with LMIs, not based on polynomial expansions, such that chordal decomposition methods can be applied effectively. The next chapter, which focusses on optimising background fields for Bénard–Marangoni convection at infinite Prandtl number, takes a first step in this direction.

Chapter 6

Bounds on heat transfer for Bénard–Marangoni convection at infinite Prandtl number[†]

Bénard–Marangoni convection describes the motion of a layer of fluid driven by shear stresses due to gradients in surface tension at the interface between the fluid and its surroundings. This type of convection arises in a variety of industrial processes, including drying of thin polymer films (Yiantsios *et al.*, 2015), fusion welding (DebRoy & David, 1995), laser cladding (Kumar & Roy, 2009), and the growth of single-crystal semiconductors (Lappa, 2010, chapter 3 and references therein). It has also been observed in distillation columns (Zuiderweg & Harmens, 1958; Patberg *et al.*, 1983) and in differentially heated fluids in microgravity environments, where buoyancy effects are negligible (Lappa, 2010, chapter 2). In addition to these applications, Bénard–Marangoni convection has recently received increasing attention as a paradigm for shear-driven turbulent transport processes (Boeck & Thess, 1998, 2001; Hagstrom & Doering, 2010), and can be considered the analogue of the stress-driven shear flows studied in chapter 5 in the context of convective flows.

In spite of its relevance, the dynamics and heat transfer properties of Bénard–Marangoni convection have been studied far less than those of its buoyancy-driven counterpart, Rayleigh–Bénard convection. One fundamental question that remains largely unanswered is how the net vertical heat transfer across the layer, described by the Nusselt number Nu, depends on the external forcing, measured by the Marangoni number Ma. A phenomenological boundary layer scaling analysis put forward by Pumir & Blumenfeld (1996) predicts a transition from $Nu \sim Ma^{1/4}$ to $Nu \sim Ma^{1/3}$ as laminar convection rolls are replaced by turbulent convection, with prefactors that depend on the Prandtl number Pr—the ratio of the fluid's

[†]The material presented in this chapter has been published in:
Fantuzzi, G., Pershin, A. and Wynn, A. (2016). Bounds on heat transfer for Bénard–Marangoni convection at infinite Prandtl number. *Journal of Fluid Mechanics* **837**, 562-596. Available from: doi:10.1017/jfm.2017.858.

kinematic viscosity and its thermal diffusivity. Two-dimensional direct numerical simulations (DNSs) at low Pr and large Ma (Boeck & Thess, 1998; Boeck, 2005) confirm the 1/3 scaling exponent for the turbulent regime when free-slip conditions are imposed on the velocity field, but $Nu \sim Ma^{1/5}$ is observed in the no-slip case. Moreover, further DNSs by Boeck & Thess (2001) indicate that Bénard–Marangoni convection at high Prandtl number may not be turbulent even when Ma is 10^4 times the value at which convection first appears. Assuming that the observed stationary convection rolls remain stable as Ma is raised when Pr is infinite, the same authors predict that $Nu \sim Ma^{2/9}$ in this limit.

Available experimental data (see Schatz & Neitzel, 2001; Eckert & Thess, 2006, and references therein) do not reach the highly nonlinear regime, where these scaling laws are thought to apply. As discussed in chapter 1, an alternative approach to confirm or disprove the phenomenological models is to derive rigorous bounds on Nu as a function of Ma directly from the governing equations using the background method. For convection problems, a decomposition of the temperature field into the sum of a steady background temperature field τ and a time-dependent fluctuation θ can be utilised to bound Nu as a function of τ alone, provided that a certain spectral constraint holds for all admissible θ. As with all applications of the background method, the problem that results is variational in nature: optimise the bound on Nu over all background fields that satisfy the spectral constraint.

The background method has been applied extensively to the Rayleigh–Bénard problem in a variety of configurations (see, for example, Doering & Constantin, 1996; Otero *et al.*, 2002; Doering *et al.*, 2006; Wittenberg & Gao, 2010; Goluskin & Doering, 2016; more references can be found in chapter 1). On the other hand, the only result for Bénard–Marangoni convection is due to Hagstrom & Doering (2010), who used a monotonically decreasing, piecewise-linear background temperature field to prove $Nu \leq 0.841 \times Ma^{1/2}$ at finite Prandtl number, while $Nu \leq 0.838 \times Ma^{2/7}$ in the infinite-Pr limit.

This chapter investigates whether Hagstrom & Doering's bound for Bénard–Marangoni convection at infinite Prandtl number can be lowered, reducing the gap with the DNS results and phenomenological predictions by Boeck & Thess (2001). The infinite-Pr limit is an attractive model for high-Prandtl-number fluids, such as the silicone oils used in experiments (de Bruyn *et al.*, 1996) or Earth's mantle (Jones, 1977), because it gives accurate quantitative predictions whilst simplifying the governing equations (Boeck & Thess, 2001). Precisely, when $Pr = \infty$ the inertial term in the momentum equation drops out and, as a result, the velocity field can be "slaved" to the temperature field, which remains the only dynamical variable in the problem (see Hagstrom & Doering, 2010 or section 6.1 below).

The primary aim of this chapter is to determine the best possible upper bound on Nu when the background method is applied to the temperature field. To this end, Hagstrom & Doering's background method analysis will be revisited, leading to a new upper-bounding variational principle for the Nusselt number that includes two *balance parameters* (Nicodemus *et al.*, 1997*a*). One of these balance parameters can be optimised analytically, while the remaining one and the background temperature field can be combined to formulate a convex bounding problem in terms of a scaled background profile, which can be optimised via semidefinite programming. For small-to-medium Marangoni numbers, say $Ma \lesssim 10^6$, this can be done efficiently using QUINOPT and the Legendre expansion methods described in chapter 4. However, in an attempt to resolve some of the computational issues discussed in section 5.5, in this chapter SDPs will be formulated with a different approach, based on piecewise-linear (rather than polynomial) expansions. Doing so makes the computation feasible for Marangoni numbers up to $Ma = 10^9$, which is sufficiently large to conjecture that the optimal bounds take the asymptotic form $Nu \lesssim Ma^{2/7}(\ln Ma)^{-1/2}$, meaning that Hagstrom & Doering's power-law bound is only logarithmically failing.

The second aim of this chapter is to identify which features of the optimal scaled background temperature field are key to lowering the bound on Nu. For instance, non-monotonicity plays an important role in the background method analysis for Rayleigh–Bénard convection at infinite Pr (Plasting & Ierley, 2005; Doering *et al.*, 2006), and it is natural to ask if the same is true for the Bénard–Marangoni problem. Another important issue is whether one can expect to improve Hagstrom & Doering's bound using a relatively simple background field, which is amenable to rigorous mathematical analysis. Answers to these questions come from the minimisation of the bound on Nu over two restricted families of scaled background fields: those that decrease monotonically, and those constrained by convexity. Observations based on these additional numerical results are supported by analysis of the variational principle for the bound, and suggest a possible way to prove rigorously the conjectured logarithmic improvement on Hagstrom & Doering's power-law bound.

The rest of this chapter is organised in the following way. Section 6.1 describes Pearson's model (Pearson, 1958) for Bénard–Marangoni convection at infinite Prandtl number. The background method with balance parameters is used in section 6.2 to derive an upper-bounding principle for Nu, which is compared to that originally formulated by Hagstrom & Doering (2010) in section 6.3. Section 6.4 is devoted to the numerical optimisation of the background field, and the results are discussed in section 6.5 with the help of additional analysis. Challenges for computations in the asymptotic regime are commented on in section 6.6, while section 6.7 summarises the main findings and offers concluding remarks.

6.1 Pearson's model

Consider a two-dimensional layer of incompressible fluid of depth h, density ρ, kinematic viscosity ν, thermal diffusivity κ and thermal conductivity λ (all results obtained in this chapter with a two-dimensional model may be extended to three dimensions as described by Hagstrom & Doering, 2010). The fluid is heated from below at constant temperature, and cooled at the surface with a fixed heat flux q. The problem is made non-dimensional using h as the length unit, h^2/κ as the time unit, and qh/λ as the temperature unit. When the Prandtl number $Pr = \nu/\kappa$ is infinite, Pearson's equations (Pearson, 1958) reduce to (Hagstrom & Doering, 2010)

$$\nabla p = \nabla^2 \boldsymbol{u}, \tag{6.1a}$$

$$\partial_t T + \boldsymbol{u} \cdot \nabla T = \nabla^2 T, \tag{6.1b}$$

$$\nabla \cdot \boldsymbol{u} = 0, \tag{6.1c}$$

where $\boldsymbol{u}(x,z,t) = u(x,z,t)\boldsymbol{e}_1 + w(x,z,t)\boldsymbol{e}_3$ is the fluid's velocity, $p(x,z,t)$ is the pressure, and $T(x,z,t)$ is the temperature. All variables are assumed to be 2π-periodic in the horizontal direction, meaning along the x axis, and satisfy the vertical boundary conditions (BCs)

$$\boldsymbol{u}|_{z=0} = 0, \qquad w|_{z=1} = 0, \qquad T|_{z=0} = 0, \qquad \partial_z T|_{z=1} = -1. \tag{6.2}$$

The fluid is driven at the top boundary by surface tension forces due to local temperature gradients, which induce motion in the bulk of the layer through the action of viscosity. Mathematically, the situation is described by the additional BC

$$\left[\partial_z u + Ma\, \partial_x T\right]_{z=1} = 0. \tag{6.3}$$

The Marangoni number $Ma = \gamma q h^2/(\lambda\rho\nu\kappa)$, where γ is the negative of the derivative of the surface tension with respect to the fluid's temperature, describes the ratio of surface tension to viscous forces, and is the governing non-dimensional parameter of the flow.

The conductive state $\boldsymbol{u}(x,z,t) = 0$, $p =$ constant, $T(x,z,t) = -z$ is asymptotically stable when $Ma \leq 66.84$ (Fantuzzi & Wynn, 2017), while for $Ma \geq 79.61$ it is linearly unstable (Pearson, 1958) and convection sets in (Boeck & Thess, 1998, 2001). Taking the divergence of (6.1a) and using (6.1c) gives $\nabla^2 p = 0$, so taking the Laplacian of (6.1a) yields $\nabla^4 \boldsymbol{u} = 0$. Thus, each component of the ensuing convective velocity is bi-harmonic and can be determined as a linear function of the temperature, which forces the velocity through

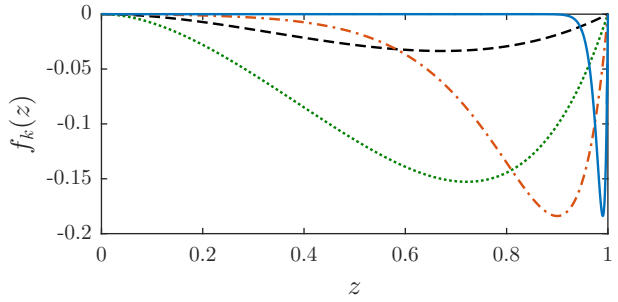


FIGURE 6.1: Plot of the function $f_k(z)$ for $k = 1$ (- - -), $k = 3$ (·····), $k = 10$ (-·-·-), and $k = 100$ (——).

the BC (6.3). In particular the horizontal Fourier coefficients $\hat{w}_k(z)$, $k \in \mathbb{Z}$, of the vertical velocity w are linear functions of the horizontal Fourier coefficients $\hat{T}_k(z)$ of the temperature. One finds (Hagstrom & Doering, 2010)

$$\hat{w}_k(z) = -Ma\, f_k(z)\, \hat{T}_k(1), \qquad k \in \mathbb{Z}, \tag{6.4}$$

where $f_0(z) = 0$ (so $\hat{w}_0 = 0$ and w has zero horizontal mean), and

$$f_k(z) = \frac{k \sinh k \left[kz \cosh(kz) - \sinh(kz) + (1 - k \coth k)\, z \sinh(kz)\right]}{\sinh(2k) - 2k}, \quad k \in \mathbb{Z} \setminus \{0\}. \tag{6.5}$$

Note that the function f_k satisfies $f_k(z) \leq 0$ for $z \in [0, 1]$, $f_k(0) = 0 = f_k(1)$, and $f_k(z) \to 0$ pointwise for all $z \in (0, 1)$ as $k \to \infty$ (see figure 6.1; note that the corresponding figure in Hagstrom & Doering's original paper is incorrect: they plot the *negative* of f_k).

Convection enhances the vertical heat transport, and since the BC $\partial_z T|_{z=1} = -1$ prescribes the heat flux through the top surface, the net effect is a reduction in the temperature drop across the layer. The key parameter to quantify this process is the Nusselt number

$$Nu := -\frac{1}{\overline{T}(1)} = \frac{1}{\langle |\nabla T|^2 \rangle}, \tag{6.6}$$

where $|\nabla T|^2 = (\partial_x T)^2 + (\partial_z T)^2$. The first equality in (6.6) defines the Nusselt number, while the second one can be proven by integrating by parts the volume and infinite-time average of $T \times$(6.1b), using (6.1c) and the BCs (see Hagstrom & Doering, 2010).

6.2 Upper-bounding principle for the Nusselt number

The analysis begins by writing $T(x, z, t) = \tau(z) + \theta(x, z, t)$, where the background field $\tau(z)$ satisfies

$$\tau(0) = 0, \qquad \tau'(1) = -1, \tag{6.7}$$

while the perturbation $\theta(x, z, t)$ is periodic in the horizontal direction and satisfies

$$\theta|_{z=0} = 0, \qquad \partial_z \theta|_{z=1} = 0. \tag{6.8}$$

Substituting this decomposition into (6.1b) yields

$$\partial_t \theta + \boldsymbol{u} \cdot \nabla \theta = \nabla^2 \theta + \tau'' - w\,\tau'. \tag{6.9}$$

Averaging $\theta\times$(6.9) over the volume and infinite time, followed by appropriate integrations by parts using (6.1c) and the BCs for θ in (6.8), shows that

$$\left\langle |\nabla \theta|^2 + \tau'\,\partial_z \theta + \tau'\,w\,\theta \right\rangle + \overline{\theta}(1) = 0. \tag{6.10}$$

Moreover, substituting the background field decomposition into (6.6) gives

$$Nu^{-1} + \overline{\theta}(1) + \tau(1) = 0, \tag{6.11a}$$

$$Nu^{-1} - \left\langle |\nabla \theta|^2 + 2\,\tau'\,\partial_z \theta \right\rangle - \left\| \tau' \right\|_2^2 = 0. \tag{6.11b}$$

After taking the linear combination $\alpha\times$(6.10)$-\beta\times$(6.11a)+(6.11b) for scalar balance parameters $\alpha, \beta \neq 1$ to be determined, using (6.8) to write $\overline{\theta}(1) = \langle \partial_z \theta \rangle$, and rearranging one arrives at

$$\frac{1}{Nu} = -\frac{\|\tau'\|_2^2 + \beta\,\tau(1)}{\beta - 1} + \frac{\alpha - 1}{\beta - 1}\,\mathcal{Q}\{\theta, w\}, \tag{6.12}$$

where

$$\mathcal{Q}\{\theta, w\} = \left\langle |\nabla \theta|^2 + \frac{\alpha}{\alpha - 1}\,\tau'\,w\,\theta + \left(\frac{\alpha - 2}{\alpha - 1}\,\tau' + \frac{\alpha - \beta}{\alpha - 1} \right) \partial_z \theta \right\rangle. \tag{6.13}$$

Then, provided that the balance parameters are chosen to satisfy

$$\frac{\alpha - 1}{\beta - 1} > 0, \tag{6.14}$$

one has

$$\frac{1}{Nu} \geq -\frac{\|\tau'\|_2^2 + \beta\,\tau(1)}{\beta - 1} + \frac{\alpha - 1}{\beta - 1} \inf_{\theta,\, w} \mathcal{Q}\{\theta, w\}, \tag{6.15}$$

where the infimum is taken over all horizontally periodic fields θ that satisfy the BCs in (6.8) and over all velocity fields w with horizontal Fourier coefficients given by (6.4). The key simplification is that θ is *not* required to satisfy the nonlinear evolution equation (6.9). As a result, it suffices to seek the infimum of $\mathcal{Q}\{\theta, w\}$ over time-independent θ (and hence w), and therefore interpret the average $\langle \cdot \rangle$ in (6.13) as a volume average only.

To compute the infimum in (6.15) one can substitute the horizontal Fourier expansions for θ and w into (6.13). Noticing that $\hat{\theta}_k = \hat{T}_k$ for $k \neq 0$, because the background field τ is independent of x, and that $f_0(z) = 0$ for all $z \in [0,1]$ in (6.4), so $\hat{w}_0 = -Ma\, f_0(z)\, \hat{T}_0(1) = 0 = -Ma\, f_0(z)\, \hat{\theta}_0(1)$, the Fourier coefficients $\hat{w}_k$ can be expressed in terms of $\hat{\theta}_k$ as

$$\hat{w}_k(z) = -Ma\, f_k(z)\, \hat{\theta}_k(1), \qquad k \in \mathbb{Z}. \tag{6.16}$$

Moreover, $\hat{\theta}_{-k} = \hat{\theta}_k^*$ (where * denotes complex conjugation) because the Fourier modes must combine into the real-valued temperature perturbation θ. Then, the quantity $\mathcal{Q}\{\theta, w\}$ can be expanded as

$$\mathcal{Q}\{\theta, w\} = \mathcal{Q}_0\{\hat{\theta}_0\} + 2 \sum_{k \geq 1} \mathcal{Q}_k\{\hat{\theta}_k\} \tag{6.17}$$

where

$$\mathcal{Q}_0\{\hat{\theta}_0\} := \int_0^1 \left|\hat{\theta}_0'(z)\right|^2 + \left(\frac{\alpha - 2}{\alpha - 1}\, \tau'(z) + \frac{\alpha - \beta}{\alpha - 1}\right) \hat{\theta}_0'(z)\, \mathrm{d}z, \tag{6.18}$$

while, since Fourier modes with $k \geq 1$ do not contribute to the last term in (6.13),

$$\mathcal{Q}_k\{\hat{\theta}_k\} := \int_0^1 \left|\hat{\theta}_k'(z)\right|^2 + k^2 \left|\hat{\theta}_k(z)\right|^2 - \frac{\alpha\, Ma}{\alpha - 1}\, \tau'(z)\, f_k(z)\, \mathrm{Re}\left[\hat{\theta}_k(1)\, \hat{\theta}_k^*(z)\right] \mathrm{d}z. \tag{6.19}$$

Now, the infimum of $\mathcal{Q}\{\theta, w\}$ must be negative semidefinite since $\mathcal{Q}\{0, 0\} = 0$, and it is finite only if each functional $\mathcal{Q}_k$, $k \geq 0$ is individually lower bounded because among all perturbations θ, w are those with only one horizontal wavenumber. Thus,

$$\inf_{\theta,\, w} \mathcal{Q}\{\theta, w\} = \inf_k \inf_{\hat{\theta}_k} \mathcal{Q}_k\{\hat{\theta}_k\}, \tag{6.20}$$

where, in light of (6.8), the infimum on the right-hand side is sought over all complex-valued functions $\hat{\theta}_k(z)$ that satisfy $\hat{\theta}_k(0) = 0 = \hat{\theta}_k'(1)$. In appendix A.7 it is shown that

$$\inf_{\hat{\theta}_0} \mathcal{Q}_0\{\hat{\theta}_0\} = -\frac{\|(\alpha - 2)\tau' + \alpha - \beta\|_2^2}{4(\alpha - 1)^2}. \tag{6.21}$$

When $k \geq 1$, instead, $\mathcal{Q}_k$ is a homogeneous functional and so if it is lower bounded, then its infimum must be exactly zero. The bound (6.15) is useful only if the infimum of $\mathcal{Q}\{\theta, w\}$ is finite, so one requires that $\mathcal{Q}_k\{\hat{\theta}_k\}$ is non-negative (which is equivalent to requiring that the infimum is zero) for all values k and obtains

$$\inf_{\theta,\, w} \mathcal{Q}\{\theta, w\} = -\frac{\|(\alpha - 2)\tau' + \alpha - \beta\|_2^2}{4(\alpha - 1)^2}. \tag{6.22}$$

Substituting this into (6.15), using the BC $\tau(0) = 0$ from (6.7) to write

$$\tau(1) = \int_0^1 \tau'(z)\,\mathrm{d}z, \tag{6.23}$$

and simplifying the resulting expression yields

$$\frac{1}{Nu} \geq -\frac{4\,\alpha(\beta-1)\tau(1) + \|\alpha\tau' + \alpha - \beta\|_2^2}{4(\alpha-1)(\beta-1)} =: \mathcal{B}\{\tau,\alpha,\beta\}. \tag{6.24}$$

This bound is valid if (6.14) holds, and if the background field τ is chosen to make the functional $\mathcal{Q}_k\{\hat{\theta}_k\}$ in (6.19) positive semidefinite for all (integer) wavenumbers $k \geq 1$. The latter set of constraints can be combined into the single condition that

$$\left\langle |\nabla\theta|^2 + \frac{\alpha}{\alpha-1}\tau'\,w\,\theta \right\rangle \geq 0 \tag{6.25}$$

for all perturbations θ, w—called *admissible* in the following—with *zero horizontal mean* and that satisfy (6.8) and (6.16). Inequality (6.25), understood over all admissible perturbations, is the expected spectral constraint on the background field τ.

Putting together all observations made so far, one concludes that the best bound on Nu available within this framework is

$$\begin{aligned}
\sup_{\tau(z),\,\alpha,\,\beta} \quad & \mathcal{B}\{\tau,\alpha,\beta\} = -\frac{4\,\alpha(\beta-1)\tau(1) + \|\alpha\tau' + \alpha - \beta\|_2^2}{4(\alpha-1)(\beta-1)} \\
\text{s.t.} \quad & \left\langle |\nabla\theta|^2 + \frac{\alpha}{\alpha-1}\tau'\,w\,\theta \right\rangle \geq 0 \quad \forall \text{ admissible } \theta,\, w, \\
& \frac{\alpha-1}{\beta-1} > 0, \\
& \tau(0) = 0, \\
& \tau'(1) = -1.
\end{aligned} \tag{6.26}$$

Note that the strict inequality $(\alpha-1)/(\beta-1) > 0$ may prevent the existence of an optimal triple (τ,α,β) that achieves the optimal value of this problem.

Remark 6.1. Problem (6.26) is interesting because the Euler–Lagrange equations that should characterise the optimal solution do not appear to be solvable. To see this, assume for simplicity that α and β are fixed and satisfy (6.14). Recall also that the spectral constraint requires $\mathcal{Q}_k\{\hat{\theta}_k\} \geq 0$ for all $\hat{\theta}_k$ satisfying $\hat{\theta}_k(0) = 0 = \hat{\theta}_k'(1)$ and all wavenumbers k. To check these conditions it suffices to consider real-valued functions $\hat{\theta}_k$, since the real and imaginary parts of $\hat{\theta}_k$ give identical and independent contributions to $\mathcal{Q}_k\{\hat{\theta}_k\}$. Then, the Lagrangian

of (6.26) can be written as

$$\mathcal{L}\{\tau, \hat{\theta}_k\} := \mathcal{B}\{\tau\} - \sum_{k\geq 1} \mathcal{Q}_k\{\hat{\theta}_k\}, \tag{6.27}$$

where the dependence on α and β has been dropped since they have been fixed for the purposes of this argument. The variation of $\mathcal{L}$ with respect to $\hat{\theta}_k$ in the direction of a function h satisfying $h(0) = 0 = h'(1)$ is

$$\frac{\delta\mathcal{L}}{\delta\hat{\theta}_k} = \int_0^1 \left[\hat{\theta}_k''(z) - k^2\hat{\theta}_k(z) + \frac{\alpha}{\alpha-1}\,\tau'(z)\,f_k(z)\,\hat{\theta}_k(1)\right] h(z)\,\mathrm{d}z \\ + \frac{\alpha}{\alpha-1}\,h(1)\int_0^1 \tau'(z)\,f_k(z)\,\hat{\theta}_k(z)\,\mathrm{d}z. \tag{6.28}$$

This variation must vanish for all h at the saddle point of the Lagrangian corresponding to the optimal solution of (6.26), so the optimal $\hat{\theta}_k$ must satisfy the differential equation

$$\hat{\theta}_k''(z) - k^2\hat{\theta}_k(z) + \frac{\alpha}{\alpha-1}\,\tau'(z)\,f_k(z)\,\hat{\theta}_k(1) = 0, \tag{6.29}$$

plus the boundary and integral conditions

$$\hat{\theta}_k(0) = 0, \qquad \hat{\theta}_k'(1) = 0, \qquad \int_0^1 \tau'(z)\,f_k(z)\,\hat{\theta}_k(z)\,\mathrm{d}z = 0. \tag{6.30}$$

This problem is over-constrained and a function $\hat{\theta}_k$ satisfying both (6.29) and (6.30) does not generally exist, except possibly for special combinations of τ and k. With all probability, therefore, the Euler–Lagrange equations for (6.26) admit no solution, and it is unclear whether numerical algorithms that rely on them, such as the time-marching algorithm of Wen *et al.* (2013, 2015), can still be employed to compute the optimal bound on *Nu*.

6.2.1 Optimisation over β

The fact that the spectral constraint is independent of β makes it possible to find the optimal β analytically. After setting to zero the first derivative of the right-hand side of (6.24) with respect to β, and using (6.23) to rearrange, one finds two stationary values,

$$\beta_+ = 1 + \left\|\alpha\,\tau' + \alpha - 1\right\|_2, \qquad \beta_- = 1 - \left\|\alpha\,\tau' + \alpha - 1\right\|_2. \tag{6.31}$$

Inspection of the second derivative of the right-hand side of (6.24) with respect to β reveals that when α is constrained by (6.14) both β_+ and β_- correspond to a local maximum. Determining the optimal β therefore requires comparing the values of such local maxima.

When $\beta = \beta_+$, let $\alpha = \lambda/(\lambda - 1)$ with $\lambda > 1$ to satisfy (6.14). Using (6.23), one can rewrite (6.24) as

$$\frac{1}{Nu} \geq \frac{1 - \|\lambda\,\tau' + 1\|_2 - \lambda\,\tau(1)}{2}. \tag{6.32}$$

The spectral constraint (6.25) can also be expressed in terms of λ as

$$\langle |\nabla\theta|^2 + \lambda\,\tau'\,w\,\theta\rangle \geq 0 \quad \forall \text{ admissible } \theta,\, w. \tag{6.33}$$

Upon introducing the scaled background field $\rho(z) := \lambda\,\tau(z) = \alpha/(\alpha - 1)\,\tau(z)$, subject to a suitably scaled version of the BCs in (6.7), the optimal bound on Nu corresponding to the choice $\beta = \beta_+$ is found by solving the variational problem

$$\begin{aligned} \sup_{\rho(z),\,\lambda} \quad & \frac{1 - \|\rho' + 1\|_2 - \rho(1)}{2} \\ \text{s.t.} \quad & \langle |\nabla\theta|^2 + \rho'\,w\,\theta\rangle \geq 0 \quad \forall \text{ admissible } \theta,\, w, \\ & \rho(0) = 0, \\ & \rho'(1) = -\lambda, \\ & \lambda > 1. \end{aligned} \tag{6.34}$$

When $\beta = \beta_-$, instead, let $\alpha = \lambda/(\lambda - 1)$ with $\lambda < 1$ to satisfy (6.14). Similar steps as above show that, when setting $\beta = \beta_-$ in (6.24), the best possible bound on Nu is given by the solution of an optimisation problem that differs from (6.34) only in the constraint for λ,

$$\begin{aligned} \sup_{\rho(z),\,\lambda} \quad & \frac{1 - \|\rho' + 1\|_2 - \rho(1)}{2}, \\ \text{s.t.} \quad & \langle |\nabla\theta|^2 + \rho'\,w\,\theta\rangle \geq 0 \quad \forall \text{ admissible } \theta,\, w, \\ & \rho(0) = 0, \\ & \rho'(1) = -\lambda, \\ & \lambda < 1. \end{aligned} \tag{6.35}$$

The key observation at this stage is that the suprema in (6.34) and (6.35) coincide despite the different constraint on λ, and furthermore they are equal to the optimal value of

$$\begin{aligned} \max_{\rho(z)} \quad & \frac{1 - \|\rho' + 1\|_2 - \rho(1)}{2}, \\ \text{s.t.} \quad & \langle |\nabla\theta|^2 + \rho'\,w\,\theta\rangle \geq 0 \quad \forall \text{ admissible } \theta,\, w, \\ & \rho(0) = 0. \end{aligned} \tag{6.36}$$

In fact, for any value of λ one can construct a feasible $\rho(z)$ for either (6.34) or (6.35) that approximates the solution of (6.36) arbitrarily accurately: simply let $\rho_0(z)$ be ε-suboptimal and strictly feasible for (6.36), and choose $\rho'(z) = \rho_0'(z)$ in (6.34) or (6.35) except for an infinitesimally thin layer near $z = 1$, where $\rho'(z) = -\lambda$. A rigorous proof is not given for brevity, but note that a similar argument was also used by Goluskin (2015) in the context of internally heated convection. The conclusion of the argument is quite satisfactory: the bound on Nu is independent of whether one sets $\beta = \beta_+$ or $\beta = \beta_-$ in (6.24).

6.2.2 An explicit value for the optimal β

The variational principle (6.36) has been obtained by optimising the balance parameter β as a function of the other balance parameter, α, and the background field $\tau(z)$. Interestingly, the optimality conditions for the solution $\rho_\star(z)$ of (6.36) allow for the derivation of a precise numerical value for the optimal β even though the optimal α and $\tau(z)$ are unknown. To show this, introduce a variable s such that $\|\rho' + 1\|_2 \leq s$ and rewrite (6.36) as

$$\begin{aligned} \max_{\rho(z),s} \quad & 1 - s - \rho(1), \\ \text{s.t.} \quad & \langle |\nabla\theta|^2 + \rho'\, w\, \theta\rangle \geq 0 \quad \forall \text{ admissible } \theta,\, w, \\ & \rho(0) = 0, \\ & \left\|\rho' + 1\right\|_2 \leq s. \end{aligned} \tag{6.37}$$

The feasible set of this problem is convex, so the linear objective function is maximised on the constraint boundary. Precisely, since for any $\rho(z)$ one can set $s = \|\rho' + 1\|_2$, the optimal bound is attained on the boundary of the feasible set of the spectral constraint, *i.e.*, when

$$\inf_{\theta, w \neq 0} \langle |\nabla\theta|^2 + \rho'\, w\, \theta\rangle = 0. \tag{6.38}$$

Given that the spectral constraint is homogeneous in θ and w, it suffices to restrict the attention to admissible θ and w satisfying some normalisation condition $\mathcal{N}\{\theta, w\} = 0$ that excludes the zero fields. The optimal scaled background field $\rho_\star(z)$ and the optimal value $s_\star$ are then those that maximise the Lagrangian functional

$$\begin{aligned} \mathcal{L}\{\rho, s, \theta, w, \zeta, \eta, \mu\} := 1 - s - \rho(1) + \zeta\, \langle |\nabla\theta|^2 + \rho'\, w\, \theta\rangle \\ + \eta\, \left(s^2 - \left\|\rho' + 1\right\|_2^2\right) + \mu\, \mathcal{N}\{\theta, w\}, \end{aligned} \tag{6.39}$$

where ζ, η and μ are scalar Lagrange multipliers.

Setting to zero the first variation of $\mathcal{L}$ with respect to $\rho(z)$ shows that the optimal scaled background field $\rho_\star(z)$ must satisfy the natural boundary condition[1]

$$1 + 2\,\eta + 2\,\eta\,\rho'_\star(1) = 0. \tag{6.40}$$

Moreover, by setting to zero the derivatives of $\mathcal{L}$ with respect to both s and η and eliminating s one obtains

$$2\,\eta\,\big\|\rho'_\star + 1\big\|_2 - 1 = 0. \tag{6.41}$$

At this point, note that if $\rho'(z) = -1$ the spectral constraint in (6.37) reduces to the "energy" stability condition of the conduction solution (Fantuzzi & Wynn, 2017). This cannot be satisfied in the convective regime, so $\|\rho'_\star + 1\|_2 \neq 0$, and using (6.41) to eliminate η from (6.40) yields

$$1 + \big\|\rho'_\star + 1\big\|_2 + \rho'_\star(1) = 0. \tag{6.42}$$

This implies that $-\rho'_\star(1) > 1$, so $\rho_\star(z)$ is also the optimal solution of (6.34) with $\lambda = -\rho'_\star(1)$. Recollecting the re-parametrisation $\alpha = \lambda/(\lambda - 1)$ one concludes that the optimal value of the balance parameter α, denoted $\alpha_\star$, is given by

$$\alpha_\star = \frac{\rho'_\star(1)}{\rho'_\star(1) + 1}. \tag{6.43}$$

Finally, since (6.34) was obtained by choosing $\beta = \beta_+$ in (6.31), and since $\alpha_\star \tau'_\star(z)/(\alpha_\star - 1) = \rho'_\star(z)$ by definition of the scaled background field, equations (6.43) and (6.42) reveal that the optimal value of the balance parameter β is

$$\beta_\star = 1 + (\alpha_\star - 1)\,\big\|\rho'_\star + 1\big\|_2 = \frac{\rho'_\star(1) + 1 - \|\rho'_\star + 1\|_2}{\rho'_\star(1) + 1} = 2. \tag{6.44}$$

6.3 Relation to Hagstrom & Doering's variational problem

The bounding principle formulated by Hagstrom & Doering (2010) can be recovered upon setting $\alpha = 2$ and $\beta = 2$ in (6.26). These values clearly satisfy (6.14), and the choice $\beta = 2$ is optimal. The variational problem for the optimal background field becomes

$$\begin{aligned} \max_{\tau(z)} \quad & -\big\|\tau'\big\|_2^2 - 2\,\tau(1), \\ \text{s.t.} \quad & \langle |\nabla\theta|^2 + 2\,\tau'\,w\,\theta\rangle \geq 0 \quad \forall \text{ admissible } \theta,\, w, \\ & \tau(0) = 0. \end{aligned} \tag{6.45}$$

[1] Of course, $\rho_\star(z)$ must also satisfy an Euler–Lagrange differential equation, but this will not be important.

Strictly speaking one should also enforce the boundary condition $\tau'(1) = -1$, but this does not limit the choice of τ for the same reasons discussed at the end of section 6.2.1.

To bring (6.45) in contact with the variational problem (6.36) for the optimal scaled background field, let $\varphi := 2\tau$ and use the BC $\varphi(0) = 0$ to rewrite (6.45) as

$$\begin{aligned} \max_{\varphi(z)} \quad & \frac{1 - \|\varphi' + 1\|_2^2 - 2\,\varphi(1)}{4}, \\ \text{s.t.} \quad & \langle |\nabla\theta|^2 + \varphi'\, w\, \theta \rangle \geq 0 \quad \forall \text{ admissible } \theta,\, w, \\ & \varphi(0) = 0. \end{aligned} \tag{6.46}$$

It is clear that (6.36) and (6.46) have the same feasible set. It is also not difficult to see that the optimal value of (6.36) is at least as large as that of (6.46), because

$$\frac{1 - \|\varphi' + 1\|_2 - \varphi(1)}{2} - \frac{1 - \|\varphi' + 1\|_2^2 - 2\,\varphi(1)}{4} = \left(\frac{1 - \|\varphi' + 1\|_2}{2}\right)^2 \geq 0. \tag{6.47}$$

In particular, using (6.36) it is almost immediate to obtain a 4.2% improvement for the prefactor of Hagstrom & Doering's bound, $Nu \leq 0.838\, Ma^{2/7}$, at least in the limit of infinite Marangoni number. Precisely, it is proven in appendix A.8 that

$$Nu \leq 0.803 \times Ma^{2/7} \quad \text{as } Ma \to \infty. \tag{6.48}$$

On the other hand, while the choice $\beta = 2$ is optimal (cf. section 6.2.2), it is not clear whether optimising α implicitly by solving (6.36) improves the asymptotic behaviour of the optimal bound on Nu compared to fixing $\alpha = 2$ *a priori*. What can be shown is that if the optimal bound obtained without optimising α has the asymptotic form $Nu \lesssim Ma^{\gamma_1} (\ln Ma)^{\gamma_2}$ for some constants $\gamma_1 > 0$ and $\gamma_2 \in \mathbb{R}$, then the bound obtained when α is optimised scales in the same way. To see this, fix $\beta = 2$ and re-parametrise $\alpha = \lambda/(\lambda - 1)$, with $\lambda > 1$ to satisfy (6.14). Optimisation over α is the same as optimisation over λ, and the bound on Nu in (6.24) can be written in terms of λ as

$$\frac{1}{Nu} \geq 1 - \frac{\lambda^2}{4(\lambda - 1)} \left\|\tau' + 1\right\|_2^2. \tag{6.49}$$

From this point onwards, the analysis is similar to that of the infinite-Pr Rayleigh–Bénard problem (Plasting, 2004, chapter 6). First, let $w = Ma\,\tilde{w}$ and define the scaled Marangoni number $M = \lambda Ma$ to rewrite the spectral constraint (6.33) as

$$\left\langle |\nabla\theta|^2 + M\,\tau'\,\tilde{w}\,\theta \right\rangle \geq 0 \quad \forall \text{ admissible } \theta,\, \tilde{w}. \tag{6.50}$$

Upon rescaling $w = Ma\,\tilde{w}$ the Marangoni number drops out of equation (6.16), so $\tilde{w}$ is a (linear) function of θ only and the admissible test functions in (6.50) are independent of M. Then, consider the family of background fields τ_M, parametrised by the scaled Marangoni number M, that maximises the right-hand side of (6.49) for a fixed value $\lambda > 1$. In other words, assume that τ_M solves the variational problem

$$\begin{aligned} \min_{\tau(z)} \quad & \big\|\tau' + 1\big\|_2^2 \\ \text{s.t.} \quad & \big\langle |\nabla\theta|^2 + M\,\tau'\,\tilde{w}\,\theta \big\rangle \geq 0 \quad \forall \text{ admissible } \theta,\, \tilde{w}. \end{aligned} \tag{6.51}$$

Moreover, write

$$\big\|\tau_M' + 1\big\|_2^2 = 1 - \sigma(M) \tag{6.52}$$

for an appropriate function $\sigma(M)$. Using (6.52) and recalling that $M = \lambda Ma$, the optimal bound on Nu becomes

$$\frac{1}{Nu} \geq \frac{\lambda^2\,\sigma(\lambda Ma) - (\lambda-2)^2}{4(\lambda-1)}. \tag{6.53}$$

It is clear that, for any fixed $\lambda > 1$, the asymptotic scaling of the function $\sigma(M)$ at large M determines the asymptotic scaling of the optimal bound on Nu with Ma, and one obtains a power-law bound with logarithmic corrections if $\sigma(M)$ has the asymptotic form

$$\sigma(M) = c\,M^{-\gamma_1}\,(\ln M)^{-\gamma_2}, \tag{6.54}$$

where $\gamma_1 > 0$, $\gamma_2 \in \mathbb{R}$ and $c > 0$ are given constants.

Assume now that the function σ obtained by solving (6.51) as a function of M has the asymptotic form (6.54). In this case, the bound on Nu in (6.53) at asymptotically large Ma can be maximised over $\lambda > 1$ by requiring that

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\frac{c\,\lambda^{2-\gamma_1}\,Ma^{-\gamma_1}\,(\ln\lambda Ma)^{-\gamma_2} - (\lambda-2)^2}{4(\lambda-1)}\right] = 0. \tag{6.55}$$

Straightforward algebra shows that the optimal $\lambda > 1$, denoted $\lambda_\star$, must satisfy

$$\lambda_\star - 1 = \frac{1 - c\,\lambda_\star^{-\gamma_1}\,Ma^{-\gamma_1}\,[\ln(\lambda_\star Ma)]^{-\gamma_2}}{1 + c(\gamma_1-1)\lambda_\star^{-\gamma_1}\,Ma^{-\gamma_1}\,[\ln(\lambda_\star Ma)]^{-\gamma_2} + \gamma_2\,c\,\lambda_\star^{-\gamma_1}\,Ma^{-\gamma_1}\,[\ln(\lambda_\star Ma)]^{-\gamma_2-1}}\,. \tag{6.56}$$

An exact formula for $\lambda_\star$ is not available, but an approximation can be found using asymptotic methods. The details can be found in appendix A.9 and the result is

$$\lambda_\star = 2 - c\,\gamma_1\,2^{-\gamma_1}\,Ma^{-\gamma_1}(\ln Ma)^{-\gamma_2} + \text{ higher-order terms}. \tag{6.57}$$

This proves that, subject to the scaling hypothesis (6.54), the optimal λ tends to a constant as $Ma \to \infty$, so optimising the balance parameters does not influence the asymptotic scaling of the bound on Nu with Ma. One also concludes that the optimal balance parameter $\alpha = \lambda/(\lambda-1)$ tends to 2 as Ma grows to infinity, meaning that the choice $\alpha = 2$ made by Hagstrom & Doering (2010)—presumably motivated only by the convenience of eliminating the linear terms when combining (6.10), (6.11a), and (6.11b) in the background method analysis—is optimal as far as the leading-order asymptotic behaviour of the bound on Nu is concerned. Provided that (6.54) holds, therefore, optimising the balance parameters not only does not improve the asymptotic scaling of the optimal bound, but also does not lower the optimal prefactor available to Hagstrom & Doering's original upper-bounding principle.

6.4 Optimal bounds

The variational problem (6.36) for the optimal scaled background field ρ can be solved numerically using SDPs. To show this, it is convenient to change variables once more and consider a function ϕ such that

$$\rho(z) := \int_0^z \phi(\xi) - 1 \,\mathrm{d}\xi, \tag{6.58}$$

so the boundary condition $\rho(0) = 0$ is satisfied. Since $\rho'(z) = \phi(z) - 1$, one can rewrite (6.36) in terms of ϕ as

$$\begin{aligned} \max_{\phi(z)} \quad & 1 - \frac{1}{2}\left\|\phi\right\|_2 - \frac{1}{2}\int_0^1 \phi(z)\,\mathrm{d}z, \\ \text{s.t.} \quad & \langle |\nabla\theta|^2 + (\phi - 1)\, w\, \theta\rangle \geq 0 \quad \forall \text{ admissible } \theta,\, w. \end{aligned} \tag{6.59}$$

Moreover, to employ semidefinite programming one needs a linear objective. Upon introducing a non-negative variable s such that $\left\|\phi\right\|_2 \leq s$ and dropping both the constant 1 and a factor of $1/2$ from the objective function, it is not difficult to see that the optimal solution of (6.59) is the same as that of the convex problem

$$\begin{aligned} \max_{\phi(z),\, s} \quad & -s - \int_0^1 \phi(z)\,\mathrm{d}z, \\ \text{s.t.} \quad & \langle |\nabla\theta|^2 + (\phi - 1)\, w\, \theta\rangle \geq 0 \quad \forall \text{ admissible } \theta,\, w, \\ & \left\|\phi\right\|_2 \leq s. \end{aligned} \tag{6.60}$$

As anticipated at the beginning of this chapter, the bound on Nu will also be optimised over the restricted classes of monotonically decreasing and convex (scaled) background fields,

i.e. such that $\rho'(z) \leq 0$ and $\rho''(z) \geq 0$. This is achieved by solving the convex problems

$$\begin{aligned} \max_{\phi(z),\, s} \quad & -s - \int_0^1 \phi(z)\,\mathrm{d}z, \\ \text{s.t.} \quad & \langle |\nabla\theta|^2 + (\phi - 1)\, w\, \theta \rangle \geq 0 \quad \forall \text{ admissible } \theta,\, w, \\ & \|\phi\|_2 \leq s, \\ & \phi(z) \leq 1, \end{aligned} \tag{6.61}$$

and

$$\begin{aligned} \max_{\phi(z),\, s} \quad & -s - \int_0^1 \phi(z)\,\mathrm{d}z, \\ \text{s.t.} \quad & \langle |\nabla\theta|^2 + (\phi - 1)\, w\, \theta \rangle \geq 0 \quad \forall \text{ admissible } \theta,\, w, \\ & \|\phi\|_2 \leq s, \\ & \phi'(z) \geq 0. \end{aligned} \tag{6.62}$$

6.4.1 Computational methodology

As in chapter 4, the derivation of SDPs that approximate (6.60)–(6.62) is based on the observation that their constraints are the infinite-dimensional equivalent of well-known types of constraints. It has already been pointed out many times in this thesis that the spectral constraint is the infinite-dimensional equivalent of an LMI. The norm constraint $\|\phi\|_2 \leq s$, instead, is the infinite-dimensional version of a second-order cone constraint (SOCC), the requirement that a vector $\boldsymbol{y} \in \mathbb{R}^{n+1}$ and a scalar s satisfy $\|\boldsymbol{y}\| \leq s$ (a more general definition of SOCCs is possible, cf. section 2.3, but is not needed here). Finally, the pointwise constraints $\phi(z) \leq 1$ and $\phi'(z) \geq 0$ are the infinite-dimensional equivalent of element-wise inequalities for a vector $\boldsymbol{y} \in \mathbb{R}^{n+1}$ of the form $\boldsymbol{A}\boldsymbol{y} \leq \boldsymbol{b}$, with $\boldsymbol{A} \in \mathbb{R}^{m\times(n+1)}$ and $\boldsymbol{b} \in \mathbb{R}^m$ given. Since SOCCs and linear inequalities are LMI-representable (cf. section 2.3), problems (6.60)–(6.62) can be solved numerically using SDPs if their constraints can be discretised into their finite-dimensional analogues.

Contrary to the approach taken in chapters 4 and 5, here the discretisation will not be carried out by considering polynomial expansions of the background field and of the test functions in the spectral constraint. Instead, given a set of $n+1$ collocation points $0 = z_0 < z_1 < \ldots < z_{n-1} < z_n = 1$ and denoting $\phi_i = \phi(z_i)$ for all $i = 1, \ldots, n$, problems (6.60)–(6.62) will be implemented using the piecewise-linear ansatz

$$\phi(z) = \sum_{i=0}^{n} \phi_i\, \psi_i(z), \tag{6.63}$$

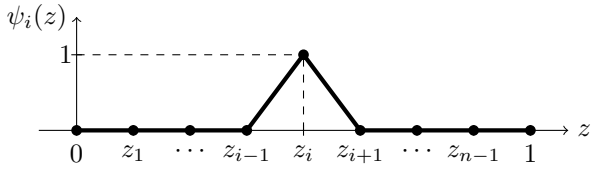


FIGURE 6.2: Sketch of the piecewise-linear function $\psi_i(z)$.

where $\psi_i(z)$ is the unique piecewise-linear function satisfying $\psi_i(z_i) = 1$ and vanishing at all other nodes (cf. figure 6.2). Similar ansatze will be introduced below to expand the spectral constraint. Clearly, the optimal solutions of problems (6.60)–(6.62) can be approximated with arbitrary accuracy by background fields of the restricted form (6.63).

This choice is made in order to address some of the computational issues outlined in 5. It will be shown that working with piecewise-linear functions leads to LMIs with chordal sparsity, for which—contrary to the LMIs encountered in chapter 5—chordal decomposition methods are effective. Using (6.63) is also advantageous because the optimal background fields turn out to be non-smooth when monotonicity or convexity are enforced, so polynomials of extreme degree are needed to approximate them accurately.[2]

When ϕ takes the form (6.63), problems (6.60)–(6.62) become optimisation problems for the vector of nodal values

$$\boldsymbol{\Phi} := [\phi_0, \ldots, \phi_n]^{\mathsf{T}} \in \mathbb{R}^{n+1}. \tag{6.64}$$

The constraint $\|\phi\|_2 \leq s$ naturally reduces to a SOCC because there exists a positive definite matrix $\boldsymbol{P} = \boldsymbol{R}^{\mathsf{T}}\boldsymbol{R}$ such that

$$\|\phi\|_2 = \left(\int_0^1 \sum_{i,j=0}^{n} \phi_i \, \phi_j \, \psi_i(z) \, \psi_j(z) \, \mathrm{d}z \right)^{1/2} = \left(\boldsymbol{\Phi}^{\mathsf{T}} \boldsymbol{P} \, \boldsymbol{\Phi} \right)^{1/2} = \|\boldsymbol{R}\,\boldsymbol{\Phi}\| \,. \tag{6.65}$$

The norm constraint $\|\phi\|_2 \leq s$ then becomes the SOCC $\|\boldsymbol{R}\,\boldsymbol{\Phi}\| \leq s$.

The spectral constraint can be reduced to a set of LMIs in a similar way, after recalling from section 6.2 that it is equivalent to the functional $\mathcal{Q}_k\{\hat{\theta}_k\}$ in (6.19) being non-negative for all integer wavenumbers $k \geq 1$ and all complex-valued functions $\hat{\theta}_k(z)$ satisfying $\hat{\theta}_k(0) = 0 = \hat{\theta}_k'(1)$. The real and imaginary parts of $\hat{\theta}_k$ give identical and independent contributions to $\mathcal{Q}_k\{\hat{\theta}_k\}$, so it suffices to consider real-valued functions $\hat{\theta}_k(z)$ in the space

$$\Gamma := \left\{ v(z) : [0,1] \to \mathbb{R}, \; \left\|v'\right\|_2^2 + \|v\|_2^2 < \infty, \; v(0) = 0, \; v'(1) = 0 \right\}. \tag{6.66}$$

[2]This was noticed after trying to solve problems (6.61) and (6.62) via the polynomial expansion methods implemented in QUINOPT.

Recalling also that variables have been changed such that

$$\frac{\alpha}{\alpha-1}\,\tau'(z) = \rho'(z) = \phi(z) - 1, \tag{6.67}$$

the spectral constraint in (6.60)–(6.62) can be replaced with the infinite set of conditions (referred to as *Fourier-transformed spectral constraints*)

$$\mathcal{Q}_k\{v\} = \int_0^1 \left|v'(z)\right|^2 + k^2\,|v(z)|^2 - Ma\,[\phi(z)-1]\,f_k(z)\,v(1)\,v(z)\,\mathrm{d}z \geq 0$$
$$\forall v \in \Gamma,\ k = 1,\ 2,\ \ldots\,. \tag{6.68}$$

Estimates detailed in appendix A.10 show that $\mathcal{Q}_k\{v\}$ is non-negative on Γ for a candidate $\phi(z)$ whenever

$$k > k_c := \left\lfloor \left(\frac{3\sqrt{3}}{128}\right)^{1/4} Ma^{1/2}\,\|\phi-1\|_\infty^{1/2} \right\rfloor, \tag{6.69}$$

where $\lfloor\cdot\rfloor$ denotes the integer part of a number. The "cutoff" wavenumber k_c represents an upper bound on the largest critical wavenumber, *i.e.* the largest values of k for which the infimum of the functional $\mathcal{Q}_k$ in (6.68) over non-zero test functions is zero.

When $k \leq k_c$, instead, the inequality $\mathcal{Q}_k\{v\} \geq 0$ can be approximated by an LMI if the test function v is represented using the piecewise-linear ansatz[3]

$$v(z) = \sum_{i=0}^{n} v_i\,\psi_i(z). \tag{6.70}$$

It should be understood that $v_0 = 0$ and $v_n = v_{n-1}$, so the boundary conditions $v(0) = 0$ and $v'(1) = 0$ are satisfied, but these substitutions are not made explicitly in (6.70) to simplify the exposition. Inserting (6.70) and (6.63) into $\mathcal{Q}_k\{v\}$ from (6.68) yields

$$\begin{aligned}\mathcal{Q}_k\{v\} = \sum_{i,j=0}^{n} v_i\,v_j \int_0^1 \psi_i(z)\,'\psi_j(z)' + k^2\,\psi_i(z)\,\psi_j(z)\,\mathrm{d}z \\ + Ma\sum_{i=0}^{n} v_n\,v_i \int_0^1 \psi_i(z)\,f_k(z)\,\mathrm{d}z \\ - Ma\sum_{i,j=0}^{n} \phi_i\,v_n\,v_j \int_0^1 \psi_i(z)\,\psi_j(z)\,f_k(z)\,\mathrm{d}z.\end{aligned} \tag{6.71}$$

Since $v_0 = 0$ and $v_n = v_{n-1}$, the right-hand side of (6.71) is a quadratic function of the vector $\boldsymbol{v} := [v_1,\ \ldots,\ v_{n-1}]^\mathsf{T}$, and there exists a symmetric matrix $\boldsymbol{Q}_k(\boldsymbol{\Phi}) \in \mathbb{S}^{n-1}$, affine with

[3] For simplicity, the test function v is discretised here using the same number of collocation points as ϕ, but this need not be the case.

respect to $\boldsymbol{\Phi}$, such that $\mathcal{Q}_k\{v\} = \boldsymbol{v}^\mathsf{T}\boldsymbol{Q}_k(\boldsymbol{\Phi})\boldsymbol{v}$. Consequently, for each wavenumber $k \leq k_c$ the Fourier-transformed spectral constraint can be approximated by the LMI $\boldsymbol{Q}_k(\boldsymbol{\Phi}) \succeq 0$.

Finally, the piecewise-linear approximation (6.63) turns the pointwise inequality $\phi(z) \leq 1$ into the $n+1$ constraints $\phi_i \leq 1$, $i = 0, \ldots, n$, written succinctly as the element-wise vector inequality $\boldsymbol{\Phi} \leq \mathbf{1}$. Similarly, the condition $\phi'(z) \geq 0$ becomes a set of n inequalities $\phi_{i-1} - \phi_i \leq 0$, $i = 1, \ldots, n$, which can be written in the vector form $\boldsymbol{A}\boldsymbol{\Phi} \leq \mathbf{0}$ with

$$\boldsymbol{A} := \begin{bmatrix} 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{bmatrix} \in \mathbb{R}^{n\times(n+1)}. \tag{6.72}$$

After substituting (6.63) into the objective function of (6.60) and defining

$$\boldsymbol{c} := \left[\int_0^1 \psi_0(z)\,\mathrm{d}z, \ldots, \int_0^1 \psi_n(z)\,\mathrm{d}z\right]^\mathsf{T}, \tag{6.73}$$

one concludes that the infinite-dimensional variational problem (6.60) is approximated by the SDP[4]

$$\begin{aligned} \max_{s,\boldsymbol{\Phi}} \quad & -s - \boldsymbol{c}^\mathsf{T}\boldsymbol{\Phi} \\ \text{s.t.} \quad & \boldsymbol{Q}_k(\boldsymbol{\Phi}) \succeq 0, \quad k = 1, \ldots, k_c, \\ & \|\boldsymbol{R}\boldsymbol{\Phi}\| \leq s. \end{aligned} \tag{6.74}$$

Similarly, (6.61) can be approximated as

$$\begin{aligned} \max_{s,\boldsymbol{\Phi}} \quad & -s - \boldsymbol{c}^\mathsf{T}\boldsymbol{\Phi} \\ \text{s.t.} \quad & \boldsymbol{Q}_k(\boldsymbol{\Phi}) \succeq 0, \quad k = 1, \ldots, k_c, \\ & \|\boldsymbol{R}\boldsymbol{\Phi}\| \leq s, \\ & \boldsymbol{\Phi} \leq \mathbf{1}, \end{aligned} \tag{6.75}$$

while (6.62) becomes

$$\begin{aligned} \max_{s,\boldsymbol{\Phi}} \quad & -s - \boldsymbol{c}^\mathsf{T}\boldsymbol{\Phi} \\ \text{s.t.} \quad & \boldsymbol{Q}_k(\boldsymbol{\Phi}) \succeq 0, \quad k = 1, \ldots, k_c, \\ & \|\boldsymbol{R}\boldsymbol{\Phi}\| \leq s, \\ & \boldsymbol{A}\boldsymbol{\Phi} \leq \mathbf{0}. \end{aligned} \tag{6.76}$$

[4]Problems (6.74)–(6.76) are not SDPs in standard form, but the terminology is justified because linear inequalities and SOCCs can be recast as LMIs (cf. section 2.3 and the discussion at the end of section 2.4).

Remark 6.2. Using (6.63) means that only lower bounds on the optimal values of (6.60)–(6.62) can be computed, because the true optimal $\phi(z)$ is unlikely to be piecewise linear. Moreover, assuming (6.70) enforces the Fourier-transformed spectral constraint only over a subset of the test function space Γ, which enlarges the set of feasible functions $\phi(z)$. Consequently, (6.74)–(6.76) estimate from above lower bounds on the true optimal values of (6.60)–(6.62), respectively. Compared to the polynomial approximation methods developed in chapter 4, this is a disadvantage, as one cannot guarantee that the numerical optimum bounds the exact one from above or below. Unless one aims at formulating a computer-assisted proof, however, this is not an issue because one expects the solutions of (6.74)–(6.76) to converge to those of (6.60)–(6.62) as the number of discretisation points increases. If rigorous computations were needed, one could try to adapt the analysis of section 4.4 and estimate the error between functions in Γ and their piecewise-linear approximation, thereby formulating SDPs to bound the optimal value of (6.60)–(6.62) rigorously from below. This analysis is left to future work.

Remark 6.3. A major advantage of using SDPs is that monotonicity and convexity can be enforced in a straightforward way using LMI-representable constraints, so problems (6.60)–(6.62) can be solved computationally with the same optimisation algorithms. One can then interrogate the bounding principle in a systematic way to identify key properties of its optimal solution, which may be used to progress with rigorous mathematical analysis. This applies not only to infinite-Pr Bénard–Marangoni convection, but to any convex upper-bounding variational problem obtained from the application of the background method. On the contrary, optimising over monotonic or convex background fields seems considerably more challenging if one follows the classical Euler–Lagrange variational approach: one has to solve a set of differential equations coupled to an inequality (in fact, a differential inequality in the convex case). However, it does not appear possible to enforce inequalities using the traditional numerical continuation strategies (Plasting & Kerswell, 2003) or the more recent time-marching methods (Wen *et al.*, 2013, 2015).

6.4.2 Comments on sparsity

The piecewise-linear basis functions ψ_i, $i = 0, \ldots, n$, used to discretise the function ϕ and the test function v in each Fourier-transformed spectral constraint have compact support. Moreover, the support of each ψ_i, $i = 1, \ldots, n-1$ overlaps only with that of $\psi_{i\pm1}$, while the supports of the boundary functions ϕ_0 and ϕ_n overlap only with those of ϕ_1 and ϕ_{n-1}, respectively. Recalling that $v_0 = 0$ and $v_n = v_{n-1}$ to enforce the BCs on v, this means that

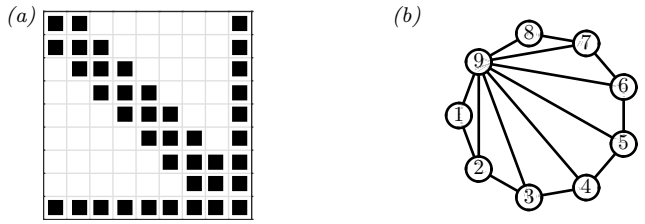


FIGURE 6.3: *(a)* Sparsity pattern of the 9×9 matrix $\boldsymbol{Q}_k(\boldsymbol{\Phi})$ obtained with $n = 10$ collocation points. *(b)* Graph representation of the matrix sparsity pattern in panel *(a)*.

in (6.71) v_1 is coupled only to v_2 and v_{n-1}, and each v_i, $i = 1, \ldots, n-2$ appears coupled only to $v_{i\pm1}$ and to v_{n-1}. As a consequence, the matrix $\boldsymbol{Q}_k(\boldsymbol{\Phi})$ has a "tridiagonal arrow" sparsity pattern, sketched in figure 6.3*(a)* for $n = 10$.

Such a sparsity pattern is chordal (cf. section 2.6) and its associated graph, shown in figure 6.3*(b)* for the case $n = 10$, has $n-3$ maximal cliques given by

$$\mathcal{C}_i = \{i,\, i+1,\, n-1\}, \quad i = 1, \ldots, n-3. \tag{6.77}$$

One can therefore apply Theorem 2.3 to replace the sparse LMI $\boldsymbol{Q}_k(\boldsymbol{\Phi}) \succeq 0$ with an equivalent set of $n-3$ LMIs, each corresponding to the 3×3 sub-matrix of $\boldsymbol{Q}_k(\boldsymbol{\Phi})$ defined by the indices in $\mathcal{C}_i$. Since the sub-matrices corresponding to consecutive cliques $\mathcal{C}_i$ and $\mathcal{C}_{i+1}$ have four common elements, three of which are independent because $\boldsymbol{Q}_k(\boldsymbol{\Phi})$ is symmetric, the decomposition procedure requires the introduction of $3(n-4)$ extra optimisation variables.

In practice, the cost of handling a very large number of such extra variables can offset the benefits of splitting the original LMI. For this reason, when n is large it is convenient to perform a partial decomposition, whereby m consecutive maximal cliques are combined (Fukuda *et al.*, 2000; Nakata *et al.*, 2003). Precisely, assume for generality that m is not a divisor of $n-3$, let $p := \left\lfloor \frac{n-3}{m} \right\rfloor$, and define the $p+1$ sets of indices

$$\mathcal{K}_j := \begin{cases} \displaystyle\bigcup_{i=(j-1)m+1}^{jm} \mathcal{C}_i, & j = 1, \ldots, p, \\ \displaystyle\bigcup_{i=pm+1}^{n-3} \mathcal{C}_i, & j = p+1. \end{cases} \tag{6.78}$$

One can check that each index set $\mathcal{K}_j$, $j = 1, \ldots, p$, has $m+2$ elements, while $\mathcal{K}_{p+1}$ has $n-1-pm$ elements (note that $n-1-pm < m+2$ by the definition of p). Decomposing the LMI $\boldsymbol{Q}_k(\boldsymbol{\Phi}) \succeq 0$ using the sets $\mathcal{K}_j$ instead of the maximal cliques $\mathcal{C}_i$ leads to an equivalent set of $p+1$ LMIs, p of size $(m+2) \times (m+2)$ and one of size $(n-1-pm) \times (n-1-pm)$, at the cost of introducing $3p$ extra optimisation variables.

6.4.3 Implementation details

The SDPs (6.74)–(6.76) were implemented and solved in MATLAB using YALMIP (Löfberg, 2004) and SDPT3 (Toh *et al.*, 1999; Tütüncü *et al.*, 2003) on a PC with a 3.40 GHz Intel® Core™ i7-4770 CPU and 16 GB of RAM. Each LMI $\boldsymbol{Q}_k(\boldsymbol{\Phi}) \succeq 0$ was decomposed as described in section 6.4.2 with $m = 8$, meaning that the largest implemented LMI had size 10×10. The Chebyshev nodes $z_i = [1-\cos(\pi i/q)]/2$, $i = 0, \ldots, q$ were utilised as collocation points in the sub-interval $(0.05, 0.98)$, while the finer distribution $z_i = [1 - \cos(\pi i/4q)]/2$, $i = 0, \ldots, 4q$ was employed in the boundary sub-intervals $[0, 0.05]$ and $[0.98, 1]$. After initial experiments, computations were run with $q = 512$, giving $n = 873$ collocation points in total. All results presented in section 6.4.4 change by less than 0.1% if larger q is used.

Chebyshev nodes were chosen because they naturally cluster near the boundaries and help resolve boundary layers near $z = 0$ and $z = 1$ in the optimal $\phi(z)$. These are expected even if no boundary conditions are imposed because to maximise the objective function in (6.60) one would like to choose $\phi(z) < 0$, but setting $\phi(z) \approx 1$ in the bulk of the domain is necessary to be able to satisfy the spectral constraint. However, it is possible to have $\phi(z) < 0$ in thin layers near the walls because the functions f_k, which act as a weight on ϕ in the Fourier-transformed spectral constraint (6.68), are small there for all k values (cf. figure 6.1). These observations are confirmed by the numerical results in section 6.4.4.

While boundary layers can in principle be resolved with a sufficiently fine distribution of Chebyshev points, refining the discretisation only near the boundaries through a secondary set of Chebyshev nodes helps reduce computational cost. In fact, since the number of rows and columns of each matrix $\boldsymbol{Q}_k(\boldsymbol{\Phi})$ grows linearly with the number of discretisation points, so does the number of LMIs obtained after the large and sparse LMI $\boldsymbol{Q}_k(\boldsymbol{\Phi}) \succeq 0$ is decomposed as described in section 6.4.2. Even ignoring the overhead due to the extra optimisation variables introduced by the decomposition procedure, therefore, the overall computational cost must grow at least linearly with the number of collocation points.

One complication to the implementation of (6.74) is that the cutoff wavenumber k_c is not known a priori, but it depends on $\boldsymbol{\Phi}$ according to (6.69). Thus, the same iterative procedure outlined in chapter 5 is employed: find the optimal $\boldsymbol{\Phi}$ using an initial guess k_0 for k_c, update the value of k_c using (6.69), check if $\boldsymbol{Q}_k(\boldsymbol{\Phi})$ is positive semidefinite for all $k \leq k_c$, and repeat the optimisation with the updated guess for k_c if any of these checks fail.

A second hurdle is that solving (6.74) with this iterative procedure becomes expensive when the Marangoni number is large because the cutoff wavenumber k_c, and therefore the number of LMI constraints, grows proportionally to $Ma^{1/2}$. For example, at $Ma = 2.5 \times 10^6$

the optimal ϕ satisfies $\|\phi - 1\|_\infty = 2$, so (6.69) gives $k_c = 1\,003$; when all 1 003 LMIs are considered in (6.74), SDPT3 takes more than 4 hours to converge. In an effort to reduce the CPU time requirements, a trial-and-error procedure in which only a subset of wavenumbers are considered in (6.74) was employed. Inspired by the numerical continuation method used by Plasting & Kerswell (2003), the Marangoni number was progressively increased according to the update rule $Ma_{i+1} = Ma_i \times 10^{1/p}$, which gives $p+1$ logarithmically spaced points between successive powers of 10. Given the critical wavenumbers $k_1, \ldots, k_m$ at one Marangoni number, the SDP for the next Ma was solved considering only wavenumbers in a window of width $2r$ around each k_i, $i = 1, \ldots, m$, *i.e.*, values of k such that

$$k \in \bigcup_{i=1}^{m} [k_i - r, k_i + r]. \tag{6.79}$$

The LMI $\boldsymbol{Q}_k(\boldsymbol{\Phi}) \succeq 0$ was subsequently checked for all remaining wavenumbers up to the cutoff value k_c. If any of these checks failed, the optimisation was repeated after adding the wavenumber with the largest constraint violation (*i.e.*, corresponding to the matrix $\boldsymbol{Q}_k$ with the most negative eigenvalue) to the list of critical values.

6.4.4 Results

The SDPs (6.74)–(6.76) were successfully solved for Marangoni numbers up to $Ma = 10^9$ using the procedure described in section 6.4.3 with $p = 19$ and $r = 10$. At each value of Ma, the optimal $\phi(z)$ was used to recover the optimal scaled background field $\rho(z)$ and the corresponding bound on the Nusselt number.

The most important results are the bounds on Nu, plotted in figure 6.4. Also shown for comparison are: the analytical bound $Nu \leq 0.803\, Ma^{2/7}$ proven in appendix A.8; the DNS results obtained by Boeck & Thess (2001); finally, the conductive value $Nu = 1$, which bounds the Nusselt number from below. The results are plotted in two ways: compensated by a factor of $Ma^{-2/7}$ to aid the visual comparison with the asymptotic scaling of the analytical bound, and compensated by $Ma^{-2/7}(\ln Ma)^{1/2}$. The main observation is that, while a gap with the DNS data remains, the fully optimal bounds and those computed after enforcing convexity grow more slowly than the analytical bound by $(\ln Ma)^{1/2}$. In particular, the fully optimal bound seems to exhibit the asymptotic behaviour

$$Nu \leq 1.285 Ma^{2/7} (\ln Ma)^{-1/2}. \tag{6.80}$$

In contrast, the bound on Nu asymptotes to $0.535 Ma^{2/7}$ when the background field is constrained to decrease monotonically. This suggests that the analytical bound attains the

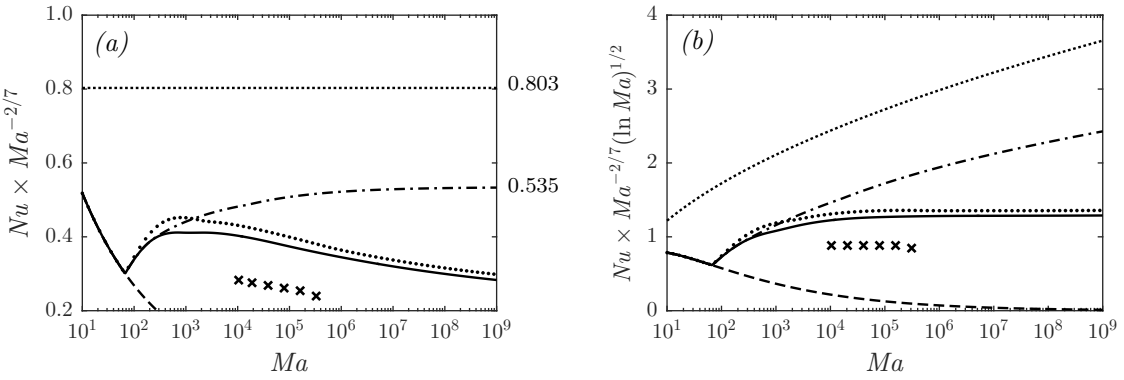


FIGURE 6.4: Comparison between: the fully optimal bounds on the Nusselt number, computed using the solution of (6.60) (——); the optimal monotonic bounds, computed using the solution of (6.61) (-·-·-); the optimal convex bounds, computed using the solution of (6.62) (·····). Also shown are the conductive Nusselt number $Nu = 1$ (- - -), the analytical bound $Nu \leq 0.803\, Ma^{2/7}$ (·····), and the DNS data by Boeck & Thess (2001) (×). Data in panel *(a)* are compensated by $Ma^{-2/7}$ to facilitate the visual comparison with the asymptotic scaling of the analytical bound. Data in panel *(b)* are compensated by $Ma^{-2/7}(\ln Ma)^{1/2}$.

optimal asymptotic scaling available when $\rho(z)$ is monotonic, but may be lowered by a logarithm upon construction of a non-monotonic background field.

Figure 6.5 shows the derivative of the optimal scaled background field, computed with each of the conic programmes (6.74)–(6.76), for a selection of values Ma. The derivative $\rho'(z)$ is plotted instead of $\rho(z)$ because, by virtue of (6.58), problems (6.60)–(6.62) can be rewritten in terms of $\rho'(z)$ alone. Since $\rho(z)$ can be recovered by integration using the boundary condition $\rho(0) = 0$, the derivative $\rho'(z)$ is the actual decision variable in (6.60)–(6.62). To ease the comparison, the profiles have been normalised by the magnitude of the boundary value $\rho'(0)$, which converges to -2 from above with increasing Ma as illustrated in figure 6.6*(a)*. Figure 6.6*(b)* demonstrates that in the fully optimal case the convergence is logarithmic. This was also observed when convexity was imposed, while power-law convergence was observed for the monotonic profiles. Such evidence corroborates the numerical conjecture that the optimal bound on Nu takes the asymptotic form (6.80).

As illustrated by figure 6.5, the optimal $\rho'(z)$ is negative for $Ma \leq 186.12$, meaning that the corresponding scaled background field decreases monotonically for sufficiently small Marangoni numbers. When Ma is raised, all profiles are characterised by boundary layers separated by a bulk region where $\rho'(z) \approx 0$. Note that the transition to the bulk region is not smooth when monotonicity or convexity are enforced, which is one of the reasons for preferring the piecewise-linear approximations of section 6.4.1 to the global polynomial approximation used chapters 4 and 5. In the fully optimal case, $\rho'(z)$ changes sign inside both boundary layers to reach positive local maxima, so the corresponding scaled background

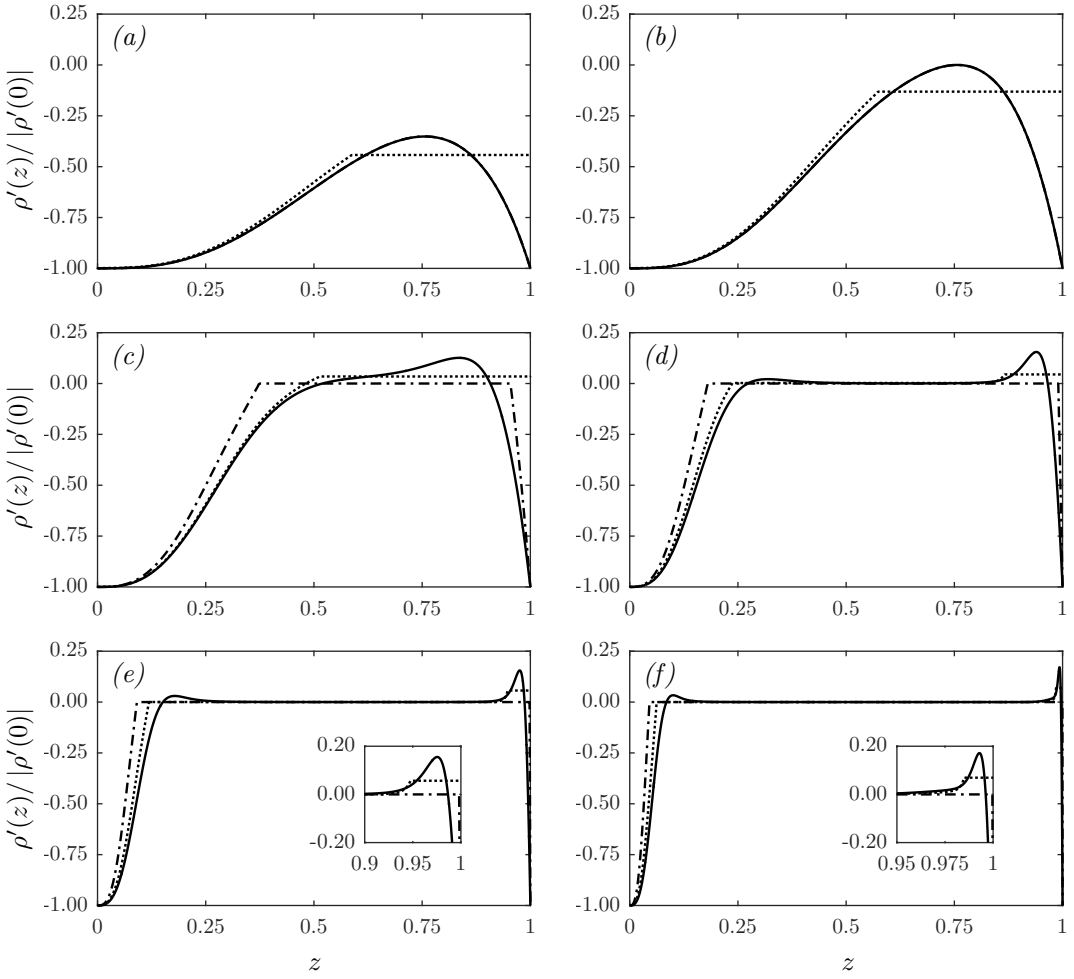


FIGURE 6.5: Normalised derivatives, $\rho'(z)/|\rho'(0)|$, of the fully optimal (——), optimal monotonic (-·-·-), and optimal convex (·····) scaled background fields. Profiles are shown for: *(a)* $Ma = 100$; *(b)* $Ma = 186.12$; *(c)* $Ma = 10^3$; *(d)* $Ma = 10^4$; *(e)* $Ma = 10^5$; and *(f)* $Ma = 10^6$. Inserts in panels *(e)–(f)* show a detailed view of the boundary layers near $z = 1$.

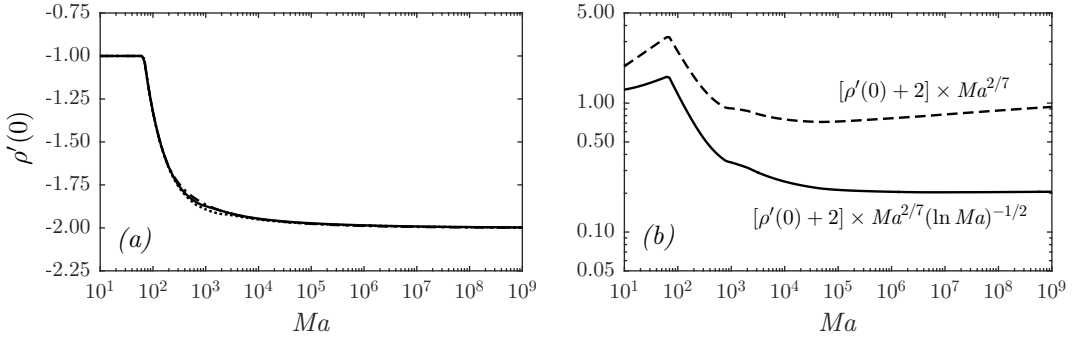


FIGURE 6.6: *(a)* The value $\rho'(0)$ for the fully optimal (——), monotonic (-·-·-), and convex (·····) background fields. All curves almost coincide. *(b)* Plot of $\rho'(0)+2$ for the fully optimal background fields, scaled by $Ma^{2/7}(\ln Ma)^{-1/2}$ (——) and by $Ma^{2/7}$ (- - -).

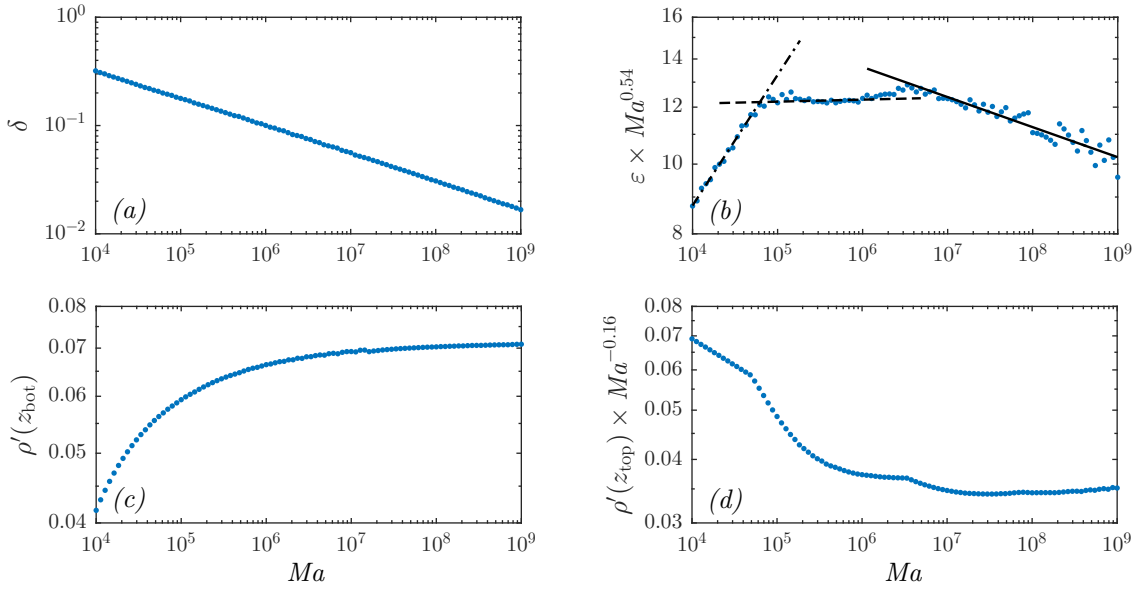


FIGURE 6.7: Details of the boundary layer structure of the fully optimal scaled background field derivative $\rho'(z)$ for $Ma \geq 10^4$. The dot-dashed, dashed, and solid lines in panel *(b)* indicate the approximate scaling laws (6.82a)–(6.82c), respectively.

field is characterised by non-monotonic boundary layers. Enforcing monotonicity removes these local maxima and makes the boundary layers thinner, while convexity prevents the local maximum near $z = 0$ and makes $\rho'(z)$ constant across the boundary layer near $z = 1$.

Figure 6.7 offers a more detailed description of the boundary layer structure of the fully optimal profiles for $Ma \geq 10^4$ (very similar results for the optimal convex profiles are not shown for brevity). Letting z_{bot} and z_{top} denote the coordinates of the positive local maxima of $\rho'(z)$ near $z = 0$ and $z = 1$, respectively, the quantities $\delta := z_{\mathrm{bot}}$ and $\varepsilon := 1 - z_{\mathrm{top}}$ can be taken as proxies for the thickness of each boundary layer. The boundary layer near the bottom of the domain ($z = 0$) is approximately self-similar at large Marangoni numbers, and least-squares power-law fits to the data in figures 6.7*(a)* and 6.7*(c)* for $Ma \geq 10^7$ return

$$\delta \approx 3.8\, Ma^{-0.26}, \qquad \rho'(z_{\mathrm{bot}}) \approx 0.07. \tag{6.81}$$

Note that the scaling exponent of δ is not far from $-2/7 \approx -0.286$, suggesting that the width of the boundary layer near $z = 0$ is one of the leading factors determining the scaling of the bound on Nu. It is tempting to conjecture that, asymptotically, $\delta \sim Ma^{-2/7}(\ln Ma)^{1/2}$, meaning that $Nu \sim \delta^{-1}$, but unfortunately the finite precision of the numerical data does not permit a clear identification of such a logarithmic trend. To obtain more precise values requires the solution of the SDPs (6.74)–(6.75) to a level of accuracy beyond the capabilities of SDPT3, and at considerably larger Ma (see section 6.6 for further discussion).

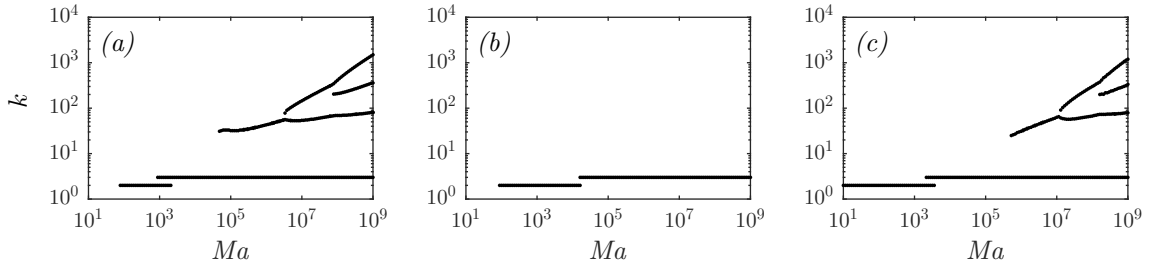


FIGURE 6.8: Bifurcation diagrams for the critical wavenumbers for: *(a)* the SDP (6.74) for the fully optimal background fields; *(b)* the SDP (6.75) for the optimal monotonic background fields; *(c)* the SDP (6.76) for the optimal convex background fields.

The situation is more complicated for the boundary layer near $z = 1$. In figure 6.7*(b)* one can identify three distinct regions, each characterised by a different scaling of ε:

$$\varepsilon \approx 1.65\, Ma^{-0.36} \qquad \text{for } Ma \lessapprox 5 \times 10^4, \tag{6.82a}$$

$$\varepsilon \approx 11.8\, Ma^{-0.54} \qquad \text{for } 5 \times 10^4 \lessapprox Ma \lessapprox 3 \times 10^6, \tag{6.82b}$$

$$\varepsilon \approx 24.3\, Ma^{-0.58} \qquad \text{for } Ma \gtrapprox 3 \times 10^6. \tag{6.82c}$$

Approximate scaling laws for $\rho'(z_{\text{top}})$ could also be determined in the first and third regions:

$$\rho'(z_{\text{top}}) \approx 0.34\, Ma^{-0.01} \qquad \text{for } Ma \lessapprox 5 \times 10^4, \tag{6.83a}$$

$$\rho'(z_{\text{top}}) \approx 0.04\, Ma^{0.16} \qquad \text{for } Ma \gtrapprox 3 \times 10^6. \tag{6.83b}$$

Once again, these scaling laws are only tentative due to the finite precision to which the SDPs for the optimal bounds could be solved. Note, however, that the large scatter in the data points in figure 6.7*(b)* is simply due to plotting ε after rescaling by $Ma^{0.54}$, which at large Ma amplifies small numerical inaccuracies.

Changes in the scaling of the boundary layer near $z = 1$ correspond to bifurcations in the critical wavenumbers for the conic programme (6.74). As illustrated in figure 6.8*(a)*, new critical wavenumbers appear at large values of k for $Ma \approx 5 \times 10^4$ and $Ma \approx 3 \times 10^6$. Another intermediate branch of critical wavenumbers appears for $Ma \approx 10^8$, but this does not seem to influence the scaling of the boundary layer. Such bifurcations can be explained in terms of the interaction, in the Fourier-transformed spectral constraint (6.68), between the boundary layer of $\rho'(z) = \phi(z) - 1$ and the function $f_k(z)$, which is almost entirely supported near $z = 1$ at large k. As illustrated by figures 6.8*(b)–(c)*, similar bifurcations were observed when solving (6.76) but not when solving (6.75), probably because the boundary layer near $z = 1$ of the optimal monotonic background fields is too thin to allow interesting interactions for wavenumbers below the cutoff value k_c.

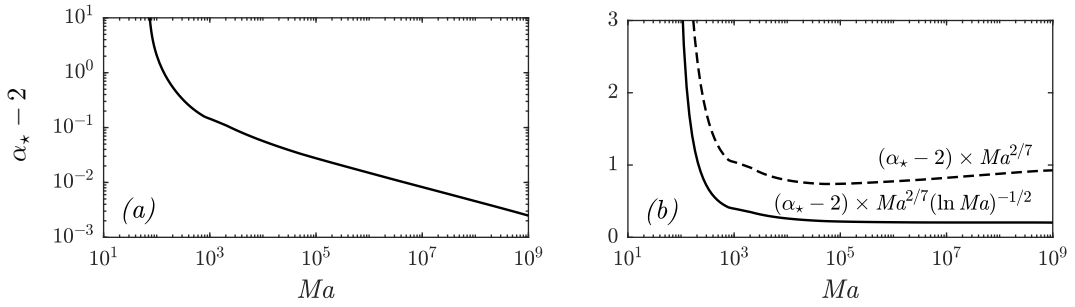


FIGURE 6.9: *(a)* Convergence of the optimal balance parameter $\alpha_\star$, computed using (6.43) and the fully optimal scaled background field, to the asymptotic value 2. *(b)* Plot of the difference $\alpha_\star - 2$, scaled by $Ma^{2/7}(\ln Ma)^{-1/2}$ (——) and by $Ma^{2/7}$ (- - -).

Finally, figure 6.9 shows the variation with Ma of the optimal balance parameter $\alpha_\star$, computed using (6.43) and the fully optimal background field. The results clearly show that $\alpha_\star$ converges to 2 as Ma is raised, and that the convergence rate is logarithmic,

$$\alpha_\star - 2 \sim Ma^{-2/7}(\ln Ma)^{1/2}. \tag{6.84}$$

This observation is consistent with the analysis of section 6.3 and corroborates the numerical conjecture that (6.80) is the correct functional form for the optimal bound on Nu.

6.5 Towards an improved analytical bound

The results presented in section 6.4.4 suggest that Hagstrom & Doering's bound $Nu \lesssim Ma^{2/7}$ may be improved by the logarithmic factor $(\ln Ma)^{-1/2}$. Despite the strong numerical evidence, however, whether the optimal bound scales logarithmically when $Ma \to \infty$ remains uncertain due to the limited range of Marangoni numbers spanned the present investigation (see section 6.6 for more on this issue). In particular, one cannot rule out the occurrence of further bifurcations in the critical wavenumbers that may cause a transition to a pure power-law behaviour with scaling exponent of $2/7$.

Uncertainty about the true asymptotic scaling notwithstanding, the numerical results demonstrate that if the current analytical bound $Nu \lesssim Ma^{2/7}$ can be improved, doing so requires a background temperature profile with non-monotonic boundary layers. More precisely, the optimal convex background fields and the corresponding bounds on Nu are evidence that what is needed is a relatively simple non-monotonic boundary layer near $z = 1$, while non-monotonicity near $z = 0$ only lowers the prefactor.

Taking advantage of these observations to improve the bound on Nu analytically, however, is likely to require a careful analysis of the sign-indefinite term in each Fourier-transformed

spectral constraint, restated here in terms of the variable $\rho(z)$ in the slightly rearranged form

$$\mathcal{Q}_k\{v\} = \left\|v'\right\|_2^2 + k^2 \left\|v\right\|_2^2 - Ma\, v(1) \int_0^1 \rho'(z)\, f_k(z)\, v(z)\, \mathrm{d}z \geq 0 \quad \forall v \in \Gamma. \tag{6.85}$$

For example, simply estimating

$$\left| Ma\, v(1) \int_0^1 \rho'(z)\, f_k(z)\, v(z)\, \mathrm{d}z \right| \leq Ma\, |v(1)| \int_0^1 \left|\rho'(z)\right|\, |f_k(z)|\, |v(z)|\, \mathrm{d}z \tag{6.86}$$

and requiring

$$\left\|v'\right\|_2^2 + k^2 \left\|v\right\|_2^2 - Ma\, |v(1)| \int_0^1 \left|\rho'(z)\right|\, |f_k(z)|\, |v(z)|\, \mathrm{d}z \geq 0 \quad \forall v \in \Gamma, \tag{6.87}$$

as done in appendix A.8 to prove the bound (6.48), forces the optimal ρ to decrease monotonically. In fact, if ρ satisfies (6.87) and $\rho'(z) \geq 0$ for $z \in \mathcal{U} \subset [0,1]$, the profile

$$\tilde{\rho}'(z) := \begin{cases} \rho'(z), & z \in [0,1] \setminus \mathcal{U}, \\ 0, & z \in \mathcal{U}, \end{cases} \tag{6.88}$$

also satisfies (6.87), but decreases monotonically and gives a larger objective value in (6.36). In light of the numerical results presented in section 6.4.4, one expects any bound obtained using the estimate (6.86) to be no better than $Nu \lesssim Ma^{2/7}$.

A better approach is to reformulate the Fourier-transformed spectral constraint (6.85) before applying any estimates. Without any loss of generality, let $\delta \in (0,1)$ and write

$$\rho'(z) = \begin{cases} g(z), & 0 \leq z \leq \delta, \\ h(z), & \delta \leq z \leq 1. \end{cases} \tag{6.89}$$

Here, δ represents the thickness of the boundary layer of the optimal background field near $z = 0$. With this choice, the Fourier-transformed spectral constraint (6.85) becomes

$$\begin{aligned} \mathcal{Q}_k\{v\} = \left\|v'\right\|_2^2 + k^2 \left\|v\right\|_2^2 - Ma\, v(1) \int_0^\delta g(z)\, f_k(z)\, v(z)\, \mathrm{d}z \\ - Ma\, v(1) \int_\delta^1 h(z)\, f_k(z)\, v(z)\, \mathrm{d}z \geq 0 \quad \forall v \in \Gamma. \end{aligned} \tag{6.90}$$

Since this inequality is homogeneous in v and holds when $v(1) = 0$, it suffices to restrict the attention to test functions normalised such that $v(1) = 1$. After adding and subtracting

$Ma\int_\delta^1 h(z)f_k(z)\,\mathrm{d}z$ one then needs to check that

$$\left\|v'\right\|_2^2 + k^2\left\|v\right\|_2^2 - Ma\int_0^\delta g(z)\,f_k(z)\,v(z)\,\mathrm{d}z \\ + Ma\int_\delta^1 h(z)\,f_k(z)\,[1-v(z)]\,\mathrm{d}z - Ma\int_\delta^1 h(z)f_k(z)\,\mathrm{d}z \geq 0. \quad (6.91)$$

If $\int_\delta^1 h(z)f_k(z)\,\mathrm{d}z < 0$, the last term in (6.91) gives a net positive contribution to the spectral constraint, and can be used to control the sign-indefinite terms. Recalling from figure 6.1 that $f_k(z) \leq 0$, this requires $h(z) > 0$ over a sufficient portion of the interval $(\delta, 1)$, meaning that the background field ρ should not decrease monotonically. Moreover, $h(z)$ should be supported in a boundary layer near $z = 1$ if the fourth term in (6.91) is to be controlled. Consequently, a non-monotonic boundary layer near $z = 1$ helps enforcing the spectral constraint. The situation is similar in infinite-Pr Rayleigh–Bénard convection (Doering *et al.*, 2006; Otto & Seis, 2011), so this observation is perhaps not surprising.

In addition to casting light on the role of the surface boundary layer, identity (6.91) may also offer a starting point to improve the bound $Nu \lesssim Ma^{2/7}$ analytically. Recalling the boundary condition $v(0) = 0$ and the normalisation condition $v(1) = 1$, one possibility is to use the fundamental theorem of calculus and the Cauchy–Schwarz inequality to bound

$$\left|\int_0^\delta g(z)\,f_k(z)\,v(z)\,\mathrm{d}z\right| \leq \int_0^\delta |g(z)\,f_k(z)|\,\left|\int_0^z v'(t)\,\mathrm{d}t\right|\,\mathrm{d}z \\ \leq \left\|v'\right\|_2 \int_0^\delta |g(z)\,f_k(z)|\,\sqrt{z}\,\mathrm{d}z \quad (6.92)$$

and

$$\left|\int_\delta^1 h(z)\,f_k(z)\,[1-v(z)]\,\mathrm{d}z\right| \leq \int_\delta^1 |h(z)\,f_k(z)|\,\left|\int_z^1 v'(t)\,\mathrm{d}t\right|\,\mathrm{d}z \\ \leq \left\|v'\right\|_2 \int_\delta^1 |h(z)\,f_k(z)|\,\sqrt{1-z}\,\mathrm{d}z. \quad (6.93)$$

Defining

$$a_k := \int_0^\delta |g(z)\,f_k(z)|\,\sqrt{z}\,\mathrm{d}z, \quad (6.94a)$$

$$b_k := \int_\delta^1 |h(z)\,f_k(z)|\,\sqrt{1-z}\,\mathrm{d}z, \quad (6.94b)$$

$$c_k := -\int_\delta^1 h(z)f_k(z)\,\mathrm{d}z \quad (6.94c)$$

to ease the notation, a sufficient condition for (6.91) is that

$$\left\|v'\right\|_2^2 - Ma\,(a_k + b_k)\left\|v'\right\|_2 + Ma\,c_k \geq 0, \quad (6.95)$$

which in turn is satisfied for all v if

$$a_k + b_k \leq 2\sqrt{\frac{c_k}{Ma}}. \tag{6.96}$$

Given a candidate background field, condition (6.96) can be checked for all wavenumbers up to the 'cutoff' wavenumber k_c in (6.69).

Improving the bound $Nu \lesssim Ma^{2/7}$ via (6.96), however, may not be straightforward. To illustrate one of the difficulties it is useful to consider a simple background field, whose form is motivated by the shape of the derivatives of the optimal convex background fields in figure 6.5, and by the fact that the corresponding bounds in figure 6.6*(b)* exhibit the same asymptotic behaviour as the fully optimal one. Precisely, fix

$$g(z) = -2, \qquad\qquad h(z) = \begin{cases} 0, & \delta \leq z < 1-\varepsilon, \\ \gamma, & 1-\varepsilon \leq z \leq 1, \end{cases} \tag{6.97}$$

with $\gamma > 0$ a constant (independent of Ma) and $\varepsilon \ll 1$ but such that $1/\varepsilon \leq k_c \sim Ma^{1/2}$.

When $k \leq 1/\varepsilon$, using the Taylor expansions $f_k(z) \sim z^2$ near $z = 0$ and $f_k(z) \sim k(z-1)$ near $z = 1$ gives

$$a_k \sim \delta^{7/2}, \qquad b_k \sim \gamma\, k\, \varepsilon^{5/2}, \qquad c_k \sim \gamma\, k\, \varepsilon^2. \tag{6.98}$$

With these estimates, condition (6.96) can be rearranged as

$$\delta^{7/2} \lesssim 2\,\varepsilon \sqrt{\frac{\gamma\, k}{Ma}} \left(1 - \sqrt{\gamma\, k\, Ma\, \varepsilon^3}\right). \tag{6.99}$$

When $k \sim 1$ the two sides of (6.99) can be balanced by taking $\varepsilon \sim Ma^{-1/3}$ and $\delta \sim Ma^{-5/21}$, which yields

$$Nu \leq \frac{2}{2\delta - \gamma\varepsilon - \sqrt{\gamma(\gamma+2)\varepsilon}} \sim \delta^{-1} \sim Ma^{5/21}. \tag{6.100}$$

Interestingly, the exponent $5/21 \approx 0.238$ is extremely close to that of the best power-law fit $Nu \sim Ma^{0.24}$ to the DNS data by (Boeck & Thess, 2001, see equation (4) in their paper). In these simulations convection takes the form of stationary rolls with energy only at low wavenumbers, and the deviation from the theoretical asymptotic scaling exponent $2/9 \approx 0.222$ can be attributed to the contribution to the heat transfer of the thermal boundary layer near the surface (see the discussion after equation (13) in Boeck & Thess, 2001). Although this contribution is expected to vanish as $Ma \to \infty$, the background method could yield a bound that agrees well with observations for a range of Marangoni numbers if the stability of the rolls were deduced rigorously from the governing equations.

The lack of such information, however, means that (6.96) must be satisfied for all values k up to $k_c \sim Ma^{1/2}$. In particular, setting $k = 1/\varepsilon$ (which is no larger than k_c by assumption) shows that one must choose $\varepsilon \lesssim Ma^{-1/2}$ and $\delta \lesssim Ma^{-2/7}$, so the eventual bound on Nu cannot grow more slowly than $\sim Ma^{2/7}$. The issue remains when one lets γ increase with Ma to mimic the behaviour of the numerically optimal profiles (cf. panel *(d)* in figure 6.7), because what is gained in (6.99) is exactly outbalanced by the need of testing wavenumbers up to $k_c \sim \gamma Ma^{1/2}$. Thus, it is expected that to improve Hagstrom & Doering's scaling using (6.96) will require careful estimates of a_k, b_k and c_k at large wavenumbers, perhaps in conjunction with a more sophisticated choice of background field.

6.6 Challenges for computations in the asymptotic regime

As mentioned at the beginning of section 6.5, the true asymptotic nature of the numerically optimal bound remains uncertain due to the limited range of Marangoni numbers that could be studied. This kind of uncertainty is inherent to any numerical investigation, but the challenges faced by SDPs in reaching the asymptotic regime deserve further discussion.

Contrary to what was observed in chapter 5, the main obstacle to computing bounds for $Ma \gg 10^9$ is cost: despite exploiting sparsity, proceeding from $Ma = 10^8$ to $Ma = 10^9$ took more than 48 hours, and to achieve significant further progress would require computational resources beyond those available to the present investigation. One difficulty—already pointed out in chapter 5—is that, at large Marangoni numbers, checking whether a candidate background field satisfies the Fourier-transformed spectral constraints up to the cutoff wavenumber k_c becomes a burden even when computations are parallelised. For example, $k_c = 20\,073$ at $Ma = 10^9$, meaning that 20 073 eigenvalue or Cholesky decompositions must be carried out after each iteration of the wavenumber-tracking procedure described in section 6.4.3. The situation is worsened by the occurrence of bifurcations in critical wavenumbers, because more iterations are needed to correctly track all critical branches. Performance could be not improved by taking smaller steps in Ma, because doing so slows progress towards higher Marangoni numbers. Increasing the parameter r in (6.79) also does not help much, because the cost of adding more LMIs to the SDP at each iteration offsets the reduction in number of iterations required to identify the critical wavenumbers.

A possible solution to the critical wavenumber identification problem could be to apply the time-marching algorithm of Wen *et al.* (2013, 2015) to the optimality conditions for the SDPs (6.74)–(6.76). This method appears to select the critical wavenumbers efficiently, although convergence to the optimal background field can be slow (Wen *et al.*, 2015). Fast

but less accurate solvers for SDPs, such as SCS (O'Donoghue *et al.*, 2016) may have similar benefits and drawbacks, with the advantage that finely tuned open-source implementations are readily available. Irrespective of which method is utilised, once the critical wavenumbers have been identified the optimal solution can be computed to higher accuracy using SDPT3.

The numerical methodology described in section 6.4 and the possible improvements discussed above can of course be applied beyond Bénard–Marangoni convection. However, studying the asymptotic regime of more complex background method problems will require overcoming some additional obstacles. Spectral constraints with multiple test functions, such as those encountered in shear flows (Plasting & Kerswell, 2003; see also chapter 5) or convection at finite Prandtl number (Doering & Constantin, 1996; Otero *et al.*, 2002), yield SDPs with larger LMIs than those considered in this chapter. Current state-of-the-art SDP solvers can handle many small LMIs efficiently, but the cost of a single LMI grows as a nonlinear function of its size. This is also an issue for problems with two- and higher-dimensional background fields, because horizontal Fourier expansions do not allow for a mode-by-mode decomposition of the spectral constraint, and after discretisation one obtains a single, large LMI instead of a set of smaller, independent LMIs corresponding to each wavevector.

Interest in the development of algorithms for large-scale SDPs has recently grown (see, for example, Sun *et al.*, 2014; Madani *et al.*, 2015; O'Donoghue *et al.*, 2016; Pakazad *et al.*, 2017; Zheng *et al.*, 2017*a*,*b*), and it is likely that more efficient solvers will become available in the near future. Meanwhile, the (current) unfavourable scalability of algorithms for SDPs can be mitigated by taking advantage of special properties of the particular background field problem at hand. For instance, symmetries can be exploited to reduce the number of degrees of freedom needed to discretise the background field or the test functions in the spectral constraint (however, this is not the case for Bénard–Marangoni convection). The choice of discretisation method also plays an important role because, as demonstrated in this and the previous chapters, it directly impacts the sparsity of the eventual LMI. In this respect, the piecewise-linear approximations considered in this chapter are more attractive than the polynomial series expansions considered previously in this thesis, because they results in LMIs with chordal sparsity pattern. The same is true when one uses multidimensional piecewise-polynomial representations in the spirit of finite-element (FE) methods: approximately speaking, chordal sparsity arises from the fact that each element in the FE mesh is coupled only to its neighbours, and only through the degrees of freedom at the boundary elements. Exploiting chordal sparsity to decompose large LMIs into multiple smaller ones proved extremely effective for the present study of Bénard–Marangoni convection, and the same should be true for other background method problems.

6.7 Conclusions

This chapter investigated the vertical heat transfer in Bénard–Marangoni convection of a fluid layer with infinite Prandtl number by means of rigorous upper bounds on the Nusselt number. First, the background method analysis by Hagstrom & Doering (2010) was extended to include balance parameters and formulate a new variational principle for the bound. Doing so led to the new analytical result $Nu \leq 0.803 \times Ma^{2/7}$, the prefactor being approximately 4.2% lower compared to the previous best bound, but it was demonstrated that optimising the balance parameters cannot affect the asymptotic scaling of the optimal bounds compared to Hagstrom & Doering's original formulation. Using SDP approximations of the upper-bounding variational problem, the bound on Nu was optimised over all background fields, as well as over two smaller families constrained, respectively, by monotonicity and by convexity. The main results were the numerical conjecture that the fully optimal bounds have the form $Nu \lesssim Ma^{2/7}(\ln Ma)^{-1/2}$ for large Marangoni numbers, and the observation that such a logarithmic bound requires a background field with a non-monotonic boundary layer near the top boundary.

Whether the logarithmic scaling observed numerically can be proven analytically remains an open question. The analysis presented in section 6.5 suggests a way forward by replacing the spectral constraint with the sufficient condition (6.96). Using (6.96) is an attractive option because it is easier to check than the spectral constraint for a candidate background field, and the role of non-monotonicity is apparent. Moreover, the fact that enforcing (6.96) at large wavenumbers seems to constrain the bound on Nu is reminiscent of the bifurcations in critical wavenumbers observed in the numerical results (cf. figure 6.8). In summary, (6.96) seems to capture the essential features of the spectral constraint.

Should (6.96) prove too strong, the analysis of the energy stability problem (Fantuzzi & Wynn, 2017) may be adapted to derive an inequality that exactly enforces each Fourier-transformed spectral constraint. The disadvantage is that such an inequality may not be analytically tractable except for very simple choices of the background field. On the other hand, it may be possible to check this condition numerically and confirm that a candidate background field can indeed achieve a logarithmic bound, leaving "only" the task of constructing the correct estimates to prove so rigorously. Alternatively, one may consider the Lagrangian dual of the variational problem obtained with background method. This amounts to constructing the temperature and velocity fields that maximise the heat transfer subject to the momentum equation (which in the infinite-Pr limit is an algebraic condition relating temperature and velocity), the boundary conditions, and suitably averaged versions

of the advection-diffusion equation for the temperature.[5] However, only the fields achieving the maximal heat transfer yield a bound on Nu, so the maximisation must be solved exactly, and an asymptotic solution using Busse's "multi-α" solution method (see for example Busse, 1979) is complicated both by the lack of vertical symmetry and by the Neumann BCs.

Irrespective of how the variational problem for the upper bound on Nu is analysed, however, the numerical results presented in this chapter reveal that applying the background method to the temperature field cannot close the gap with the phenomenological prediction $Nu \sim Ma^{2/9}$ by Boeck & Thess (2001). It is possible that Boeck & Thess's assumption that steady convection rolls remain stable as $Ma \to \infty$ is incorrect, making a scaling exponent of $2/9$ unattainable with any bounding method. Proving so rigorously requires a lower bound on Nu that grows faster than $Ma^{2/9}$, which can also not be achieved with the background method because the unstable conduction solution saturates the constant lower bound $Nu \geq 1$. Further numerical simulations at high Ma seem essential to investigate the issue, and the observation of steady convection rolls would provide further supporting evidence for Boeck & Thess's phenomenological prediction. Determining the stability of the steady rolls is of interest also to reveal if the bifurcations in critical wavenumbers observed in the present computations correspond to yet unobserved physical instabilities.

In case the scaling law $Nu \sim Ma^{2/9}$ were confirmed by further DNSs, the derivation of rigorous bounds on Nu that exhibit the same asymptotic scaling will necessarily require bounding techniques beyond the background method. Unfortunately, the formulation of a wall-to-wall optimal transport problem (Hassanzadeh *et al.*, 2014; Tobasco & Doering, 2017) is not suited to the study Bénard–Marangoni convection at infinite-Pr. In fact, the optimal transport approach treats the temperature as a passively advected and diffusing scalar, and one looks for the (generally time-dependent) incompressible velocity field that maximises the passive vertical transport of heat subject to a maximum power budget. However, in infinite-Pr Bénard–Marangoni convection the flow velocity is a linear function of the temperature field, which is effectively the only dynamical variable. This coupling is crucial in the background method analysis, so improving the bound on Nu without taking it into account seems unlikely.

It would then be tempting to formulate the "ultimate" optimal wall-to-wall transport problem using the temperature as the decision variable, and let the flow velocity be specified as a function of it. However, this corresponds to searching for the exact solution of the equations of motion (6.1a)–(6.1c) with maximal heat transfer, so progress does not appear

[5] The duality between the two approaches was pointed out in chapter 1, and for a detailed discussion in the context of Rayleigh–Bénard convection, the reader is referred to Plasting & Ierley (2005).

possible. Difficulties remain when one drops the time dependence: maximising the heat transfer among the steady solutions is not much easier, and in any case the eventual bound would rely on the unproven assumption that unsteady flows cannot transport more heat than steady ones. Nonetheless, the construction of exact solutions is of interest because knowledge of a (possibly unstable) flow with Nusselt number Nu_{ss} places a strict limit on what can be achieved by upper-bounding theory. In particular, any bounds that apply equally to all solutions of (6.1a)–(6.1c) cannot be better than Nu_{ss}. Moreover, any flow with heat transfer $Nu_{\mathrm{ss}} \gg \text{const.} \times Ma^{2/9}$ would demonstrate that Boeck & Thess's phenomenological scaling applies at most to a particular subset of all possible convective flows.

While improving the rigorous upper bound on Nu using the "ultimate" wall-to-wall optimal transport approach described above appears challenging, it may be possible to consider successively weaker, tractable relaxations of it. The idea stems from the aforementioned realisation that the background method analysis is dual to the problem of maximising the heat transfer over all temperature (and associated velocity) fields that satisfy a set of constraints obtained by averaging the heat equation (Plasting & Ierley, 2005). The upper bound on Nu may therefore be improved by including additional constraints implied by the heat equation, but not the heat equation itself. A simple way to do so is through a general bounding framework that encompasses the background method (Chernyshenko *et al.*, 2014; Chernyshenko, 2017). The essence of this approach is to construct a functional $\mathcal{V}$ of the flow variables subject to a positivity condition akin to the spectral constraint in the background method. Each term in this functional can be interpreted as enforcing a particular constraint implied by the governing equations. Taking $\mathcal{V}$ to be the volume average of a quadratic polynomial of the flow variables gives the same bound as the background method (Chernyshenko, 2017), but experience with finite-dimensional systems (Fantuzzi *et al.*, 2016; Goluskin, 2017; Tobasco *et al.*, 2018) indicates that considering more general functionals—for instance, volume averages of higher-than-quadratic polynomials of the flow variables—could yield significant improvements. Although the construction of suitable functionals may be beyond the reach of purely analytical work, in this case too can progress be guided by solving SDPs. Whether bounds can be computed in the asymptotic regime is highly dependent on the availability of efficient algorithms for semidefinite programming, but recent developments in this field (see, for example, Sun *et al.*, 2014; O'Donoghue *et al.*, 2016; Zheng *et al.*, 2017*a*,*b*) give hope that Bénard–Marangoni convection and other turbulent hydrodynamic systems may be studied successfully in the near future.

Chapter 7

Conclusions and outlook

The underpinning theme of this thesis has been the development and application of numerical techniques for the optimisation of bounds, derived using the background method, on asymptotic or time-averaged quantities that describe a turbulent system. The key idea of the background method is to consider the evolution of the system around a steady background field and show that, if a certain integral quadratic form that depends affinely on the background field is positive semidefinite (a condition known as the spectral constraint because it requires that the eigenvalues of a self-adjoint operator are non-negative), then the quantity of interest can be bounded in terms of the background field alone. Numerical optimisation of the background field and of the corresponding bound is essential if the background method is to be truly useful, either to test phenomenological theories against rigorous analysis of the system's governing equations, or to make quantitative predictions when direct numerical simulations necessitate prohibitively large resources. Until recently, however, the construction of optimal background fields has demanded sophisticated numerical strategies because, although optimality conditions in the form of Euler–Lagrange (EL) equations can be formulated relatively easily, they often admit multiple solutions. Since only one of these satisfies the spectral constraint, care must be taken to avoid the computation of so-called "spurious" solutions, which satisfy the EL equations, but violate the spectral constraint.

For three classical fluid dynamical systems (namely, two-dimensional porous media convection, plane Couette flow, and two-dimensional Rayleigh–Bénard convection between stress-free isothermal plates), Wen *et al.* (2013, 2015) demonstrated that spurious solutions can be avoided by evolving a time-dependent version of the EL equations until a steady state is reached. One question investigated in this thesis is whether the same is true for other systems, and in chapter 3 this time-marching approach was applied to bound the asymptotic energy of solutions of the Kuramoto–Sivashinsky (KS) equation. For this system, the variational problem for the optimal background field has two spectral constraints, and it was

observed that convergence to spurious solutions can occur unless possible degeneracies in the eigenvalue problems associated with the spectral constraints are taken into account when deriving the EL equations. Precisely, an informal analogy with finite-dimensional optimisation problems subject to linear matrix inequalities (LMIs) revealed that, when constructing the Lagrangian functional for the optimal background field problem, one should include as many independent "copies" of each spectral constraint as the largest possible multiplicity of the ground-state eigenvalue of the corresponding linear operator. Ground-state eigenvalues can have multiplicity 2 for the KS equation, and considering two copies of each spectral constraint in the Lagrangian yields extra terms in the EL equations. These appear to destabilise any spurious steady states, and thereby allow for the robust computation of the optimal background field. Unfortunately, it was also demonstrated that the argument put forward by Wen *et al.* (2015) to prove rigorously that spurious solutions are unstable does not extend to the KS equation. Consequently, one cannot guarantee *a priori* that if the time-marching algorithm converges to a steady state, then the corresponding background field is the optimal one. Similar issues are likely to arise when optimising background fields beyond the KS equation, and while it is conjectured that in most cases the time-marching method will indeed converge to the desired optimal solution, its performance should be assessed on a problem-by-problem basis until a more comprehensive theoretical convergence analysis is carried out.

Motivated by the difficulties encountered in the study of the KS equation, the rest of this thesis has focussed on the development of an alternative numerical approach to optimise background fields. This was born out of the observations that the spectral constraint is the infinite-dimensional equivalent of an LMI, and that it can often be posed as the condition that a set of integral quadratic forms with one-dimensional compact domain of integration are positive semidefinite. For these reasons, chapter 4 considered the problem of minimising a linear cost function subject to a one-dimensional affine homogeneous integral inequality, *i.e.*, the requirement that an integral functional, affinely dependent on the decision variables, is non-negative for all functions subject to prescribed homogeneous boundary conditions. Legendre series expansions were used to show that the feasible set of an affine homogeneous quadratic integral inequality can be approximated, either from the inside or from the outside, by sets represented by LMIs. The optimal cost value can therefore be bounded rigorously (modulo numerical roundoff errors), both from above and from below, through the solution of SDPs. It was also proven that if the optimal cost value is attained by an optimal point, then arbitrarily accurate lower bounds can be obtained with outer LMI approximations constructed via truncated series expansions with sufficiently many terms. Convergence

of the upper bounds computed with inner LMI approximations to the true optimal cost, instead, could not be established, but numerical experiments demonstrate that it is often observed in practice. Future work should try to determine for which sub-classes of integral inequalities, if any, convergence of the upper bounds can be proven rigorously.

The series expansion techniques developed in chapter 4, implemented in an open-source MATLAB toolbox called QUINOPT, were employed in chapter 5 to optimise bounds on the energy dissipation coefficient C_ε for two- and three-dimensional stress-driven shear flows. Semidefinite programming proved robust and efficient when the Grashoff number Gr—the non-dimensional measure of the strength of the imposed shear—was not too large ($Gr \leq 10^5$ in two dimensions and $Gr \leq 10^4$ in three dimensions). Optimisation of the background field improves the constant (*i.e.*, independent of Gr) analytical bounds proven by Hagstrom & Doering (2014) by more than 10 times at the largest values of Gr considered in the computations. On the other hand, the optimal bounds on C_ε appear to approach a constant value as Gr grows, suggesting that Hagstrom & Doering's results are optimal as far as their asymptotic scaling with Gr is concerned. Constant C_ε would be consistent with Kolmogorov's theory of turbulence, according to which dissipation becomes independent of the fluid's viscosity at large Grashoff numbers. Of course, whether bounds obtained with the background method capture the asymptotic scaling of the dissipation measured in real flows remains to be seen, but this question could not be answered in this work due to the lack of readily available experimental/numerical data. In addition, it should be stressed that the asymptotic behaviour of the optimal bounds was not determined with any degree of certainty, due to the limited range of Gr that could be studied. In particular, in light of the results obtained for shear-driven Bénard–Marangoni convection in chapter 6 (and summarised below), it may be possible that the optimal bounds do not approach a constant, but rather decrease logarithmically with Gr. To settle this matter, optimal bounds should be computed at values of Gr well within the asymptotic regime, but it was not possible to do so here because the SDPs set up by QUINOPT became increasingly ill conditioned as Gr was raised. This issue should be investigated thoroughly in the future if semidefinite programming is to be applied successfully to a wider range of optimal background field problems. As discussed in section 5.5, however, accurate computations that extend far into the asymptotic regime are also likely to be prevented by the prohibitively large computational resources currently required to solve large SDPs accurately via general-purpose solvers.

A first attempt to alleviate the current poor scalability of algorithms for semidefinite programming was made in chapter 6, which studied Bénard–Marangoni convection—the

motion of a layer of fluid driven by shear stresses due to thermally induced gradients in surface tension—at infinite Prandtl number. The background method with balance parameters was applied to bound the Nusselt number Nu, the non-dimensional measure of the vertical heat transfer enhancement due to convection, as a function of the Marangoni number Ma, the non-dimensional measure of the thermal forcing. After changing variables to obtain a convex variational problem for a scaled background temperature field, piecewise-linear approximations were utilised to turn the spectral constraint into a set of sparse LMIs that can be implemented efficiently through chordal decomposition methods (cf. section 2.6 and references therein). This made it possible to optimise the background field for Marangoni numbers up to 10^9, giving compelling numerical evidence that the optimal bounds satisfy $Nu \lesssim Ma^{2/7}(\ln Ma)^{-1/2}$ at large Ma. Further optimisation over the two restricted classes of monotonic and convex background fields revealed that the asymptotic scaling of the analytical bound $Nu \leq 0.803 Ma^{2/7}$, also proven in this thesis, is optimal within the class of monotonic background fields. On the contrary, non-monotonic boundary layers near the surface of the fluid layer help to enforce the spectral constraint and lower the bound by a logarithmic factor. The numerical results also suggest that a boundary layer profile with constant slope suffices to lower a pure power-law bound, giving hope that a rigorous proof of the logarithmic correction observed numerically may be within the reach of analytical work.

The overall conclusion stemming from the results presented throughout this thesis is that, provided that the computational challenges associated with the solution of large SDPs can be overcome in the future, semidefinite programming provides a very attractive framework for the construction of optimal background fields. One reason is that the availability of software packages for the formulation and solution of the relevant SDPs, such as the MATLAB toolbox QUINOPT developed as part of this work, eliminates the need for the implementation of sophisticated, problem-specific numerical strategies and enables researchers to concentrate on modelling and analysis. Another reason is that, since many types of constraints can be represented as LMIs, semidefinite programming offers a flexible optimisation framework, in which constraints on the background field (for instance, monotonicity) can be added or removed at will without requiring any changes to the numerical algorithm. As demonstrated in chapter 6, one can therefore interrogate the bounding principles obtained with the background method in a systematic way, in order to identify key properties of the optimal background fields and guide rigorous mathematical analysis. However, it should be remarked that the fluid dynamical systems studied in this thesis—shear flows driven by surface stresses and Bénard–Marangoni convection at infinite Prandtl number—are among

the simplest, and a successful application of semidefinite programming to more complex systems is highly dependent on the resolution of a few outstanding issues.

From a theoretical perspective, the SDP approximations of integral inequalities developed in chapter 4 should be extended to allow for the discretisation of multi-dimensional spectral constraints that cannot be reduced to one-dimensional integral inequalities through Fourier expansions. Outer approximations, obtained when spectral constraints are enforced only over finite-dimensional subsets of the test function spaces on which they are formally required to hold, can be formulated in a relatively straightforward way upon discretising the test functions using well-established spectral or finite-element methods. The construction of numerically tractable inner approximations, instead, is expected to be more challenging because it requires infinite series expansions and functional estimates that, crucially, depend on the properties of the chosen expansion basis. Obstacles are especially likely to arise when working with geometrically complex domains (e.g., not "boxes" or similarly simple geometries), because a suitable expansion basis may not be available analytically. For polytopic domains that can be "triangulated", or covered exactly using a mesh of elements with simple geometries, it may be possible to rigorously bound the difference between any function defined on the domain and a piecewise-polynomial approximation on the mesh. In fact, estimates of this kind may already be available because similar ideas lie at the core of convergence analysis for finite-element methods. To derive inner approximations of integral inequalities on non-polytopic domains, however, a fundamentally different strategy may be needed.

From the point of view of numerical implementation, instead, one should address the challenges posed by the degradation of numerical conditioning and by the considerable computational resources currently required by state-of-the-art software packages to solve large SDPs accurately. Given the growing interest in SDPs across a wide range of disciplines (fluid mechanics, control, operations research, machine learning, and artificial intelligence just to name a few) it is expected that more robust and more efficient algorithms for large-scale SDPs will become available in the near future. In the meantime, it remains imperative to reduce the size of the SDP approximation of a given background field problem as much as possible, by taking advantage of any available special structures or symmetries. It also seems fundamental to try to bridge the gap between the formulation of the SDP on the one hand, and its numerical solution on the other. That this can significantly improve performance has already been demonstrated in chapter 6, where the spectral constraint was discretised using piecewise-linear approximations so as to obtain LMIs with chordal sparsity patterns. Without doubt, the question of how an optimal background field problem can be

reformulated into an SDP with chordal sparsity or other similar "computationally friendly" structural properties should be addressed by future research.

Finally, it has recently become apparent that the background method is a particular instance of a more general convex framework to bound time-averaged properties of dynamical systems (Chernyshenko *et al.*, 2014; Fantuzzi *et al.*, 2016; Chernyshenko, 2017; Goluskin, 2017; Tobasco *et al.*, 2018). Just like the construction of a background field satisfying one or more spectral constraints lies at the heart of the background method, this more general framework is centred around the construction of a so-called *auxiliary functional*, subject to a certain inequality condition that implies the desired bound. Interestingly, for many systems governed by ODEs (Tobasco *et al.*, 2018) and perhaps also by PDEs, there exist auxiliary functionals that yield sharp bounds, meaning that there are solutions of the governing equations for which the bounds are exact. An even more exciting observation is that, just as demonstrated in this work for the background method, near-optimal auxiliary functionals can be searched for through the solution of SDPs—albeit often very large ones. For this reason, the methods and results presented in this thesis can be considered a stepping stone to the construction of such near-optimal auxiliary functionals. Naturally, it remains to be seen exactly how far these ideas can be pushed, and whether complex systems of physical or engineering relevance can be studied successfully without the investment of considerable time and computing power. The hope, however, is that the insights gained by optimising background fields using SDPs will help to guide the future development of powerful, computer-assisted methods based on auxiliary functionals, which will enable accurate and inexpensive quantitative analysis of complex systems across a wide range of fields and applications.

Bibliography

Agarwal, R. P. and O'Regan, D. (2009), *Ordinary and Partial Differential Equations — With Special Functions, Fourier Series and Boundary Value Problems*, Universitext, 1st edn, Springer, New York. Available from: doi:10.1007/978-0-387-79146-3.

Agler, J., Helton, J. W., McCullough, S. and Rodman, L. (1988) Positive definite matrices with a given sparsity pattern. *Linear Algebra and its Applications* **107**, 101–149. Available from: doi:10.1016/0024-3795(88)90240-6.

Ahmadi, M., Valmorbida, G. and Papachristodoulou, A. (2014) Input-output analysis of distributed parameter systems using convex optimization. In: *Proceedings of the 53rd IEEE Conference on Decision and Control*, Los Angeles, CA, IEEE. pp. 4310–4315. Available from: doi:10.1109/CDC.2014.7040061.

Ahmadi, M., Valmorbida, G. and Papachristodoulou, A. (2016) Dissipation inequalities for the analysis of a class of PDEs. *Automatica* **66**, 163–171. Available from: doi:10.1016/j.automatica.2015.12.010.

Andersen, E. D., Jensen, B., Jensen, J., Sandvik, R. and Worsøe, U. (2009) *MOSEK version 6*, MOSEK ApS, Copenhagen. Report number: TR-2009-3. Available from: `http://docs.mosek.com/whitepapers/mosek6.pdf` [Accessed: 01 May 2015].

Bertsimas, D. and Caramanis, C. (2006) Bounds on linear PDEs via semidefinite optimization. *Mathematical Programming Series A* **108**(1), 135–158. Available from: doi:10.1007/s10107-006-0702-z.

Biggin, A. J., Steinberger, B., Aubert, J., Suttie, N., Holme, R., Torsvik, T. H., van der Meer, D. G. and van Hinsbergen, D. J. J. (2012) Possible links between long-term geomagnetic variations and whole-mantle convection processes. *Nature Geoscience* **5**, 526–533. Available from: doi:10.1038/ngeo1558.

Blair, J. R. S. and Peyton, B. (1993), An introduction to chordal graphs and clique trees, In: George, A., Gilbert, J. R. and Liu, J. W. H. (eds), *Graph Theory and Sparse Matrix Computation*, vol. 56 of *The IMA Volumes in Mathematics and its Applications*, Springer, New York, USA, pp. 1–26. Available from: doi:10.1007/978-1-4613-8369-7_1.

Boeck, T. (2005) Bénard–Marangoni convection at large Marangoni numbers: Results of numerical simulations. *Advances in Space Research* **36**(1), 4–10. Available from: doi:10.1016/j.asr.2005.02.083.

Boeck, T. and Thess, A. (1998) Turbulent Bénard–Marangoni convection: Results of two-dimensional simulations. *Physical Review Letters* **80**(6), 1216–1219. Available from: doi:10.1103/PhysRevLett.80.1216.

Boeck, T. and Thess, A. (2001) Power-law scaling in Bénard–Marangoni convection at large Prandtl numbers. *Physical Review E* **64**(2), 027303. Available from: doi:10.1103/PhysRevE.64.027303.

Boyd, S., El Ghaoui, L., Feron, E. and Balakrishnan, V. (1994), *Linear Matrix Inequalities in System and Control Theory*, vol. 15 of *SIAM Studies in Applied and Numerical Mathematics*, SIAM, Philadelphia. Available from: doi:10.1137/1.9781611970777.

Boyd, S. and Vandenberghe, L. (2004), *Convex Optimization*, 1st edn, Cambridge University Press, Cambridge. Available from: `https://doi.org/10.1017/CBO9780511804441`.

Brezis, H. (2010), *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Universitext, 1 edn, Springer–Verlag, New York. Available from: doi:10.1007/978-0-387-70914-7.

Bronski, J. C. and Gambill, T. N. (2006) Uncertainty estimates and L_2 bounds for the Kuramoto–Sivashinsky equation. *Nonlinearity* **19**(9), 2023–2039. Available from: doi:10.1088/0951-7715/19/9/002.

Burer, S. and Choi, C. (2006) Computational enhancements in low-rank semidefinite programming. *Optimization Methods and Software* **21**(3), 493–512. Available from: doi:10.1080/10556780500286582.

Burer, S. and Monteiro, R. D. C. (2003) A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming, Series B* **95**(2), 329–357. Available from: doi:10.1007/s10107-002-0352-8.

Burer, S. and Monteiro, R. D. C. (2005) Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming, Series A* **103**(3), 427–444. Available from: doi:10.1007/s10107-004-0564-1.

Burer, S., Monteiro, R. D. C. and Zhang, Y. (2002) Solving a class of semidefinite programs via nonlinear programming. *Mathematical Programming, Series A* **93**(1), 97–122. Available from: doi:10.1007/s101070100279.

Busse, F. H. (1970) Bounds for turbulent shear flow. *Journal of Fluid Mechanics* **41**(1), 219–240. Available from: doi:10.1017/S0022112070000599.

Busse, F. H. (1979) The optimum theory of turbulence. *Advances in Applied Mechanics* **18**, 77–121. Available from: doi:10.1016/S0065-2156(08)70265-5.

Carlson, J., Jaffe, A. and Wiles, A. (eds) (2006), *The Millennium Prize Problems*, American Mathematical Society, Providence, R.I. (USA). Available from: `http://www.claymath.org/library/monographs/MPPc.pdf`.

Chan, S.-K. (1971) Infinite Prandtl number turbulent convection. *Studies in Applied Mathematics* **50**(1), 13–49. Available from: doi:10.1002/sapm197150113.

Chernyshenko, S. I. (2017) Relationship between the methods of bounding time averages. arXiv:1704.02475 [physics.flu-dyn]. Available from: `https://arxiv.com/abs/1704.02475` [Accessed: 23 Aug 2014].

Chernyshenko, S. I., Goulart, P. J., Huang, D. and Papachristodoulou, A. (2014) Polynomial sum of squares in fluid dynamics: a review with a look ahead. *Philosophical Transactions of the Royal Society A* **372**(2020), 20130350. Available from: doi:10.1098/rsta.2013.0350.

Collet, P., Eckmann, J.-P., Epstein, H. and Stubbe, J. (1993) A Global Attracting Set for the Kuramoto-Sivashinsky Equation. *Communications in Mathematical Physics* **152**(1), 203–214. Available from: doi:10.1007/BF02097064.

Constantin, P. and Doering, C. R. (1995*a*) Variational bounds in dissipative systems. *Physica D: Nonlinear Phenomena* **82**(3), 221–228. Available from: doi:10.1016/0167-2789(94)00237-K.

Constantin, P. and Doering, C. R. (1995*b*) Variational bounds on energy dissipation in incompressible flows. II. Channel flow. *Physical Review E* **51**(4), 3192–3198. Available from: doi:10.1103/PhysRevE.51.3192.

Constantin, P. and Doering, C. R. (1996) Heat transfer in convective turbulence. *Nonlinearity* **9**(4), 1049–1060. Available from: doi:10.1088/0951-7715/9/4/013.

Constantin, P. and Doering, C. R. (1999) Infinite Prandtl number convection. *Journal of Statistical Physics* **94**(1–2), 159–172. Available from: doi:10.1023/A:1004511312885.

Courant, R. and Hilbert, D. (1953), *Methods of Mathematical Physics*, vol. 1, 1st edn, Interscience Publisher Inc., New York.

Cushman-Roisin, B. and Beckers, J.-M. (2011), *Introduction to Geophysical Fluid Dynamics — Physical and Numerical Aspects*, vol. 101 of *International Geophysics Series*, 2nd edn, Academic Press.

de Bruyn, J. R., Bodenschatz, E., Morris, S. W., Trainoff, S. P., Hu, Y., Cannell, D. S. and Ahlers, G. (1996) Apparatus for the study of Rayleigh–Bénard convection in gases under pressure. *Review of Scientific Instruments* **67**(6), 2043–2067. Available from: doi:10.1063/1.1147511.

DebRoy, T. and David, S. A. (1995) Physical processes in fusion welding. *Reviews of Modern Physics* **67**(1), 85–112. Available from: doi:10.1103/RevModPhys.67.85.

Doering, C. R. and Constantin, P. (1992) Energy dissipation in shear driven turbulence. *Physical Review Letters* **69**(11), 1648–1651. Available from: doi:10.1103/PhysRevLett.69.1648.

Doering, C. R. and Constantin, P. (1994) Variational bounds on energy dissipation in incompressible flows: Shear flow. *Physical Review E* **49**(5), 4087–4099. Available from: doi:10.1103/PhysRevE.49.4087.

Doering, C. R. and Constantin, P. (1996) Variational bounds on energy dissipation in incompressible flows. III. Convection. *Physical Review E* **53**(6), 5957–5981. Available from: doi:10.1103/PhysRevE.53.5957.

Doering, C. R. and Constantin, P. (1998) Bounds for heat transport in a porous layer. *Journal of Fluid Mechanics* **376**, 263–296. Available from: doi:10.1017/S002211209800281X.

Doering, C. R. and Constantin, P. (2001) On upper bounds for infinite Prandtl number convection with or without rotation. *Journal of Mathematical Physics* **42**(2), 784–795. Available from: doi:10.1063/1.1336157.

Doering, C. R. and Gibbon, J. D. (1995), *Applied analysis of the Navier–Stokes equations*, vol. 1 of *Cambridge Texts in Applied Mathematics*, 1st edn, Cambridge University Press, Cambridge. Available from: doi:10.1017/CBO9780511608803.

Doering, C. R. and Hyman, J. M. (1997) Energy stability bounds on convective heat transport: Numerical study. *Physical Review E* **55**(6), 7775–7778. Available from: doi:10.1103/PhysRevE.55.7775.

Doering, C. R., Otto, F. and Reznikoff, M. G. (2006) Bounds on vertical heat transport for infinite Prandtl number Rayleigh–Bénard convection. *Journal of Fluid Mechanics* **560**, 229–241. Available from: doi:10.1017/S0022112006000097.

Dougall, J. (1953) The product of two Legendre polynomials. *Proceedings of the Glasgow Mathematical Association* **1**(3), 121–125. Available from: doi:10.1017/S2040618500035590.

Eckert, K. and Thess, A. (2006), Secondary instabilities in surface-tension-driven Bénard–Marangoni convection, In: Mutabazi, I., Wesfreid, J. E. and Guyon, E. (eds), *Dynamics of spatio-temporal cellular structures*, Springer tracts in modern physics, Springer New York, pp. 163–176. Available from: doi:10.1007/978-0-387-25111-0_9.

Evans, L. C. (2010), *Partial Differential Equations*, vol. 19 of *Graduate Studies in Mathematics*, 1st edn, American Mathematical Society, Providence, R.I. (USA).

Everitt, W. N. (1957) The Sturm–Liouville problem for fourth-order differential equations. *The Quarterly Journal of Mathematics* **8**(1), 146–160. Available from: doi:10.1093/qmath/8.1.146.

Fantuzzi, G., Goluskin, D., Huang, D. and Chernyshenko, S. I. (2016) Bounds for deterministic and stochastic dynamical systems using sum-of-squares optimization. *SIAM Journal on Applied Dynamical Systems* **15**(4), 1962–1988. Available from: doi:10.1137/15M1053347.

Fantuzzi, G. and Wynn, A. (2015) Construction of an optimal background profile for the Kuramoto–Sivashinsky equation using semidefinite programming. *Physics Letters A* **379**(1–2), 23–32. Available from: doi:10.1016/j.physleta.2014.10.039.

Fantuzzi, G. and Wynn, A. (2017) Exact energy stability of Bénard–Marangoni convection at infinite Prandtl number. *Journal of Fluid Mechanics* **822**, R1. Available from: doi:10.1017/jfm.2017.323.

Fujisawa, K., Kim, S., Kojima, M., Okamoto, Y. and Yamashita, M. (2009) *User's manual for SparseCoLO: conversion methods for SPARSE COnic-form Linear Optimization problems*, Deptartment of Mathematical and Computing Sciences, Tokyo Institute of Technology, Tokyo, Japan. Report number: B-453. Available from: `http://www.optimization-online.org/DB_HTML/2009/02/2234.html` [Accessed: 11 Dec 2015].

Fukuda, M., Kojima, M., Murota, K. and Nakata, K. (2000) Exploiting sparsity in semidefinite programming via matrix completion I: General framework. *SIAM J. Optim.* **11**(3), 647–674. Available from: doi:10.1137/S1052623400366218.

Gambill, T. N. (2006) Application of uncertainty inequalities to bound the radius of the attractor for the Kuramoto-Sivashinsky equation. PhD thesis. University of Illinois at Urbana-Champaign. Available from: `https://www.ideals.illinois.edu/handle/2142/86875` [Accessed: 25 Sep 2014].

Giacomelli, L. and Otto, F. (2005) New bounds for the KuramotoâĂŘ-Sivashinsky equation. *Communications on Pure and Applied Mathematics* **58**(3), 297–318. Available from: doi:10.1002/cpa.20031.

Giaquinta, M. and Hildebrandt, S. (1996), *Calculus of Variations I*, vol. 310 of *Grundlehren der mathematischen Wissenschaften*, 2nd edn, Springer–Verlag, Berlin. Available from: doi:10.1007/978-3-662-03278-7.

Goluskin, D. (2015) Internally heated convection beneath a poor conductor. *Journal of Fluid Mechanics* **771**, 36–56. Available from: doi:10.1017/jfm.2015.140.

Goluskin, D. (2016), *Internally heated convection and Rayleigh–Bénard convection*, Springer Briefs in Thermal Engineering and Applied Science, 1st edn, Springer International Publishing, Cham, Switzerland. Available from: doi:10.1007/978-3-319-23941-5.

Goluskin, D. (2017) Bounding averages rigorously using semidefinite programming: mean moments of the lorenz system. *Journal of Nonlinear Science* (in press). Available from: doi:10.1007/s00332-017-9421-2.

Goluskin, D. and Doering, C. R. (2016) Bounds for convection between rough boundaries. *Journal of Fluid Mechanics* **804**, 370–386. Available from: doi:10.1017/jfm.2016.528.

Goodman, J. (1994) Stability of the Kuramoto–Sivashinsky and related systems. *Communications on Pure and Applied Mathematics* **47**(3), 293–306. Available from: doi:10.1002/cpa.3160470304.

Hagstrom, G. I. and Doering, C. R. (2010) Bounds on heat transport in Bénard–Marangoni convection. *Physical Review E* **81**(4), 047301. Available from: doi:10.1103/PhysRevE.81.047301.

Hagstrom, G. I. and Doering, C. R. (2014) Bounds on surface stress-driven shear flow. *Journal of Nonlinear Science* **24**(1), 185–199. Available from: doi:10.1007/s00332-013-9183-4.

Hassanzadeh, P., Chini, G. P. and Doering, C. R. (2014) Wall to wall optimal transport. *Journal of Fluid Mechanics* **751**, 627–662. Available from: doi:10.1017/jfm.2014.306.

Howard, L. N. (1963) Heat transport by turbulent convection. *Journal of Fluid Mechanics* **17**(3), 405–432. Available from: doi:10.1017/S0022112063001427.

Howard, L. N. (1972) Bounds on Flow Quantities. *Annual Review of Fluid Mechanics* **4**(1), 473–494. Available from: doi:10.1146/annurev.fl.04.010172.002353.

Hyman, J. M. and Nicolaenko, B. (1986) The Kuramoto–Sivashinsky equation: a bridge between PDE's and dynamical systems. *Physica D: Nonlinear Phenomena* **18**(1-3), 113–126. Available from: doi:10.1016/0167-2789(86)90166-1.

Hyman, J. M., Nicolaenko, B. and Zaleski, S. (1986) Order and complexity in the Kuramoto–Sivashinsky model of weakly turbuent interfaces. *Physica D: Nonlinear Phenomena* **23**(1-3), 265–292. Available from: doi:10.1016/0167-2789(86)90136-3.

Ierley, G. R., Kerswell, R. R. and Plasting, S. C. (2006) Infinite-Prandtl-number convection. Part 2. A singular limit of upper bound theory. *Journal of Fluid Mechanics* **560**, 159–227. Available from: doi:10.1017/S0022112006000450.

Jackson, D. (1930), *The Theory of Approximation*, vol. 11 of *American Mathematical Society Colloquium Publications*, American Mathematical Society, New York. Available from: doi:10.1002/zamm.19310110117.

Jones, G. M. (1977) Thermal interaction of the core and the mantle and long-term behavior of the geomagnetic field. *Journal of Geophysical Research* **82**(11), 1703–1709. Available from: doi:10.1029/JB082i011p01703.

Kakimura, N. (2010) A direct proof for the matrix decomposition of chordal-structured positive semidefinite matrices. *Linear Algebra and its Applications* **433**(4), 819–823. Available from: doi:10.1016/j.laa.2010.04.012.

Kerswell, R. R. (1997) Variational bounds on shear-driven turbulence and turbulent Boussinesq convection. *Physica D: Nonlinear Phenomena* **100**(3–4), 355–376. Available from: doi:10.1016/S0167-2789(96)00227-8.

Kerswell, R. R. (1998) Unification of variational principles for turbulent shear flows: the background method of Doering-Constantin and the mean-fluctuation formulation of Howard-Busse. *Physica D: Nonlinear Phenomena* **121**(1–2), 175–192. Available from: doi:10.1016/S0167-2789(98)00104-3.

Kerswell, R. R. (1999) Variational principle for the Navier-Stokes equations. *Physical Review E* **59**(5), 5482–5494. Available from: doi:10.1103/PhysRevE.59.5482.

Kerswell, R. R. (2001) New results in the variational approach to turbulent Boussinesq convection. *Physics of Fluids* **13**(1), 192–209. Available from: doi:10.1063/1.1327295.

Kim, S., Kojima, M., Mevissen, M. and Yamashita, M. (2011) Exploiting sparsity in linear and nonlinear matrix inequalities via positive semidefinite matrix completion. *Math. Program. Ser. B* **129**(1), 33–68. Available from: doi:10.1007/s10107-010-0402-6.

Kumar, A. and Roy, S. (2009) Effect of three-dimensional melt pool convection on process characteristics during laser cladding. *Computational Materials Science* **46**(2), 495–506. Available from: doi:10.1016/j.commatsci.2009.04.002.

Kuramoto, Y. and Tsuzuki, T. (1975) On the formation of dissipative structures in reaction-diffusion systems: Reductive perturbation approach. *Progress of Theoretical Physics* **54**(3), 687–699. Available from: doi:10.1143/PTP.54.687.

Kuramoto, Y. and Tsuzuki, T. (1976) Persistent propagation of concentration waves in dissipative media far from thermal equilibrium. *Progress of Theoretical Physics* **55**(2), 356–369. Available from: doi:10.1143/PTP.55.356.

Lappa, M. (2010), *Thermal convection: patterns, evolution and stability*, Wiley. Available from: doi:10.1002/9780470749982.

Löfberg, J. (2004) YALMIP: A toolbox for modeling and optimization in MATLAB. In: *IEEE International Symposium on Computer Aided Control Systems Design*, Taipei, IEEE. pp. 284–289. Available from: doi:10.1109/CACSD.2004.1393890.

Löfberg, J. (2009) Pre- and post-processing sum-of-squares programs in practice. *IEEE Transactions on Automatic Control* **54**(5), 1007–1011. Available from: doi:10.1109/TAC.2009.2017144.

Madani, R., Kalbat, A. and Lavaei, J. (2015) ADMM for sparse semidefinite programming with applications to optimal power flow problem. In: *Proceedings of the 54th IEEE Conference on Decision and Control*, Osaka, Japan, IEEE. pp. 5932–5939. Available from: doi:10.1109/CDC.2015.7403152.

Malkus, W. V. R. (1954) The Heat Transport and Spectrum of Thermal Turbulence. *Proceedings of the Royal Society A* **225**(1161), 196. Available from: doi:10.1098/rspa.1954.0197.

Marchioro, C. (1994) Remark on the energy dissipation in shear driven turbulence. *Physica D: Nonlinear Phenomena* **74**(3–4), 395–398. Available from: doi:10.1016/0167-2789(94)90203-8.

Michelson, D. M. and Sivashinsky, G. I. (1977) Nonlinear analysis of hydrodynamic instability in laminar flames—II. Numerical experiments. *Acta Astronautica* **4**(11–12), 1207–1221. Available from: doi:10.1016/0094-5765(77)90097-2.

Molinet, L. (2000) Local dissipativity in L^2 for the Kuramoto–Sivashinsky equation in spatial dimension 2. *Journal of Dynamics and Differential Equations* **12**(3), 533–556. Available from: doi:10.1023/A:1026459527446.

Nakata, K., Fujisawa, K., Fukuda, M., Kojima, M. and Murota, K. (2003) Exploiting sparsity in semidefinite programming via matrix completion II: Implementation and numerical results. *Math. Program. Ser. B* **95**(2), 303–327. Available from: doi:10.1007/s10107-002-0351-9.

Nicodemus, R., Grossmann, S. and Holthaus, M. (1997*a*) Improved variational principle for bounds on energy dissipation in turbulent shear flow. *Physica D: Nonlinear Phenomena* **101**(1–2), 178–190. Available from: doi:10.1016/S0167-2789(96)00210-2.

Nicodemus, R., Grossmann, S. and Holthaus, M. (1997*b*) Variational bound on energy dissipation in plane Couette flow. *Physical Review E* **56**(6), 6774–6786. Available from: doi:10.1103/PhysRevE.56.6774.

Nicodemus, R., Grossmann, S. and Holthaus, M. (1998) The background flow method. Part 1. Constructive approach to bounds on energy dissipation. *Journal of Fluid Mechanics* **363**, 281–300. Available from: doi:10.1017/S0022112098001165.

Nicolaenko, B., Scheurer, B. and Temam, R. (1985) Some global dynamical properties of the Kuramoto–Sivashinsky equations: Nonlinear stability and attractors. *Physica D: Nonlinear Phenomena* **16**(2), 155–183. Available from: doi:10.1016/0167-2789(85)90056-9.

O'Donoghue, B., Chu, E., Parikh, N. and Boyd, S. (2016) Conic Optimization via Operator Splitting and Homogeneous Self-Dual Embedding. *Journal of Optimization Theory and Applications* **169**(3), 1–27. Available from: doi:10.1007/s10957-016-0892-3.

Otero, J. (2002) Bounds for the heat transport in turbulent convection. PhD thesis. University of Michigan. Available from: https://deepblue.lib.umich.edu/handle/2027.42/132721 [Accessed: 07 Feb 2017].

Otero, J., Dontcheva, L. A., Johnston, H., Worthing, R. A., Kurganov, A., Petrova, G. and Doering, C. R. (2004) High-Rayleigh-number convection in a fluid-saturated porous layer. *Journal of Fluid Mechanics* **500**, 263–281. Available from: doi:10.1017/S0022112003007298.

Otero, J., Wittenberg, R. W., Worthing, R. A. and Doering, C. R. (2002) Bounds on Rayleigh–Bénard convection with an imposed heat flux. *Journal of Fluid Mechanics* **473**, 191–199. Available from: doi:10.1017/S0022112002002410.

Otto, F. and Seis, C. (2011) Rayleigh–Bénard convection: Improved bounds on the Nusselt number. *Journal of Mathematical Physics* **52**(8), 083702. Available from: doi:10.1063/1.3623417.

Pakazad, S. K., Hansson, A., Andersen, M. S. and Rantzer, A. (2017) Distributed semidefinite programming with application to large-scale system analysis. *IEEE Transactions on Automatic Control* (in press). Available from: doi:10.1109/TAC.2017.2739644.

Papachristodoulou, A. and Peet, M. M. (2006) On the Analysis of Systems Described by Classes of Partial Differential Equations. In: *Proceedings of the 45th IEEE Conference on Decision and Control*, San Diego, CA, USA, IEEE. pp. 747–752. Available from: doi:10.1109/CDC.2006.377815.

Parrilo, P. A. (2003) Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming Series B* **96**(2), 293–320. Available from: doi:10.1007/s10107-003-0387-5.

Parrilo, P. A. (2013), Semidefinite optimization, In: Blekherman, G., Parrilo, P. A. and Thomas, R. R. (eds), *Semidefinite optimization and convex algebraic geometry*, MOS-SIAM Series on Optimization, SIAM, pp. 3–46. Available from: doi:10.1137/1.9781611972290.ch2.

Patberg, W. B., Koers, A., Steenge, W. D. E. and Drinkenburg, A. A. H. (1983) Effectiveness of mass transfer in a packed distillation column in relation to surface tension gradients. *Chemical Engineering Science* **38**(6), 917–923. Available from: doi:10.1016/0009-2509(83)80013-X.

Pearson, J. R. A. (1958) On convection cells induced by surface tension. *Journal of Fluid Mechanics* **4**(5), 489–500. Available from: doi:10..1017/S0022112058000616.

Peet, M. M. and Bliman, P.-A. (2007) An Extension of the Weierstrass Theorem to Linear Varieties: Application to Delayed Systems. In: *7th IFAC Workshop on Time-Delay Systems*, Nantes, France. pp. 1–4. Available from: https://asu.pure.elsevier.com/en/publications/an-extension-of-the-weierstrass-theorem-to-linear-varieties-appli [Accessed: 05 Jun 2016].

Plasting, S. C. (2004) Turbulence has its limits: *a priori* estimates of transport properties in turbulent fluid flows. PhD thesis. University of Bristol. Available from: http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.399391 [Accessed: 07 Feb 2017].

Plasting, S. C. and Ierley, G. R. (2005) Infinite-Prandtl-number convection. Part 1. Conservative bounds. *Journal of Fluid Mechanics* **542**(2005), 343–363. Available from: doi:10.1017/S0022112005006555.

Plasting, S. C. and Kerswell, R. R. (2003) Improved upper bound on the energy dissipation rate in plane Couette flow: the full solution to Busse's problem and the Constantin-Doering-Hopf problem with one-dimensional background field. *Journal of Fluid Mechanics* **477**, 363–379. Available from: doi:10.1017/S0022112002003361.

Pope, S. B. (2000), *Turbulent Flows*, Cambridge University Press, Cambridge.

Pumir, A. and Blumenfeld, L. (1996) Heat transport in a liquid layer locally heated on its free surface. *Physical Review E* **54**(5), R4528–R4531. Available from: doi:10.1103/PhysRevE.54.R4528.

Robinson, J. C. (2001), *Infinite-dimensional dynamical systems: an introduction to dissipative parabolic PDEs and the theory of global attractors*, Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge.

Schatz, M. F. and Neitzel, G. P. (2001) Experiments on thermocapillary instabilities. *Annual Review of Fluid Mechanics* **33**, 93–127. Available from: doi:10.1146/annurev.fluid.33.1.93.

Sivashinsky, G. I. (1977) Nonlinear analysis of hydrodynamic instability in laminar flames—I. Derivation of basic equations. *Acta Astronautica* **4**(11–12), 1177–1206. Available from: doi:10.1016/0094-5765(77)90096-0.

Sivashinsky, G. I. (1980) On Flame Propagation Under Conditions Of Stoichiometry. *SIAM Journal on Applied Mathematics* **39**(1), 67–82. Available from: doi:10.1137/0139007.

Sivashinsky, G. I. and Michelson, D. M. (1980) On irregular wavy flow of a liquid film down a vertical plane. *Progress of Theoretical Physics* **63**(6), 2112–2114. Available from: doi:10.1143/PTP.63.2112.

Straughan, B. (2004), *The Energy Method, Stability, and Nonlinear Convection*, vol. 91 of *Applied Mathematical Sciences*, 2nd edn, Springer-Verlag, New York. Available from: doi:10.1007/978-0-387-21740-6.

Sturm, J. F. (1999) Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization methods and software* **11**(1-4), 625–653. Available from: doi:10.1080/10556789908805766.

Sturm, J. F. (2002) Implementation of interior point methods for mixed semidefinite and second order cone optimization problems. *Optimization Methods and Software* **17**(6), 1105–1154. Available from: doi:10.1080/1055678021000045123.

Sun, Y., Andersen, M. S. and Vandenberghe, L. (2014) Decomposition in conic optimization with partially separable structure. *SIAM Journal on Optimization* **24**(2), 873–897. Available from: doi:10.1137/130926924.

Tan, W. and Packard, A. (2006) Stability region analysis using sum of squares programming. In: *Proceedings of the American Control Conference*, Minneapolis, MN, USA, June 14-16, IEEE. pp. 2297–2302. Available from: doi:10.1109/ACC.2006.1656562.

Tang, W., Caulfield, C. P. and Young, W. R. (2004) Bounds on dissipation in stress-driven flow. *Journal of Fluid Mechanics* **510**, 333–352. Available from: doi:10.1017/S0022112004009589.

Tobasco, I. and Doering, C. R. (2017) Optimal wall-to-wall transport by incompressible flows. *Physical Review Letters* **118**(26), 264502. Available from: doi:10.1103/PhysRevLett.118.264502.

Tobasco, I., Goluskin, D. and Doering, C. R. (2018) Optimal bounds and extremal trajectories for time averages in nonlinear dynamical systems. *Physics Letters A* **382**(6), 382–386. Available from: doi:10.1016/j.physleta.2017.12.023.

Toh, K.-C., Todd, M. J. and Tütüncü, R. H. (1999) SDPT3—a MATLAB software package for semidefinite programming, version 1.3. *Optimization Methods and Software* **11**(1), 545–581. Available from: doi:10.1080/10556789908805762.

Tütüncü, R. H., Toh, K.-C. and Todd, M. J. (2003) Solving semidefinite-quadratic-linear programs using SDPT3. *Mathematical Programming Series B* **95**(2), 189–217. Available from: doi:10.1007/s10107-002-0347-5.

Valmorbida, G., Ahmadi, M. and Papachristodoulou, A. (2014*a*) Semi-definite programming and functional inequalities for distributed parameter systems. In: *Proceedings of the 53rd IEEE Conference on Decision and Control*, Los Angeles, CA, USA, IEEE. pp. 4304–4309. Available from: doi:10.1109/CDC.2014.7040060.

Valmorbida, G., Ahmadi, M. and Papachristodoulou, A. (2014*b*) Semi-definite programming and functional inequalities for distributed parameter systems. arXiv:1403.6882 [mathOC]. Available from: `https://arxiv.com/abs/1403.6882` [Accessed: 07 Jun 2014]. (Extended version of Ref. Valmorbida *et al.* (2014*a*)).

Valmorbida, G., Ahmadi, M. and Papachristodoulou, A. (2015) Convex Solutions to Integral Inequalities in Two-Dimensional Domains. In: *IEEE 54th Annual Conference on Decision and Control*, Osaka, Japan, IEEE. pp. 7268–7273. Available from: doi:10.1109/CDC.2015.7403366.

Valmorbida, G., Ahmadi, M. and Papachristodoulou, A. (2016) Stability Analysis for a Class of Partial Differential Equations via Semidefinite Programming. *IEEE Transactions on Automatic Control* **61**(6), 1649–1654. Available from: doi:10.1109/TAC.2015.2479135.

Vandenberghe, L. and Boyd, S. (1996) Semidefinite Programming. *SIAM Review* **38**(1), 49–95. Available from: doi:10.1137/1038003.

Wen, B., Chini, G. P., Dianati, N. and Doering, C. R. (2013) Computational approaches to aspect-ratio-dependent upper bounds and heat flux in porous medium convection. *Physics Letters A* **377**(41), 2931–2938. Available from: doi:10.1016/j.physleta.2013.09.009.

Wen, B., Chini, G. P., Kerswell, R. R. and Doering, C. R. (2015) Time-stepping approach for solving upper-bound problems: Application to two-dimensional Rayleigh–Bénard convection. *Physical Review E* **92**(4), 043012. Available from: doi:10.1103/PhysRevE.92.043012.

Wen, Z., Goldfarb, D. and Yin, W. (2010) Alternating direction augmented Lagrangian methods for semidefinite programming. *Mathematical Programming Computation* **2**(3-4), 203–230. Available from: doi:10.1007/s12532-010-0017-1.

Whitehead, J. P. and Doering, C. R. (2011) Ultimate State of Two-Dimensional Rayleigh–Bénard Convection between Free-Slip Fixed-Temperature Boundaries. *Physical Review Letters* **106**(24), 244501. Available from: doi:10.1103/PhysRevLett.106.244501.

Whitehead, J. P. and Doering, C. R. (2012) Rigid bounds on heat transport by a fluid between slippery boundaries. *Journal of Fluid Mechanics* **707**, 241–259. Available from: doi:10.1017/jfm.2012.274.

Whitehead, J. P. and Wittenberg, R. W. (2014) A rigorous bound on the vertical transport of heat in Rayleigh–Bénard convection at infinite Prandtl number with mixed thermal boundary conditions. *Journal of Mathematical Physics* **55**(9), 093104. Available from: doi:10.1063/1.4896223.

Wittenberg, R. W. (2002) Dissipativity, analyticity and viscous shocks in the (de) stabilized Kuramoto–Sivashinsky equation. *Physics Letters A* **300**(4–5), 407–416. Available from: doi:10.1016/S0375-9601(02)00861-7.

Wittenberg, R. W. (2010) Bounds on Rayleigh–Bénard convection with imperfectly conducting plates. *Journal of Fluid Mechanics* **665**, 158–198. Available from: doi:10.1017/S0022112010003897.

Wittenberg, R. W. and Gao, J. (2010) Conservative bounds on Rayleigh–Bénard convection with mixed thermal boundary conditions. *European Physical Journal B* **76**(4), 565–580. Available from: doi:10.1140/epjb/e2010-00227-x.

Yan, X. (2004) On limits to convective heat transport at infinite Prandtl number with or without rotation. *Journal of Mathematical Physics* **45**(7), 2718–2743. Available from: doi:10.1063/1.1763246.

Yiantsios, S. G., Serpetsi, S. K., Doumenc, F. and Guerrier, B. (2015) Surface deformation and film corrugation during drying of polymer solutions induced by Marangoni phenomena. *International Journal of Heat and Mass Transfer* **89**, 1083–1094. Available from: doi:10.1016/j.ijheatmasstransfer.2015.06.015.

Zeidler, E. (1995), *Applied Functional Analysis—Applications to Mathematical Physics*, vol. 108 of *Applied Mathematical Sciences*, 1st edn, Springer, New York. Available from: doi:10.1007/978-1-4612-0815-0.

Zheng, Y., Fantuzzi, G., Papachristodoulou, A., Goulart, P. J. and Wynn, A. (2017*a*) Fast ADMM for Semidefinite Programs with Chordal Sparsity. In: *Proceedings of the 2017 American Control Conference*, Seattle, WA, USA, IEEE. pp. 3335–3340. Available from: doi:10.23919/ACC.2017.7963462.

Zheng, Y., Fantuzzi, G., Papachristodoulou, A., Goulart, P. and Wynn, A. (2017*b*) Fast ADMM for homogeneous self-dual embedding of sparse SDPs. *IFAC-PapersOnLine* **50**(1), 8411–8416. Available from: doi:10.1016/j.ifacol.2017.08.1569.

Zuiderweg, F. J. and Harmens, A. (1958) The influence of surface phenomena on the performance of distillation columns. *Chemical Engineering Science* **9**(2-3), 89–103. Available from: doi:http://dx.doi.org/10.1016/0009-2509(58)80001-9.

Appendix A

Miscellaneous proofs

A.1 Proof of Theorem 3.1

The proof of Theorem 3.1 presented here is an adaptation of arguments presented in chapter 6 of the book by Evans (2010) in the context of second-order elliptic PDEs.

Let u be any sufficiently smooth, odd and periodic function on $[-\ell, \ell]$ such that $Du = \partial^4 u + \partial^2 u + fu$ is well defined. For any $v \in H_{\mathrm{p,o}}$—where $H_{\mathrm{p,o}}$ is the space of square-integrable, odd, and periodic functions on $(-\ell, \ell)$ with two square-integrable periodic derivatives defined in (3.6)—integration by parts implies

$$\int v\, Du \,\mathrm{d}x = \int u''\, v'' - u'\, v' + f\, u\, v \,\mathrm{d}x =: \mathcal{Q}_0\{u, v\}. \tag{A.1}$$

As in chapter 3, here and for the rest of this section all integrals are over the interval $(-\ell, \ell)$.

The symmetric bilinear form $\mathcal{Q}_0\{u, v\}$ is well defined for all $u, v \in H_{\mathrm{p,o}}$, and $u \in H_{\mathrm{p,o}}$ is a *weak solution* of the eigenvalue problem (3.17) with eigenvalue σ if

$$\mathcal{Q}_0\{u, v\} = \sigma \int u\, v \,\mathrm{d}x \quad \forall v \in H_{\mathrm{p,o}}. \tag{A.2}$$

Clearly, if u satisfies (A.2), then for any $\lambda \in \mathbb{R}$ it also satisfies

$$\mathcal{Q}_\lambda\{u, v\} = (\sigma + \lambda) \int u\, v \,\mathrm{d}x \quad \forall v \in H_{\mathrm{p,o}}, \tag{A.3}$$

where

$$\mathcal{Q}_\lambda\{u, v\} := \int u''\, v'' - u'\, v' + (f + \lambda) u\, v \,\mathrm{d}x. \tag{A.4}$$

Consequently, if u is an eigenfunction of D with eigenvalue σ, then it is also an eigenfunction of the operator D_λ defined by

$$D_\lambda u := Du + \lambda u, \tag{A.5}$$

with corresponding eigenvalue $\sigma + \lambda$. Conversely, if u is an eigenfunction of D_λ with eigenvalue η, then it is also an eigenfunction of D with eigenvalue $\sigma = \eta - \lambda$. Thus, to study the eigenvalue problem (3.17) it suffices to analyse the eigenvalues of D_λ for some convenient value of λ. This is possible thanks to the following result.

Lemma A.1. *Let $\|u\|_{\mathrm{s}}$ be the Sobolev-type norm on $H_{\mathrm{p,o}}$ defined according to*

$$\|u\|_{\mathrm{s}} := \left(\|u''\|_2^2 + \|u'\|_2^2 + \|u\|_2^2 \right)^{\frac{1}{2}}. \tag{A.6}$$

There exist a constant $\alpha > 0$ such that, for any $\lambda \geq \alpha$,

$$\mathcal{Q}_\lambda\{u, u\} \geq \frac{1}{2} \|u\|_{\mathrm{s}}^2, \qquad u \in H_{\mathrm{p,o}}. \tag{A.7}$$

Proof. The assumption that $f \in L^\infty(-\ell, \ell)$, integration by parts, the Cauchy–Schwarz inequality, and the elementary inequality $ab \leq a^2/3 + 3b^2/4$ can be used to estimate

$$\begin{aligned} \|u''\|_2^2 + \frac{1}{2}\|u'\|_2^2 &= \mathcal{Q}_0\{u, u\} + \frac{3}{2}\int |u'|^2 \,\mathrm{d}x - \int f\, u^2 \,\mathrm{d}x \\ &\leq \mathcal{Q}_0\{u, u\} - \frac{3}{2}\int u''\, u \,\mathrm{d}x + \|f\|_\infty \|u\|_2^2 \\ &\leq \mathcal{Q}_0\{u, u\} + \frac{1}{2}\|u''\|_2^2 + \left(\frac{9}{8} + \|f\|_\infty\right)\|u\|_2^2. \end{aligned} \tag{A.8}$$

To conclude the proof, simply add $\frac{1}{2}\|u\|_2^2$ to both sides, rearrange, and observe that $\mathcal{Q}_\lambda\{u, u\} = \mathcal{Q}_0\{u, u\} + \lambda \|u\|_2^2$ for any λ. □

Lemma A.1 can be used to show that the eigenvalues of the linear operator D must be bounded from below. Indeed, letting $\lambda = -\sigma$ in (A.3) implies that any eigenfunction-eigenvalue pair (u, σ) satisfies

$$\mathcal{Q}_{-\sigma}\{u, v\} = 0 \quad \forall v \in H_{\mathrm{p,o}}. \tag{A.9}$$

The bilinear form $\mathcal{Q}_{-\sigma}\{u, v\}$ is obviously bounded on $H_{\mathrm{p,o}}$ (using $\|\cdot\|_{\mathrm{s}}$ as the underlying norm) and Lemma A.1 implies that it is also positive definite for all $\sigma \leq -\alpha$. Then, the Lax-Milgram theorem (Evans, 2010, section 6.2.1) guarantees that $u = 0$ is the only function satisfying (A.9) for $\sigma \leq -\alpha$, so any eigenvalues of D must be such that $\sigma > -\alpha$.

Now, fix any $\lambda > \alpha$. The bilinear form $\mathcal{Q}_\lambda\{u, v\}$ satisfies the hypotheses of the Lax-Milgram theorem, and for any $g \in L^2(-\ell, \ell)$ there exists a unique $u \in H_{\mathrm{p,o}}$ such that

$$\mathcal{Q}_\lambda\{u, v\} = \int g\, v \,\mathrm{d}x \quad \forall v \in H_{\mathrm{p,o}}. \tag{A.10}$$

Let $K : L^2(-\ell,\ell) \to L^2(-\ell,\ell)$ be the linear operator such that the unique u satisfying (A.10) is given by $u = Kg$. Note that if $u \in H_{\mathrm{p,o}}$ is an eigenfunction of problem (3.17) with eigenvalue $\sigma \neq -\lambda$, then it is also an eigenfunction of the operator D_λ with eigenvalue $\sigma + \lambda \neq 0$, and an eigenfunction of K with eigenvalue $1/(\sigma+\lambda)$. The latter observation follows from the fact that, for all $v \in H_{\mathrm{p,o}}$,

$$\int g\, v \,\mathrm{d}x = \mathcal{Q}_\lambda\{u, v\} = \int v\, D_\lambda u \,\mathrm{d}x = (\sigma+\lambda)\int u\, v \,\mathrm{d}x = (\sigma+\lambda)\int Kg\, v \,\mathrm{d}x. \tag{A.11}$$

Then, part (i) of Theorem 3.1 is a direct consequence of the following result.

Proposition A.2. *The linear operator K has at most a countable sequence of real and non-negative eigenvalues that converge to zero. In addition, there exist a sequence $\{w_k\}_{k\in\mathbb{N}}$ of eigenfunctions of K that form an orthonormal basis for $L^2(-\ell,\ell)$.*

Proof. The result follows from the theory of self-adjoint compact operators (Evans, 2010, appendix D.6) if one can show that K is a bounded, compact, self-adjoint operator mapping $L^2(-\ell,\ell)$ to $L^2(-\ell,\ell)$, and that $\int g\, Kg \,\mathrm{d}x \geq 0$ for all $g \in L^2(-\ell,\ell)$.

The assumption that $\lambda > \alpha$ means that one can use (A.10), Lemma A.1, and the Cauchy–Schwarz inequality to estimate

$$\frac{1}{2}\left\|Kg\right\|_{\mathrm{s}}^2 \leq \mathcal{Q}_\lambda\{u,u\} = \int g\, u \,\mathrm{d}x \leq \left\|g\right\|_2 \left\|u\right\|_2 \leq \left\|g\right\|_2 \left\|u\right\|_{\mathrm{s}} = \left\|g\right\|_2 \left\|Kg\right\|_{\mathrm{s}}. \tag{A.12}$$

Consequently, there exists a constant $C > 0$ such that $\left\|Kg\right\|_{\mathrm{s}} \leq C \left\|g\right\|_2$ and, since $H_{\mathrm{p,o}}$ is compactly embedded in $L^2(-\ell,\ell)$,[1] one concludes that K is a bounded and compact operator mapping $L^2(-\ell,\ell)$ to itself.

To show that K is self-adjoint, consider $g, h \in L^2(-\ell,\ell)$ and let $u_1 = Kg$, $u_2 = Kh$. The bilinear form $\mathcal{Q}_\lambda$ is symmetric, and it follows from (A.10) that

$$\int h\, Kg \,\mathrm{d}x = \int h\, u_1 \,\mathrm{d}x = \mathcal{Q}_\lambda\{u_2, u_1\} = \mathcal{Q}_\lambda\{u_1, u_2\} = \int g\, u_2 \,\mathrm{d}x = \int g\, Kh \,\mathrm{d}x. \tag{A.13}$$

Finally, for any $g \in L^2(-\ell,\ell)$ let $u = Kg$ and use (A.10) and Lemma A.1 to obtain

$$\int g\, Kg \,\mathrm{d}x = \int g\, u \,\mathrm{d}x = \mathcal{Q}_\lambda\{u,u\} \geq 0. \tag{A.14}$$

□

In order to prove part (ii) of Theorem 3.1, recall that the sequence $\{w_k\}_{k\in\mathbb{N}}$ of eigenfunctions of K is also a sequence of eigenfunctions for the operator D_λ in (A.5), as well as

[1] An equivalent result for Sobolev spaces of periodic functions is proven by Robinson (2001, appendix A) using Fourier series expansions. Replacing the Fourier series with a sine series yields the result for $H_{\mathrm{p,o}}$.

for the operator $D = D_0$ in problem (3.17). Recall also that, if $\{\eta_k\}_{k\in\mathbb{N}}$ and $\{\sigma_k\}_{k\in\mathbb{N}}$ are the sequences of eigenvalues of D_λ and D, respectively, then $\sigma_k = \eta_k - \lambda$. Note that $\eta_k > 0$ for all $k \in \mathbb{N}$ and $\eta_k \to \infty$ as $k \to \infty$ because they are the reciprocal of the eigenvalues of the operator K, which are non-negative and tend to zero according to Proposition A.2.

First, observe that $u = w_k$ satisfies (A.10) with $g = \eta_k w_k$ for any $k \in \mathbb{N}$. Moreover, setting $u = w_k$ and $g = \eta_k w_k$ in (A.10), integrating by parts, and rearranging gives

$$\int w_k'' \, v'' \, \mathrm{d}x = \int \left[(\eta_k - \lambda - f) w_k - w_k''\right] v \, \mathrm{d}x. \tag{A.15}$$

The term in square brackets on the right-hand side of this identity is the second weak derivative of w_k'' by the very definition of weak derivatives, meaning that w_k has at least four weak derivatives. Consequently, for any $u \in H_{\mathrm{p,o}}$ it makes sense to use integration by parts to write

$$\mathcal{Q}_\lambda\{w_k, u\} = \int u \, D_\lambda w_k \, \mathrm{d}x = \eta_k \int u \, w_k \, \mathrm{d}x, \tag{A.16}$$

where the second inequality follows because w_k is an eigenfunction of D_λ with eigenvalue η_k. This identity and the fact that the functions w_k are orthonormal imply that, for all $k, l = 0, 1, \ldots$ with $k \neq l$,

$$\mathcal{Q}_\lambda\{w_k, w_k\} = \eta_k \int w_k^2 \, \mathrm{d}x = \eta_k, \tag{A.17a}$$

$$\mathcal{Q}_\lambda\{w_k, w_l\} = \eta_k \int w_k \, w_l \, \mathrm{d}x = 0. \tag{A.17b}$$

Second, note that $\{w_k\}_{k\in\mathbb{N}}$ is an orthonormal basis for $L^2(-\ell, \ell)$, so any $u \in H_{\mathrm{p,o}} \subset L^2(-\ell, \ell)$ can be written as

$$u = \sum_{k\geq 0} c_k w_k, \tag{A.18}$$

the series converging in the L^2 sense. If, moreover, $\|u\|_2 = 1$, then the expansion coefficients $c_k := \int u \, w_k \, \mathrm{d}x$ are such that

$$\sum_{k\geq 0} {c_k}^2 = \|u\|_2^2 = 1. \tag{A.19}$$

In addition, the sequence $\{\eta_k^{-1/2} w_k\}_{k\in\mathbb{N}}$ is an orthonormal basis for $H_{\mathrm{p,o}}$ when endowed with the inner product defined by the positive definite[2], symmetric, and bilinear form $\mathcal{Q}_\lambda$. In fact, since $\{w_k\}_{k\in\mathbb{N}}$ is an orthonormal basis of $L^2(-\ell, \ell)$ it follows from (A.16) that $u = 0$ is the only member of $H_{\mathrm{p,o}}$ satisfying

$$\mathcal{Q}_\lambda\{\eta_k^{-1/2} w_k, u\} = 0, \quad k = 0, 1, \ldots. \tag{A.20}$$

[2]Positive definiteness follows from Lemma A.1 because it has been assumed that $\lambda > \alpha$.

Consequently, any $u \in H_{\mathrm{p,o}}$ can be expanded as

$$u = \sum_{k\geq 0} d_k \frac{w_k}{\sqrt{\eta_k}}, \qquad d_k := \mathcal{Q}_\lambda\{u, \eta_k^{-1/2} w_k\}, \tag{A.21}$$

the series converging in $H_{\mathrm{p,o}}$. But (A.16) implies that

$$d_k = \mathcal{Q}_\lambda\{u, \eta_k^{-1/2} w_k\} = \mathcal{Q}_\lambda\{\eta_k^{-1/2} u, w_k\} = \eta_k^{1/2} \int u\, w_k \,\mathrm{d}x = \eta_k^{1/2} c_k, \tag{A.22}$$

meaning that the series (A.18) converges not only in $L^2(-\ell, \ell)$, but also in $H_{\mathrm{p,o}}$. Thus, for any $u \in H_{\mathrm{p,o}}$ with $\|u\|_2 = 1$, equations (A.17a), (A.17b), (A.19) and the fact that $0 < \eta_0 \leq \eta_1 \leq \ldots$ imply

$$\mathcal{Q}_\lambda\{u, u\} = \sum_{k\geq 0} \eta_k \, {c_k}^2 \geq \eta_0, \tag{A.23}$$

and hence

$$\inf_{\substack{u\in H_{\mathrm{p,o}} \\ \|u\|_2=1}} \mathcal{Q}_0\{u, u\} = \inf_{\substack{u\in H_{\mathrm{p,o}} \\ \|u\|_2=1}} \mathcal{Q}_\lambda\{u, u\} - \lambda \geq \eta_0 - \lambda = \sigma_0. \tag{A.24}$$

Equality holds for $u = w_0$ so the infimum is attained, proving part (ii) of Theorem 3.1.

A.2 Proof of Theorem 4.2

Define the norm $\|\boldsymbol{w}\|_k^2 := \int_{-1}^1 (\mathcal{D}^k \boldsymbol{w})^\mathsf{T} \mathcal{D}^k \boldsymbol{w} \,\mathrm{d}x$, consider the functional

$$\mathcal{H}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\} := \frac{\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}}{\|\boldsymbol{w}\|_k^2}, \tag{A.25}$$

and let

$$t(\boldsymbol{\gamma}) := \inf_{\boldsymbol{w}\in H\setminus\{\boldsymbol{0}\}} \mathcal{H}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}, \qquad t_N(\boldsymbol{\gamma}) := \inf_{\boldsymbol{w}\in S_N\setminus\{\boldsymbol{0}\}} \mathcal{H}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}. \tag{A.26}$$

These infima need not be achieved. Clearly, the sets T and T_N^{out} are described by the inequalities $t(\boldsymbol{\gamma}) \geq 0$ and $t_N(\boldsymbol{\gamma}) \geq 0$, respectively, and one has the following result.

Lemma A.3. *Suppose $\boldsymbol{\gamma} \notin T$, meaning that there exists $\varepsilon_{\boldsymbol{\gamma}} > 0$ such that $t(\boldsymbol{\gamma}) \leq -2\varepsilon_{\boldsymbol{\gamma}}$. Then, there exists an integer $N_{\boldsymbol{\gamma}}$ such that $t_N(\boldsymbol{\gamma}) \leq -\varepsilon_{\boldsymbol{\gamma}}$ for all $N \geq N_{\boldsymbol{\gamma}}$.*

Proof. Let $\{\boldsymbol{w}_n\}_{n\geq 0}$, $\boldsymbol{w}_n \in H$, $\boldsymbol{w}_n \neq \boldsymbol{0}$ be a minimising sequence for $\mathcal{H}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}$, such that

$$\lim_{n\to\infty} \mathcal{H}_{\boldsymbol{\gamma}}\{\boldsymbol{w}_n\} = t(\boldsymbol{\gamma}), \tag{A.27}$$

and for each n define $\mu_n := \|\boldsymbol{w}_n\|_k^2 / (n+1)$. Note that $F\big(x, \mathcal{D}^{\boldsymbol{k}}\boldsymbol{w}(x)\big)$, the integrand of $\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\}$, and the product $(\mathcal{D}^{\boldsymbol{k}}\boldsymbol{w}(x))^\mathsf{T} \mathcal{D}^{\boldsymbol{k}}\boldsymbol{w}(x)$ are continuous with respect to all entries of

the vector $\mathcal{D}^{\boldsymbol{k}}\boldsymbol{w}(x)$ at each fixed $x \in [-1, 1]$. Using

$$|\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\} - \mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}_n\}| \leq 2 \left\| F\left(x, \mathcal{D}^{\boldsymbol{k}}\boldsymbol{w}\right) - F\left(x, \mathcal{D}^{\boldsymbol{k}}\boldsymbol{w}_n\right) \right\|_\infty \tag{A.28}$$

and a similar inequality for $\left| \|\boldsymbol{w}\|_k^2 - \|\boldsymbol{w}_n\|_k^2 \right|$, it is then not difficult to show that there exists $\delta_n > 0$ such that if

$$\max_{0 \leq \alpha \leq k} \|\partial^\alpha \boldsymbol{w} - \partial^\alpha \boldsymbol{w}_n\|_\infty \leq \delta_n, \tag{A.29}$$

then

$$|\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}\} - \mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}_n\}| \leq \mu_n, \tag{A.30a}$$

$$\left| \|\boldsymbol{w}\|_k^2 - \|\boldsymbol{w}_n\|_k^2 \right| \leq \mu_n. \tag{A.30b}$$

Since the Weierstrass approximation theorem can be extended to linear subspaces of continuously differentiable functions with prescribed boundary conditions (Peet & Bliman, 2007, Proposition 2), there exists a vector-valued polynomial $\boldsymbol{P}_n \in H$ of degree d_n, $\boldsymbol{P}_n \neq \boldsymbol{0}$, that satisfies (A.29). It may be assumed without loss of generality that $d_n < d_{n+1}$. Using (A.30b) it can be shown that

$$\frac{n+1}{n+2} \leq \frac{\|\boldsymbol{w}_n\|_k^2}{\|\boldsymbol{P}_n\|_k^2} \leq \frac{n+1}{n} \tag{A.31}$$

and since $\boldsymbol{P}_n \in S_N$ for all $N \in \{d_n, \ldots, d_{n+1} - 1\}$ one can utilise (A.30a), (A.30b), and the definition of μ_n to write

$$t_N(\boldsymbol{\gamma}) \leq \mathcal{H}_{\boldsymbol{\gamma}}\{\boldsymbol{P}_n\} \leq \frac{|\mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{P}_n\} - \mathcal{F}_{\boldsymbol{\gamma}}\{\boldsymbol{w}_n\}|}{\|\boldsymbol{P}_n\|_k^2} + \frac{\|\boldsymbol{w}_n\|_k^2}{\|\boldsymbol{P}_n\|_k^2} \mathcal{H}_{\boldsymbol{\gamma}}\{\boldsymbol{w}_n\} \leq \frac{1}{n} + \frac{\|\boldsymbol{w}_n\|_k^2}{\|\boldsymbol{P}_n\|_k^2} \mathcal{H}_{\boldsymbol{\gamma}}\{\boldsymbol{w}_n\}. \tag{A.32}$$

Since $t(\boldsymbol{\gamma}) \leq t_N(\boldsymbol{\gamma})$, it follows from (A.27), (A.31), and (A.32) that $t_N(\boldsymbol{\gamma}) - t(\boldsymbol{\gamma}) \downarrow 0$ as N (hence n) tends to infinity. Then, since $t(\boldsymbol{\gamma}) \leq -2\varepsilon_{\boldsymbol{\gamma}}$ by assumption, there exists $N_{\boldsymbol{\gamma}} \in \mathbb{N}$ such that

$$t_N(\boldsymbol{\gamma}) \leq \varepsilon_{\boldsymbol{\gamma}} + t(\boldsymbol{\gamma}) \leq \varepsilon_{\boldsymbol{\gamma}} - 2\varepsilon_{\boldsymbol{\gamma}} = -\varepsilon_{\boldsymbol{\gamma}} < 0 \quad \forall N \geq N_{\boldsymbol{\gamma}}. \tag{A.33}$$

□

Armed with Lemma A.3, one can prove Theorem 4.2. That the sequence of optimal values $\{p_N^\star\}_{N \geq 0}$ is non-decreasing follows from the inclusion $T_{N+1}^{\text{out}} \subset T_N^{\text{out}}$. All that is left to prove is convergence when (4.14) achieves its optimal value at an optimal point $\boldsymbol{\gamma}^\star$. To do so, suppose that the feasible set T of (4.14) is *bounded*. If this were not the case, one could formulate an equivalent problem with bounded feasible set and for which $\boldsymbol{\gamma}^\star$ remains an optimal solution simply by adding the constraint $\|\boldsymbol{\gamma}\| \leq r$ for a sufficiently large $r > 0$.

Since T is bounded, for any $\varepsilon > 0$ (different from that used in Lemma A.3), the set

$$K := \{\boldsymbol{\gamma} : \varepsilon \leq \operatorname{dist}(\boldsymbol{\gamma}, T) \leq 2\,\varepsilon\}, \tag{A.34}$$

where $\operatorname{dist}(\boldsymbol{\gamma}, T) = \min_{\boldsymbol{\eta} \in T} \|\boldsymbol{\eta} - \boldsymbol{\gamma}\|$ is the usual euclidean distance of $\boldsymbol{\gamma}$ from T, is compact. In addition, it only contains points that are infeasible for (4.14). By Lemma A.3, for each $\boldsymbol{\gamma} \in K$ there exists an integer $N_{\boldsymbol{\gamma}}$ such that $t_N(\boldsymbol{\gamma}) < 0$ for all $N \geq N_{\boldsymbol{\gamma}}$, meaning that $\boldsymbol{\gamma}$ is infeasible for (4.24) for all $N \geq N_{\boldsymbol{\gamma}}$.

At this stage, one can combine this observation with the compactness of K and the continuity of $t_N(\boldsymbol{\gamma})$ with $\boldsymbol{\gamma}$—the proof of this fact is not difficult and is left to the reader—to find an integer $N_0 = N_0(\varepsilon)$, a finite collection of points $\{\boldsymbol{\gamma}_i\}_{i=1}^{N_0}$ in K, and positive values $\{\delta_i\}_{i=1}^{N_0}$ such that (i) the balls $B(\boldsymbol{\gamma}_i, \delta_i)$ with centre $\boldsymbol{\gamma}_i$ and radius δ_i cover K, and (ii) $t_N(\boldsymbol{\gamma}) < 0$ in each ball for all $N \geq N_0$. Consequently, all points in K are infeasible for the outer SDP relaxation (4.24) when $N \geq N_0$.

Now, the feasible set T_N^{out} of the outer SDP relaxation is convex, so in particular it is connected. Then, it must be contained within an ε-neighbourhood of T for all $N \geq N_0$:

$$\forall N \geq N_0, \quad \max_{\boldsymbol{\gamma} \in T_N^{\text{out}}} \operatorname{dist}(\boldsymbol{\gamma}, T) < \varepsilon. \tag{A.35}$$

In particular, T_N^{out} is bounded, and there exists a point $\boldsymbol{\gamma}_N^\star$ with $\boldsymbol{c}^T \boldsymbol{\gamma}_N^\star = p_N^\star$ whose projection onto T, denoted $\mathbb{P}_T(\boldsymbol{\gamma}_N^\star)$, satisfies $\|\mathbb{P}_T(\boldsymbol{\gamma}_N^\star) - \boldsymbol{\gamma}_N^\star\| < \varepsilon$. Then, for all $N \geq N_0$,

$$\begin{aligned} p^\star - \|\boldsymbol{c}\|\,\varepsilon &\leq \boldsymbol{c}^{\mathsf{T}} \mathbb{P}_T(\boldsymbol{\gamma}_N^\star) - \|\boldsymbol{c}\|\,\varepsilon \\ &= \boldsymbol{c}^{\mathsf{T}} \left[\mathbb{P}_T(\boldsymbol{\gamma}_N^\star) - \boldsymbol{\gamma}_N^\star\right] + p_N^\star - \|\boldsymbol{c}\|\,\varepsilon \\ &\leq \|\boldsymbol{c}\|\ \|\mathbb{P}_T(\boldsymbol{\gamma}_N^\star) - \boldsymbol{\gamma}_N^\star\| + p_N^\star - \|\boldsymbol{c}\|\,\varepsilon \\ &< p_N^\star. \end{aligned} \tag{A.36}$$

Since $p_N^\star \leq p^\star$ then $p^\star - \|\boldsymbol{c}\|\,\varepsilon \leq \lim_{N\to\infty} p_N^\star \leq p^\star$ for any ε, and the proof of Theorem 4.2 is concluded by letting $\varepsilon \to 0$.

A.3 Proof of Lemma 4.3

The statement is trivial when $\alpha = k$. For $\alpha \leq k - 1$, the fundamental theorem of calculus can be applied to show that

$$\left(\partial^\alpha u\right)(x) = \partial^\alpha u(-1) + \int_{-1}^{x} \partial^{\alpha+1} u(t)\,\mathrm{d}t = \partial^\alpha u(-1) + \sum_{n\geq 0} \hat{u}_n^{\alpha+1} \int_{-1}^{x} L_n(t)\,\mathrm{d}t. \tag{A.37}$$

Integration and summation can be exchanged because $u \in C^m([-1,1])$ and $k \leq m-1$, so the Legendre expansions of $\partial^\alpha u$, $\alpha \in \{0, ..., k\}$, converge uniformly (cf. section 4.1).

The last expression in (A.37) can be integrated recalling that $L_0(x) = 1$, $L_1(x) = x$, $L_n(\pm 1) = (\pm 1)^n$ and using the recurrence relation (4.8). Then,

$$\partial^\alpha u = \partial^\alpha u(-1) + [L_1 + L_0]\, \hat{u}_0^{\alpha+1} + \sum_{n\geq 1} [L_{n+1} - L_{n-1}]\, \frac{\hat{u}_n^{\alpha+1}}{2n+1}. \tag{A.38}$$

Rearranging the series and comparing coefficients with the Legendre expansion of $\partial^\alpha u$ yields

$$\hat{u}_0^\alpha = \partial^\alpha u(-1) + \hat{u}_0^{\alpha+1} - \frac{1}{3}\hat{u}_1^{\alpha+1}, \tag{A.39a}$$

$$\hat{u}_n^\alpha = \frac{\hat{u}_{n-1}^{\alpha+1}}{2n-1} - \frac{\hat{u}_{n+1}^{\alpha+1}}{2n+3}, \quad n \geq 1. \tag{A.39b}$$

Using these identities one can easily construct matrices $\boldsymbol{C}^\alpha$ and $\boldsymbol{E}^\alpha$ such that

$$\hat{\boldsymbol{u}}_{[r,s]}^\alpha = \boldsymbol{E}^\alpha \mathcal{D}^{k-1} u(-1) + \boldsymbol{C}^\alpha \hat{\boldsymbol{u}}_{[r-1,s+1]}^{\alpha+1}. \tag{A.40}$$

Here and in the following, it should be understood that negative indices should be replaced by 0. Note that the matrices $\boldsymbol{C}^\alpha$ and $\boldsymbol{E}^\alpha$ depend on r and s, but this is not indicated explicitly to ease the notation. Note also that (A.39) implies $\boldsymbol{E}^\alpha = \boldsymbol{0}$ if $r \geq 1$.

Expressions similar to (A.40) can be built for all vectors $\hat{\boldsymbol{u}}_{[r-i,s+i]}^{\alpha+i}$, $i \in \{0, \ldots, k-\alpha-1\}$. After some algebra, it is therefore possible to write

$$\hat{\boldsymbol{u}}_{[r,s]}^\alpha = \boldsymbol{B}_{[r,s]}^\alpha \mathcal{D}^{k-1} u(-1) + \left(\prod_{i=0}^{k-\alpha-1} \boldsymbol{C}^{\alpha+i} \right) \hat{\boldsymbol{u}}_{[r-k+\alpha,s+k-\alpha]}^k, \tag{A.41}$$

where

$$\boldsymbol{B}_{[r,s]}^\alpha := \boldsymbol{E}^\alpha + \boldsymbol{C}^\alpha \boldsymbol{E}^{\alpha+1} + \cdots + \left(\prod_{i=0}^{k-\alpha-2} \boldsymbol{C}^{\alpha+i} \right) \boldsymbol{E}^{k-1}. \tag{A.42}$$

Note that, in light of (A.39), all matrices $\boldsymbol{E}^{\alpha+i}$, $i \in \{0, \ldots, k-\alpha-1\}$, are zero if $r \geq k-\alpha$. Since $s + k - \alpha \leq M$ by assumption, the last term in (A.41) can be rewritten in terms of $\hat{\boldsymbol{u}}_{[0,M]}^k$ (recall that $r - k + \alpha$ is replaced by 0 if it is negative). Defining

$$\boldsymbol{D}_{[r,s]}^\alpha := \left[\boldsymbol{0}_{(s-r+1)\times(r-k+\alpha)}, \quad \prod_{i=0}^{k-\alpha-1} \boldsymbol{C}^{\alpha+i}, \quad \boldsymbol{0}_{(s-r+1)\times(M-s-k+\alpha)} \right], \tag{A.43}$$

where subscripts indicate the size of the zero matrices, concludes the proof.

A.4 Proof of Lemma 4.4

Recalling the definition of the vector of boundary values $\mathcal{B}^{k-1}u$ from (4.5), one only needs to show that $\mathcal{D}^{k-1}u(1)$ can be expressed as linear combination of the entries of $\check{\boldsymbol{u}}_M$. Applying the fundamental theorem of calculus as in appendix A.3, for any $\alpha \in \{0, \dots, k-1\}$ it may be shown that

$$\partial^\alpha u(1) = \partial^\alpha u(-1) + 2\hat{u}_0^{\alpha+1}. \tag{A.44}$$

According to Lemma 4.3, $\partial^\alpha u(1)$ can then be written as a linear combination of the entries of $\check{\boldsymbol{u}}_M$. Repeating this argument for all $\alpha \in \{0, \dots, k-1\}$ shows that the same is true for all entries of $\mathcal{D}^{k-1}u(1)$, proving the existence of the matrix $\boldsymbol{G}_M$ as desired.

A.5 Proof of Lemma 4.6

(i) Recall (4.26) and expand

$$\mathcal{Q}_{uv}^{\alpha\beta} = \sum_{m=0}^{N_\alpha} \sum_{n=N_\beta+1}^{+\infty} \hat{u}_m^\alpha \hat{v}_n^\beta \int_{-1}^{1} f\, L_m\, L_n \,\mathrm{d}x + \sum_{m=N_\alpha+1}^{\infty} \sum_{n=0}^{N_\beta} \hat{u}_m^\alpha \hat{v}_n^\beta \int_{-1}^{1} f\, L_m\, L_n \,\mathrm{d}x, \tag{A.45}$$

where $N_\alpha = N + \alpha$ and $N_\beta = N + \beta$. Since f is a polynomial of degree at most d_F, the product $f\,L_m$ is a polynomial of degree at most $m + d_F$, so it is orthogonal to any Legendre polynomial L_n with $n > m + d_F$. In particular, it may be shown (Dougall, 1953) that the integral $\int_{-1}^{1} f L_n L_m \,\mathrm{d}x$ vanishes if $|m - n| > d_F$. Using the short-hand notation $\overline{n} = n + 1 - d_F$, one can therefore write

$$\mathcal{Q}_{uv}^{\alpha\beta} = \begin{bmatrix} \hat{u}^\alpha_{\overline{N_\beta}} \\ \vdots \\ \hat{u}^\alpha_{N_\alpha} \end{bmatrix}^{\mathsf{T}} \boldsymbol{\Phi}_{[\overline{N_\beta}, N_\alpha]}^{[N_\beta+1, N_\alpha+d_F]} \begin{bmatrix} \hat{v}^\beta_{N_\beta+1} \\ \vdots \\ \hat{v}^\beta_{N_\alpha+d_F} \end{bmatrix} + \begin{bmatrix} \hat{v}^\beta_{\overline{N_\alpha}} \\ \vdots \\ \hat{v}^\beta_{N_\beta} \end{bmatrix}^{\mathsf{T}} \boldsymbol{\Phi}_{[\overline{N_\alpha}, N_\beta]}^{[N_\alpha+1, N_\beta+d_F]} \begin{bmatrix} \hat{u}^\alpha_{N_\alpha+1} \\ \vdots \\ \hat{u}^\alpha_{N_\beta+d_F} \end{bmatrix}. \tag{A.46}$$

For generality, it has been assumed here that α, β and d_F satisfy $1 - d_F \leq \alpha - \beta \leq d_F - 1$, so the vectors in (A.46) are well-defined. If the first (resp. second) of these conditions is violated, then the first (resp. second) term in (A.46) simply vanishes.

Since $N_\alpha + d_F \leq M + \beta - k$ and $N_\beta + d_F \leq M + \alpha - k$, the conditions of Lemma 4.3 hold. In addition, the assumption that $N \geq d_F + k - 1$ guarantees that $\overline{N_\alpha} \geq k - \beta$ and $\overline{N_\beta} \geq k - \alpha$, so Lemma 4.3 can be applied with no dependence on the boundary values. Consequently, there exists a matrix $\boldsymbol{Q}(\boldsymbol{\gamma})$ such that

$$\mathcal{Q}_{uv}^{\alpha\beta} = \left(\hat{\boldsymbol{u}}^k_{[0,M]}\right)^{\mathsf{T}} \boldsymbol{Q}(\boldsymbol{\gamma})\, \hat{\boldsymbol{v}}^k_{[0,M]}. \tag{A.47}$$

Finally, the matrix $\boldsymbol{Q}_{uv}^{\alpha\beta}$ is constructed by using (4.38) after taking the symmetric part of the right-hand side of (A.47).

(ii) Let

$$\boldsymbol{\nu} := \left[\hat{u}_{M+1}^k, \ldots, \hat{u}_{M+d_F}^k, \hat{v}_{M+1}^k, \ldots, \hat{v}_{M+d_F}^k\right]^{\mathsf{T}}. \tag{A.48}$$

After replacing N_α and N_β with M in (A.46), it may be verified using (4.38) that

$$\mathcal{Q}_{uv}^{kk} = 2\,\boldsymbol{\psi}_M{}^{\mathsf{T}}\,\boldsymbol{\Xi}_{\overline{M}}{}^{\mathsf{T}}\boldsymbol{Y}\boldsymbol{\nu}. \tag{A.49}$$

By (4.42),

$$\begin{aligned} 0 &\leq \begin{bmatrix} \boldsymbol{\Xi}_{\overline{M}}\boldsymbol{\psi}_M \\ \boldsymbol{\nu} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \boldsymbol{Q}_{uv}^{kk} & \boldsymbol{Y} \\ \boldsymbol{Y}^{\mathsf{T}} & \boldsymbol{\Sigma}_{uv}^{kk} \otimes \boldsymbol{\Delta} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Xi}_{\overline{M}}\boldsymbol{\psi}_M \\ \boldsymbol{\nu} \end{bmatrix} \\ &= \boldsymbol{\psi}_M{}^{\mathsf{T}} \left(\boldsymbol{\Xi}_{\overline{M}}{}^{\mathsf{T}}\boldsymbol{Q}_{uv}^{kk}\boldsymbol{\Xi}_{\overline{M}}\right)\boldsymbol{\psi}_M + \boldsymbol{\nu}^{\mathsf{T}}\left(\boldsymbol{\Sigma}_{uv}^{kk} \otimes \boldsymbol{\Delta}\right)\boldsymbol{\nu} + \mathcal{Q}_{uv}^{kk}. \end{aligned} \tag{A.50}$$

Now, $\boldsymbol{\Sigma}_{uv}^{kk}$ is a diagonal matrix by assumption. Then, the definitions of $\boldsymbol{\Delta}$ and $\boldsymbol{\nu}$ imply

$$\mathcal{Q}_{uv}^{kk} \geq -\boldsymbol{\psi}_M{}^{\mathsf{T}}\left(\boldsymbol{\Xi}_{\overline{M}}{}^{\mathsf{T}}\boldsymbol{Q}_{uv}^{kk}\boldsymbol{\Xi}_{\overline{M}}\right)\boldsymbol{\psi}_M - (\boldsymbol{\Sigma}_{uv}^{kk})_{1,1}\sum_{n=M+1}^{M+d_F}\frac{2|\hat{u}_n^k|^2}{2n+1} - (\boldsymbol{\Sigma}_{uv}^{kk})_{2,2}\sum_{n=M+1}^{M+d_F}\frac{2|\hat{v}_n^k|^2}{2n+1}. \tag{A.51}$$

The sums in (A.51) can be bounded by $\left\|U_M^k\right\|_2^2$ and $\left\|V_M^K\right\|_2^2$ by virtue of (4.11), giving (4.43).

A.6 Proof of Lemma 4.7

For each $\alpha \leq k$, the quantity $\|U_{N_\alpha}^\alpha\|_2^2$ can be bounded in terms of the vector $\hat{\boldsymbol{u}}_{[0,M]}^k$ and $\|U_M^k\|_2^2$ (a similar bound can be found for $V_{N_\beta}^\beta$). To derive this bound, begin by noticing that (4.26), (4.27), and (4.31) imply

$$\begin{aligned} \frac{1}{2}\left\|U_{N_\alpha}^\alpha\right\|_2^2 &= \sum_{n=N_\alpha+1}^{M-k+\alpha}\frac{(\hat{u}_n^\alpha)^2}{2n+1} + \sum_{n=M-k+\alpha+1}^{+\infty}\frac{(\hat{u}_n^\alpha)^2}{2n+1} \\ &= \left(\hat{\boldsymbol{u}}_{[0,M]}^k\right)^{\mathsf{T}}\boldsymbol{H}_\alpha\,\hat{\boldsymbol{u}}_{[0,M]}^k + \sum_{n=M-k+\alpha+1}^{+\infty}\frac{(\hat{u}_n^\alpha)^2}{2n+1}, \end{aligned} \tag{A.52}$$

where the matrix $\boldsymbol{H}_\alpha$ can be obtained from Lemma 4.3. Since (A.39b) is applied $k-\alpha$ times to $(\hat{u}_n^\alpha)^2$ to compute $\boldsymbol{H}_\alpha$, and since $n > N_\alpha \geq N$, it follows that $\|\boldsymbol{H}_\alpha\|_{\mathrm{F}} \sim N^{-2(k-\alpha)-1}$.

When $\alpha = k$, the last term in (A.52) is $\frac{1}{2}\left\|U_M^k\right\|_2^2$, so

$$\frac{1}{2}\left\|U_{N_k}^k\right\|_2^2 = \left(\hat{\boldsymbol{u}}_{[0,M]}^k\right)^T\boldsymbol{H}_k\,\hat{\boldsymbol{u}}_{[0,M]}^k + \frac{1}{2}\left\|U_M^k\right\|_2^2. \tag{A.53}$$

When $\alpha \leq k-1$, instead, define

$$\omega_\eta := \frac{4}{[2(M-k+\eta)+1][2(M-k+\eta)+5]}, \quad \eta \in \{0, \dots, k-1\}. \tag{A.54}$$

Using (A.39), the elementary inequality $(a-b)^2 \leq 2(a^2+b^2)$, and appropriate changes of indices one can estimate

$$\begin{aligned}
\sum_{n=M-k+\alpha+1}^{+\infty} \frac{(\hat{u}_n^\alpha)^2}{2n+1} &\leq \sum_{n=M-k+\alpha+1}^{+\infty} \frac{2}{2n+1}\left[\frac{|\hat{u}_{n-1}^{\alpha+1}|^2}{(2n-1)^2} + \frac{|\hat{u}_{n+1}^{\alpha+1}|^2}{(2n+3)^2}\right] \\
&\leq \sum_{n=M-k+\alpha}^{M-k+\alpha+1} \frac{2|\hat{u}_n^{\alpha+1}|^2}{(2n+3)(2n+1)^2} + \sum_{n=M-k+\alpha+2}^{\infty} \frac{4|\hat{u}_n^{\alpha+1}|^2}{(2n-1)(2n+1)(2n+3)} \\
&\leq \sum_{n=M-k+\alpha}^{M-k+\alpha+1} \frac{2|\hat{u}_n^{\alpha+1}|^2}{(2n+3)(2n+1)^2} + \omega_{\alpha+1} \sum_{n=M-k+\alpha+2}^{+\infty} \frac{|\hat{u}_n^{\alpha+1}|^2}{2n+1}.
\end{aligned} \tag{A.55}$$

Applying Lemma 4.3 to the first term on the right-hand side of (A.55) and substituting back into (A.52) reveals that there exists a matrix $\boldsymbol{T}_\alpha$ such that

$$\frac{1}{2}\left\|U_{N_\alpha}^\alpha\right\|_2^2 \leq \left(\hat{\boldsymbol{u}}_{[0,M]}^k\right)^{\mathsf{T}} \boldsymbol{T}_\alpha \hat{\boldsymbol{u}}_{[0,M]}^k + \omega_{\alpha+1} \sum_{n=M-k+\alpha+2}^{+\infty} \frac{|\hat{u}_n^{\alpha+1}|^2}{2n+1}. \tag{A.56}$$

As for $\boldsymbol{H}_\alpha$, it may be verified that $\|\boldsymbol{T}_\alpha\|_{\mathrm{F}} \sim N^{-2(k-\alpha)-1}$.

Similar estimates can be carried out for the infinite sum on the right-hand side of (A.56). By recursion, one can eventually construct a matrix $\boldsymbol{Z}_\alpha$ and a constant λ_α such that

$$\frac{1}{2}\left\|U_{N_\alpha}^\alpha\right\|_2^2 \leq \left(\hat{\boldsymbol{u}}_{[0,M]}^k\right)^{T} \boldsymbol{Z}_\alpha \hat{\boldsymbol{u}}_{[0,M]}^k + \lambda_\alpha \left\|U_M^k\right\|_2^2. \tag{A.57}$$

Note that $\|\boldsymbol{Z}_\alpha\|_{\mathrm{F}} \sim N^{-2(k-\alpha)-1}$ and $\lambda_\alpha \sim N^{-2(k-\alpha)}$, because every step of the recursion procedure introduces a factor of N^{-2} according to (A.54). Moreover, the right-hand side of (A.57) has the same form as (A.53), so for the rest of this section no distinction will be made between the case $\alpha \leq k-1$ and the case $\alpha = k$.

The estimate (A.57) can be used in conjunction with Young's inequality and (4.38) to show that, for any $\varepsilon > 0$,

$$\begin{aligned}
|\mathcal{R}_{uv}^{\alpha\beta}| \leq \|f\|_\infty \boldsymbol{\psi}_M{}^{\mathsf{T}} \left(\boldsymbol{\Xi}_0{}^{\mathsf{T}} \begin{bmatrix} \varepsilon \boldsymbol{Z}_\alpha & \boldsymbol{0} \\ \boldsymbol{0} & \frac{1}{\varepsilon}\boldsymbol{Z}_\beta \end{bmatrix} \boldsymbol{\Xi}_0\right) \boldsymbol{\psi}_M \\
+ \|f\|_\infty \left(\varepsilon\lambda_\alpha \|U_k\|_2^2 + \frac{\lambda_\beta}{\varepsilon}\|V_k\|_2^2\right).
\end{aligned} \tag{A.58}$$

At this stage, set $\varepsilon = (N+1)^{\beta-\alpha}$, such that $\varepsilon\lambda_\alpha \sim \varepsilon^{-1}\lambda_\beta \sim N^{\alpha+\beta-2k}$ and $\|\varepsilon \boldsymbol{Z}_\alpha\|_{\text{F}} \sim \|\varepsilon^{-1}\boldsymbol{Z}_\beta\|_{\text{F}} \sim N^{\alpha+\beta-2k-1}$. Additionally, let

$$\boldsymbol{R}_{uv}^{\alpha\beta} := \boldsymbol{\Xi}_0{}^{\mathsf{T}} \begin{bmatrix} \varepsilon \boldsymbol{Z}_\alpha & \boldsymbol{0} \\ \boldsymbol{0} & \frac{1}{\varepsilon}\boldsymbol{Z}_\beta \end{bmatrix} \boldsymbol{\Xi}_0, \qquad \boldsymbol{\Sigma}_{uv}^{\alpha\beta} := \begin{bmatrix} \varepsilon\lambda_\alpha & 0 \\ 0 & \varepsilon^{-1}\lambda_\beta \end{bmatrix}. \tag{A.59}$$

Recalling that the Legendre polynomials satisfy $\|L_n\|_\infty \leq 1$ for all $n \geq 0$ (Jackson, 1930), equation (4.44) follows from the estimate

$$\|f\|_\infty = \sup_{x\in[-1,1]} \left| \sum_{n=0}^{p} \hat{f}_n(\gamma)\mathcal{L}_n(x) \right| \leq \sum_{n=0}^{p} \left| \hat{f}_n(\gamma) \right| = \left\| \hat{\boldsymbol{f}}(\gamma) \right\|_1. \tag{A.60}$$

A.7 Proof of (6.21)

Let $\hat{\theta}_0(z) = v(z)$ to simplify the notation. It is not difficult to check using the calculus of variations that the infimum of $\mathcal{Q}_0$ over all test functions v that satisfy $v(0) = 0$ and $v'(1) = 0$ is not attained unless $\beta = 2$. This difficulty can be resolved by noticing that

$$\inf_{\substack{v(0)=0,\\ v'(1)=0}} \mathcal{Q}_0\{v\} = \min_A \min_{\substack{v(0)=0,\\ v(1)=A}} \mathcal{Q}_0\{v\}. \tag{A.61}$$

In other words, one can replace the Neumann BC $v'(0) = 0$ with the Dirichlet condition $v(1) = A$, solve the Dirichlet problem

$$\mathcal{Q}_0^\star(A) := \min_{\substack{v(0)=0,\\ v(1)=A}} \mathcal{Q}_0\{v\}, \tag{A.62}$$

and minimise $\mathcal{Q}_0^\star(A)$ over A. Equation (A.61) is justified because for each value A, the minimum of the Dirichlet problem can be approximated with arbitrary accuracy by a function that satisfies $v'(1) = 0$; for example, if $v^\star$ is the minimiser of the Dirichlet problem (A.62) for a given A, take

$$v(z) = \begin{cases} v^\star(z), & 0 \leq z \leq 1-\delta, \\ v^\star(1-\delta), & 1-\delta \leq z \leq 1 \end{cases} \tag{A.63}$$

for $\delta > 0$ sufficiently small. A rigorous proof is omitted for brevity, but a similar argument was given by Fantuzzi & Wynn (2017, appendix C).

The minimiser of the Dirichlet problem (A.62) satisfies the Euler–Lagrange equation

$$-2\,v'' - \frac{\alpha-2}{\alpha-1}\,\tau'' = 0 \tag{A.64}$$

subject to the BCs $v(0) = 0$ and $v(1) = A$, and is given by

$$v^\star(z) = \frac{\alpha - 2}{2(\alpha - 1)} \left[\tau(1)\, z - \tau(z)\right] + A\, z. \tag{A.65}$$

The corresponding minimum is

$$\mathcal{Q}_0^\star(A) = A^2 + \frac{(\alpha - 2)\,\tau(1) + \alpha - \beta}{\alpha - 1}\, A + \frac{(\alpha - 2)^2 \left[|\tau(1)|^2 - \|\tau'\|_2^2\right]}{4(\alpha - 1)^2}. \tag{A.66}$$

An expression for the minimum over A is readily found, and it can be rearranged in the form (6.21) after noticing that $\tau(1) = \int_0^1 \tau'(z)\,\mathrm{d}z$ by virtue of (6.7).

A.8 Proof of the bound (6.48)

Consider a piecewise-linear scaled background field of the form

$$\rho(z) = \begin{cases} -R\, z & 0 \le z \le \delta, \\ -R\, \delta, & \delta \le z \le 1. \end{cases} \tag{A.67}$$

The boundary layer slope $R > 0$ and thickness $\delta > 0$ should be chosen to satisfy the spectral constraint (6.25) whilst optimising the bound on the Nusselt number,

$$\frac{1}{Nu} \ge \frac{1 - \|\rho' + 1\|_2 - \rho(1)}{2} = \frac{1 - \sqrt{1 + R\,(R - 2)\,\delta} + R\,\delta}{2}. \tag{A.68}$$

Recall from section 6.2 that the spectral constraint is equivalent to the quadratic form $\mathcal{Q}_k\{\hat\theta_k\}$ in (6.19) being positive semidefinite for all wavenumbers $k \ge 1$, and recall the change of variables $\alpha/(\alpha - 1)\tau(z) = \rho(z)$. Although the test function $\hat\theta_k$ is complex valued, the contributions of its real and imaginary parts to $\mathcal{Q}_k\{\hat\theta_k\}$ are identical and independent, so it suffices to consider real-valued test functions. Consequently, R and δ must be chosen such that, for all $k \ge 1$,

$$\mathcal{Q}_k\{v\} = \left\|v'\right\|_2^2 + k^2\, \|v\|_2^2 - Ma\, R\, v(1) \int_0^\delta f_k(z)\, v(z)\,\mathrm{d}z \ge 0 \tag{A.69}$$

for all real-valued functions $v(z)$ that satisfy the BCs $v(0) = 0$ and $v'(1) = 0$.

To bound the sign-indefinite term in (A.69), note that the BC $v(0) = 0$ and the Cauchy–Schwarz inequality imply

$$|v(1)| = \left| \int_0^1 v'(z)\,\mathrm{d}z \right| \le \left\|v'\right\|_2. \tag{A.70}$$

Moreover, since $|f_k(z)| = -f_k(z) \leq c\, z^2$ for $c \approx 0.943$ (Hagstrom & Doering, 2010),

$$\begin{aligned}
\left| Ma\, R\, v(1) \int_0^\delta f_k(z)\, v(z)\, \mathrm{d}z \right| &\leq Ma\, R\, c \left| \int_0^\delta \int_0^z z^2\, v'(\xi)\, \mathrm{d}\xi\, \mathrm{d}z \right| \left\| v' \right\|_2 \\
&= Ma\, R\, c \left| \int_0^\delta \int_\xi^\delta z^2\, v'(\xi)\, \mathrm{d}z\, \mathrm{d}\xi \right| \left\| v' \right\|_2 \\
&= \frac{Ma\, R\, c}{3} \left| \int_0^\delta \left(\delta^3 - \xi^3\right)\, v'(\xi)\, \mathrm{d}\xi \right| \left\| v' \right\|_2 \\
&\leq \frac{Ma\, R\, c}{3} \sqrt{\int_0^\delta \left(\delta^3 - \xi^3\right)^2\, \mathrm{d}\xi}\, \left\| v' \right\|_2^2 \\
&= \frac{Ma\, R\, c\, \delta^{7/2}}{\sqrt{14}}\, \left\| v' \right\|_2^2 . \qquad \text{(A.71)}
\end{aligned}$$

Inequality (A.69) therefore holds if

$$\delta = \left(\frac{Ma\, R\, c}{\sqrt{14}} \right)^{-2/7} . \qquad \text{(A.72)}$$

With this choice of δ, the asymptotic behaviour of the bound (A.68) as Ma tends to infinity is

$$\frac{1}{Nu} \geq \left(\frac{\sqrt{14}}{c} \right)^{2/7} \frac{R\,(4-R)}{4\, R^{2/7}} Ma^{-2/7}, \qquad \text{(A.73)}$$

and upon choosing $R = 5/3$ to maximise the prefactor one obtains

$$Nu \leq \frac{36}{35} \left(\frac{5\, c}{3\sqrt{14}} \right)^{2/7} Ma^{2/7} \approx 0.803\, Ma^{2/7} \quad \text{as } Ma \to \infty. \qquad \text{(A.74)}$$

A.9 Proof of (6.57)

Drop the suffix $\star$ from $\lambda_\star$ to ease the notation and rearrange (6.57) as

$$(\lambda - 1) \left\{ \lambda^{\gamma_1} Ma^{\gamma_1} + \frac{c}{[\ln(\lambda Ma)]^{\gamma_2}} \left[\gamma_1 - 1 + \frac{\gamma_2}{\ln(\lambda Ma)} \right] \right\} = \lambda^{\gamma_1} Ma^{\gamma_1} - \frac{c}{[\ln(\lambda Ma)]^{\gamma_2}}. \qquad \text{(A.75)}$$

To solve this equation when $Ma \gg 1$ using asymptotic expansions, introduce the ansatz

$$\lambda = k_0 + k_1 Ma^{-\gamma_1} (\ln Ma)^{-\gamma_2} + \cdots , \qquad \text{(A.76)}$$

where the notation $+\cdots$ means that higher-order terms are omitted. The exact form of the higher-order terms is not important to determine the constants k_0 and k_1. Equation (A.75) can be rewritten in a more convenient form using the fact that $Ma \gg 1$. To do this, one begins by considering $(\lambda Ma)^{\gamma_1}$, $\ln(\lambda Ma)$, and $[\ln(\lambda Ma)]^p$ for a generic exponent p.

First, use the fact that $(1+\varepsilon)^{\gamma_1} = 1 + \gamma_1\varepsilon + \cdots$ for $\varepsilon \ll 1$ to expand

$$\begin{aligned}(\lambda Ma)^{\gamma_1} &= Ma^{\gamma_1} k_0^{\gamma_1} \left[1 + \frac{k_1}{k_0} Ma^{-\gamma_1} (\ln Ma)^{-\gamma_2} + \cdots\right]^{\gamma_1} \\ &= Ma^{\gamma_1} k_0^{\gamma_1} \left[1 + \gamma_1 \frac{k_1}{k_0} Ma^{-\gamma_1} (\ln Ma)^{-\gamma_2} + \cdots\right] \\ &= Ma^{\gamma_1} k_0^{\gamma_1} + \gamma_1 k_0^{\gamma_1 - 1} k_1 (\ln Ma)^{-\gamma_2} + \cdots . \end{aligned} \tag{A.77}$$

Second, apply the properties of logarithms and the expansion $\ln(1+\varepsilon) = \varepsilon + \cdots$ for $\varepsilon \ll 1$ to write

$$\begin{aligned}\ln(\lambda Ma) &= \ln\left\{ Ma\, k_0 \left[1 + \frac{k_1}{k_0} Ma^{-\gamma_1} (\ln Ma)^{-\gamma_2} + \cdots\right]\right\} \\ &= \ln Ma + \ln k_0 + \ln\left[1 + \frac{k_1}{k_0} Ma^{-\gamma_1} (\ln Ma)^{-\gamma_2} + \cdots\right] \\ &= \ln Ma + \ln k_0 + \frac{k_1}{k_0} Ma^{-\gamma_1} (\ln Ma)^{-\gamma_2} + \cdots . \end{aligned} \tag{A.78}$$

Finally, (A.78) and the expansion $(1+\varepsilon)^p = 1 + p\varepsilon + \cdots$ yield

$$\begin{aligned}[\ln(\lambda Ma)]^p &= (\ln Ma)^p \left[1 + \ln k_0\, (\ln Ma)^{-1} + \frac{k_1}{k_0} Ma^{-\gamma_1} (\ln Ma)^{-\gamma_2 - 1} + \cdots\right]^p \\ &= (\ln Ma)^p \left[1 + p\, \ln k_0\, (\ln Ma)^{-1} + p\, \frac{k_1}{k_0} Ma^{-\gamma_1} (\ln Ma)^{-\gamma_2 - 1} + \cdots\right] \\ &= (\ln Ma)^p + p\, \ln k_0\, (\ln Ma)^{p-1} + p\, \frac{k_1}{k_0} Ma^{-\gamma_1} (\ln Ma)^{p - \gamma_2 - 1} + \cdots . \end{aligned} \tag{A.79}$$

Using (A.77) and (A.79) with $p = -\gamma_2$, the right-hand side of (A.75), denoted by R, becomes

$$\begin{aligned} R &:= \lambda^{\gamma_1} Ma^{\gamma_1} - \frac{c}{[\ln(\lambda Ma)]^{\gamma_2}} \\ &= Ma^{\gamma_1} k_0^{\gamma_1} + \gamma_1 k_0^{\gamma_1 - 1} k_1 (\ln Ma)^{-\gamma_2} + \cdots - c\, (\ln Ma)^{-\gamma_2} + \cdots \\ &= Ma^{\gamma_1} k_0^{\gamma_1} + \left[\gamma_1 k_0^{\gamma_1 - 1} k_1 - c\right] (\ln Ma)^{-\gamma_2} + \cdots . \end{aligned} \tag{A.80}$$

Similarly,

$$\frac{c}{[\ln(\lambda Ma)]^{\gamma_2}} \left[\gamma_1 - 1 + \frac{\gamma_2}{\ln(\lambda Ma)}\right] = c(\gamma_1 - 1)(\ln Ma)^{-\gamma_2} + \cdots , \tag{A.81}$$

so the term in curly braces on the left-hand side of (A.75), denoted S for convenience, becomes

$$\begin{aligned} S &:= \lambda^{\gamma_1} Ma^{\gamma_1} + \frac{c}{[\ln(\lambda Ma)]^{\gamma_2}} \left[\gamma_1 - 1 + \frac{\gamma_2}{\ln(\lambda Ma)}\right] \\ &= Ma^{\gamma_1} k_0^{\gamma_1} + \gamma_1 k_0^{\gamma_1 - 1} k_1 (\ln Ma)^{-\gamma_2} + \cdots + c(\gamma_1 - 1)(\ln Ma)^{-\gamma_2} + \cdots \\ &= Ma^{\gamma_1} k_0^{\gamma_1} + \left[\gamma_1 k_0^{\gamma_1 - 1} k_1 + c(\gamma_1 - 1)\right] (\ln Ma)^{-\gamma_2} + \cdots . \end{aligned} \tag{A.82}$$

The left-hand side of (A.75), denoted L for simplicity, can then be expanded as

$$\begin{aligned} L &:= (\lambda - 1)S \\ &= \left(k_0 - 1 + k_1 Ma^{-\gamma_1} (\ln Ma)^{-\gamma_2} + \cdots\right) \times \\ &\qquad \left\{ Ma^{\gamma_1} k_0^{\gamma_1} + \left[\gamma_1 k_0^{\gamma_1 - 1} k_1 + c(\gamma_1 - 1)\right] (\ln Ma)^{-\gamma_2} + \cdots \right\} \\ &= (k_0 - 1) k_0^{\gamma_1} Ma^{\gamma_1} \\ &\qquad + \left\{ (k_0 - 1) \left[\gamma_1 k_0^{\gamma_1 - 1} k_1 + c(\gamma_1 - 1)\right] + k_0^{\gamma_1} k_1 \right\} (\ln Ma)^{-\gamma_2} + \cdots . \end{aligned} \tag{A.83}$$

Equation (A.75) requires that $L = R$. Upon matching terms proportional to Ma^{γ_1} in (A.80) and (A.83) one finds

$$k_0^{\gamma_1} = (k_0 - 1) k_0^{\gamma_1} \quad \Rightarrow \quad k_0 = 2. \tag{A.84}$$

Similarly, after equating terms proportional to $(\ln Ma)^{-\gamma_2}$ in (A.80) and (A.83) and using the fact that $k_0 = 2$ one concludes that

$$\gamma_1 2^{\gamma_1 - 1} k_1 - c = \gamma_1 2^{\gamma_1 - 1} k_1 + c(\gamma_1 - 1) + 2^{\gamma_1} k_1 \quad \Rightarrow \quad k_1 = -c\,\gamma_1\, 2^{-\gamma_1}. \tag{A.85}$$

Substituting the values of k_0 and k_1 into (A.76) gives (6.57).

A.10 Proof of (6.69)

Since any test function $v \in \Gamma$ vanishes at $z = 0$, integration by parts shows that for any constant $\gamma \geq 0$

$$\gamma\, |v(1)|^2 - 2\,\gamma \int_0^1 v\, v'\, \mathrm{d}z = 0. \tag{A.86}$$

Adding this to the quadratic form $\mathcal{Q}_k\{v\}$ in (6.68) and using the Cauchy–Schwarz inequality to estimate the sign-indefinite terms yields

$$\mathcal{Q}_k\{v\} \geq \left\|v'\right\|_2^2 + k^2 \left\|v\right\|_2^2 + \gamma\, |v(1)|^2 - 2\,\gamma \left\|v'\right\|_2 \left\|v\right\|_2 - Ma\, \|(\phi - 1)\, f_k\|_2\, |v(1)|\, \left\|v\right\|_2, \tag{A.87}$$

so $\mathcal{Q}_k\{v\} \geq 0$ if

$$\left\|v'\right\|_2^2 - 2\,\gamma \left\|v'\right\|_2 \left\|v\right\|_2 + \omega k^2 \left\|v\right\|_2^2 \geq 0, \tag{A.88a}$$

$$(1 - \omega)\, k^2 \left\|v\right\|_2^2 - Ma\, \|(\phi - 1) f_k\|_2\, |v(1)|\, \left\|v\right\|_2 + \gamma\, |v(1)|^2 \geq 0, \tag{A.88b}$$

for some scalar $\omega \in (0, 1)$. Recalling that a quadratic form $ax^2 + bxy + cy^2$ is non-negative for all x and y if $b^2 \leq 4ac$, and choosing $\gamma = \sqrt{\omega} k$ to complete the square in (A.88a), one

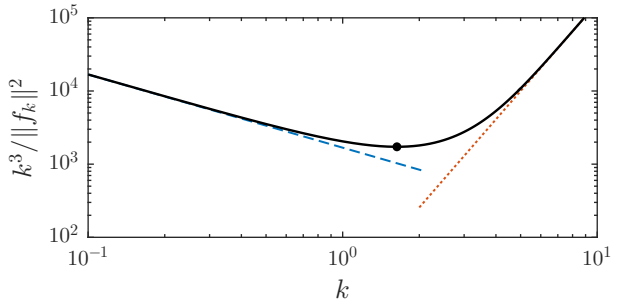


FIGURE A.1: Plot of $k^3/\|f_k\|_2^2$ (——) along with its small-k asymptote, $1680k^{-1}$ (- - -), and its large-k asymptote, $16k^4$ (·····). The minimum (•) is at $k \approx 1.633$.

obtains that $\mathcal{Q}_k\{v\} \geq 0$ if

$$Ma^2 \|(\phi - 1)\, f_k\|_2^2 \leq 4\,(1-\omega)\,\sqrt{\omega}\,k^3. \tag{A.89}$$

After setting $\omega = 1/3$ to maximise the right-hand side, estimating

$$\|(\phi - 1)\, f_k\|_2 \leq \|\phi - 1\|_\infty \|f_k\|_2\,, \tag{A.90}$$

and rearranging, one arrives at

$$\frac{k^3}{\|f_k\|_2^2} \geq \frac{3\sqrt{3}}{8}\, Ma^2 \|\phi - 1\|_\infty^2\,. \tag{A.91}$$

Figure A.1 demonstrates that $k^3/\|f_k\|_2^2$ has a minimum at $k = k_{\text{crit}} \approx 1.633$, grows asymptotically to $1680\,k^{-1}$ as $k \to 0$, and quickly asymptotes to $16\,k^4$ for $k > k_{\text{crit}}$. In fact $k^3/\|f_k\|_2^2 \geq 16k^4$ so (A.91)—and hence the Fourier-transformed spectral constraint $\mathcal{Q}_k\{v\} \geq 0$—holds for all wavenumbers larger than the critical value

$$k_c := \left\lfloor \left(\frac{3\sqrt{3}}{128}\right)^{1/4} Ma^{1/2} \|\phi - 1\|_\infty^{1/2} \right\rfloor. \tag{A.92}$$

Appendix B

Energy stability of stress-driven shear flows in finite periodic domains

Consider the same stress-driven shear flow described in section 5.1. Energy stability analysis (Hagstrom & Doering, 2014) shows that the laminar flow $\boldsymbol{u}_\ell = Gr\, z\, \boldsymbol{e}_1$ is globally stable for all Grashoff numbers up to the critical value

$$\begin{aligned} Gr_E := \sup_{Gr} \quad & Gr, \\ \text{s.t.} \quad & \mathcal{E}\{\boldsymbol{u}\} := \int_{\Omega_d} \|\nabla \boldsymbol{u}\|_{\text{F}}^2 + Gr\, u\, w \, \mathrm{d}^d\boldsymbol{x} \geq 0 \quad \forall \boldsymbol{u} \in H, \end{aligned} \tag{B.1}$$

where $d = 2$ or 3 depending on whether the two- or the three-dimensional flow model is being considered. The space H, defined in equation (5.12), is the space of smooth functions that satisfy incompressibility and the homogeneous version of the flow's boundary conditions (BCs), given explicitly in (5.7).

The energy stability problem (B.1) for two- and three-dimensional fluid layers that extend to infinity in the horizontal directions (*i.e.*, $\Gamma_x, \Gamma_y \to \infty$) was solved by Hagstrom & Doering (2014), who established that $Gr_E \approx 51.7300$ in three dimensions and $Gr_E \approx 139.5396$ in two dimensions. These values are (sharp) lower bounds on the largest Gr_E for a periodic layer with finite aspect ratios Γ_x, Γ_y, because in this case perturbations are defined only by a countable set of horizontal Fourier modes. This appendix describes how the critical Grashoff number Gr_E for energy stability of the laminar flow in finite periodic layers can be estimated accurately, from above and from below, using the methods of chapter 4. This is done for two reasons. First, knowing Gr_E for prescribed finite values of the horizontal periods Γ_x, Γ_y (Γ_x only in two dimensions) is useful to verify that the bounds on the energy dissipation coefficients C_ε computed in chapter 5 are correct. Second, it is demonstrated that using SDPs is an attractive alternative to the traditional approach to energy stability problems in fluid mechanics, based on the solution of boundary-eigenvalue problems.

B.1 Energy stability in two dimensions

In two spatial dimensions, each function $\boldsymbol{u} \in H$ can be expanded using the same Fourier series introduced in section 5.3.3. In fact, steps similar to those outlined in section 5.3.3 show that the functional $\mathcal{E}\{\boldsymbol{u}\}$ is positive semidefinite for all $\boldsymbol{u} \in H$ if and only if, for all positive integers m, the quadratic form

$$\mathcal{E}_m\{W_m\} := \int_0^1 \frac{1}{\alpha_m^2} \left|W_m''(z)\right|^2 + 2\left|W_m'(z)\right|^2 + \alpha_m^2 \left|W_m(z)\right|^2 - \frac{Gr}{\alpha_m} \mathrm{Im}\left[W_m'(z) W_m^*(z)\right] \, \mathrm{d}z \tag{B.2}$$

is non-negative for all complex-valued functions W_m that satisfy the BCs

$$W_m(-1) = W_m(1) = W_m'(-1) = W_m''(1) = 0. \tag{B.3}$$

In these expressions, $\alpha_m = 2\pi m/\Gamma_x$ is the wavenumber in the horizontal direction, W_m is the vertical component of the m-th mode in the Fourier expansion of $\boldsymbol{u}$, and the superscript $*$ denotes complex conjugation.

For any value of Gr, only a finite number of values m need be considered explicitly, because $\mathcal{E}_m\{W_m\}$ is positive when m is large. In fact, upon applying the Cauchy-Schwarz inequality and the elementary inequality $ab \leq a^2/(\sqrt{2}\alpha_m) + \alpha_m b^2/(2\sqrt{2})$ one can bound

$$\begin{aligned} \mathcal{E}_m\{W_m\} &\geq 2\left\|W_m'\right\|_2^2 + \alpha_m^2 \left\|W_m\right\|_2^2 - \frac{Gr}{\alpha_m} \left\|W_m'\right\|_2 \left\|W_m\right\|_2 \\ &\geq \left(1 - \frac{Gr}{2\sqrt{2}\alpha_m^2}\right) \left(2\left\|W_m'\right\|_2^2 + \alpha_m^2 \left\|W_m\right\|_2^2\right). \end{aligned} \tag{B.4}$$

Consequently, for any fixed Gr the inequality $\mathcal{E}_m\{W_m\} \geq 0$ holds for all m such that $\alpha_m^2 \geq Gr/(2\sqrt{2})$ or, equivalently, for all m larger than the critical value

$$m_{\mathrm{cr}}(Gr) := \left\lfloor \frac{\Gamma_x}{\pi} \sqrt{\frac{Gr}{8\sqrt{2}}} \right\rfloor. \tag{B.5}$$

Then, upon replacing the inequality[1] $\mathcal{E}\{\boldsymbol{u}\} \geq 0$ in problem (B.1) with the set of constraints $\mathcal{E}_m\{W_m\} \geq 0$, $m \leq m_{\mathrm{cr}}(Gr)$, one concludes that

$$\begin{aligned} Gr_E = \sup_{Gr} \quad & Gr, \\ \text{s.t.} \quad & \mathcal{E}_m\{W_m\} \geq 0 \quad \forall W_m \text{ satisfying (B.3)}, \quad m = 1, 2, \ldots, m_{\mathrm{cr}}(Gr). \end{aligned} \tag{B.6}$$

[1] All functional inequalities should be understood as being imposed over the set of admissible argument functions, even when this is not explicitly specified. For instance, the inequality $\mathcal{E}\{\boldsymbol{u}\} \geq 0$ is imposed for all $\boldsymbol{u} \in H$, while the inequality $\mathcal{E}_m\{W_m\} \geq 0$ is imposed for all W_m that satisfy (B.3).

TABLE B.1: Upper and lower bounds on Gr_{E} for the two-dimensional flow model and two values of the horizontal period, $\Gamma_x = 2$ and $\Gamma_x = 3$. The tabulated upper and lower bounds were computed, respectively, through the solution of outer and inner SDP approximations of (B.6) set up with QUINOPT using degree-N Legendre expansions.

	$\Gamma_x = 2$		$\Gamma_x = 3$	
N	Lower bound	Upper bound	Lower bound	Upper bound
2	0.000000	∞	0.000000	∞
4	0.000000	∞	131.254405	∞
6	46.831187	140.177761	148.653886	149.387859
8	139.539920	139.545729	148.662378	148.673814
10	139.539934	139.539943	148.662419	148.662430
12	139.539935	139.539935	148.662419	148.662419

This is an optimisation problem with affine homogeneous integral inequality constraints of the type studied in chapter 4, so rigorous (modulo numerical roundoff error) upper and lower bounds for Gr_E can be computed using semidefinite programming. More precisely, after replacing each integral inequality $\mathcal{E}_m\{W_m\} \geq 0$ with an LMI derived using the outer approximation method of section 4.3 one calculates an upper bound on Gr_E, because the constraints are relaxed. Instead, a lower bound is obtained when the inner approximation technique of section 4.4 is employed because the constraints are strengthened.

The only complication preventing a direct implementation of inner and outer SDP approximations of (B.6) is that the number of constraints depends on the optimisation variable, and is therefore unknown. However, one can use the same iterative procedure described at the end of section 5.3.3. First, fix and integer m_0 and compute an upper (resp. lower) bound B on Gr_E by solving an outer (resp. inner) SDP approximation of (B.6) considering only constraints with $m \leq m_0$. If $m_0 < m_{\mathrm{cr}}(B)$, check that the candidate bound B is actually feasible by verifying that, for all $m = m_0 + 1, \ldots, m_{\mathrm{cr}}(B)$, the quadratic forms $\mathcal{E}_m\{W_m\}$ are positive semidefinite for all functions W_m satisfying (B.3). If these checks fail, repeat the optimisation with larger m_0.

These steps were implemented in MATLAB for a range of values of the horizontal period Γ_x, using SDPT3 to solve the inner and outer SDP approximations of (B.6) set up with QUINOPT. As demonstrated in table B.1 for the cases $\Gamma_x = 2$ and $\Gamma_x = 3$, the upper and lower bounds on Gr_E converge to each other as the degree of the Legendre expansions carried out by QUINOPT, denoted by N in the table and in the following, is raised. The bounds agree to 6 decimal places for N as low as 12, meaning that the optimal solution Gr_E of the original infinite-dimensional problem (B.1) can be estimated accurately through the solution of small SDPs.

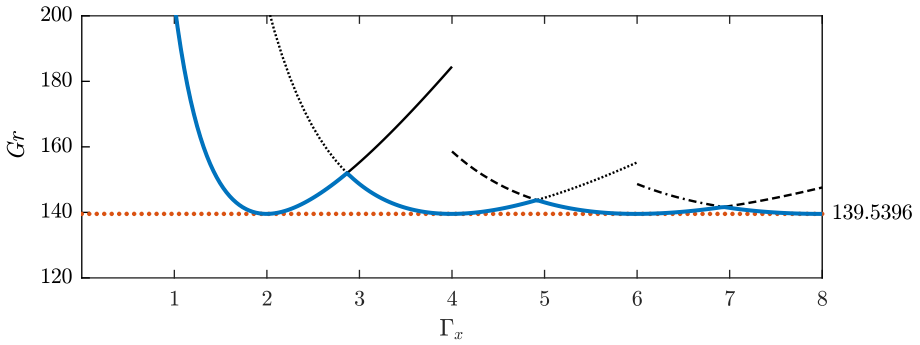


FIGURE B.1: Converged lower bounds on Gr_{E} for the two-dimensional flow model as a function of the horizontal period Γ_x (——). Also plotted are the asymptotic value $Gr_{\mathrm{E}} \approx 139.5396$ computed by Hagstrom & Doering (2014) (·····) and converged lower bounds on the critical Grashoff number for energy stability of individual Fourier modes: $m = 1$ (——), $m = 2$ (·····), $m = 3$ (- - -), and $m = 4$ (-·-·-).

Converged lower bounds on Gr_E, meaning that they improve by less than 0.1% when N is increased, are shown in figure B.1 for $\Gamma_x \leq 8$. Similarly converged upper bounds are not plotted as they would be visually indistinguishable. As Γ_x grows to infinity, Gr_E tends to the asymptotic value 139.5396 computed by Hagstrom & Doering (2014), which bounds Gr_E from below at any Γ_x. The local minima of Gr_E saturate this lower bound and correspond to pairs (m, Γ_x) such that $\alpha_m = 2\pi m/\Gamma_x \approx 3.1469$. This value is in excellent agreement with the critical wavenumber 3.146899 reported by Hagstrom & Doering (2014) for the infinite-layer case. Finally, corner points correspond to values of Γ_x at which the critical Fourier mode in (B.6) changes. To illustrate this more clearly, converged lower bounds on the critical Grashoff number for the energy stability of individual Fourier modes were computed by solving (B.6) after discarding the constraints for all but a single value m. The results obtained when considering only $m = 1$, 2, 3, or 4 are also plotted in figure B.1.

B.2 Energy stability in three dimensions

To simplify the energy stability analysis in three dimensions, it will be assumed that the critical perturbations are streamwise independent (Tang *et al.*, 2004; Hagstrom & Doering, 2014). Then, a Fourier expansion in the cross-stream (y) direction similar to that used in section 5.3.4 shows that $\mathcal{E}\{\boldsymbol{u}\}$ is non-negative for all streamwise-independent perturbations $\boldsymbol{u} \in H$ if and only if, for each positive integer m, the quadratic form

$$\mathcal{E}_m\{U_m, W_m\} := \int_0^1 \left|U_m'(z)\right|^2 + \beta_m^2 \left|U_m(z)\right|^2 + \frac{1}{\beta_m^2}\left|W_m''(z)\right|^2 + 2\left|W_m'(z)\right|^2 + \beta_m^2 \left|W_m(z)\right|^2 + Gr\, U_m(z) W_m(z)\,\mathrm{d}z \quad \text{(B.7)}$$

is positive semidefinite for all real-valued functions U_m and W_m that satisfy the BCs

$$U_m(0) = W_m'(0) = W_m(0) = U_m'(1) = W_m''(1) = W_m(1) = 0. \tag{B.8}$$

In these expressions, $\beta_m = 2\pi m/\Gamma_y$ is the wavenumber in the cross-stream direction, while U_m and W_m correspond to the streamwise and vertical components of the m-th Fourier mode in the expansion of $\boldsymbol{u}$.[2]

Upon estimating

$$\mathcal{E}_m\{U_m, W_m\} \geq \beta_m^2 \|U_m\|_2^2 + Gr\|U_m\|_2 \|W_m\|_2 + \beta_m^2 \|W_m\|_2^2, \tag{B.9}$$

one concludes that, for any value Gr, the quadratic form $\mathcal{E}_m\{U_m, W_m\}$ is non-negative if $\beta_m^2 \geq Gr/2$, which is true if m is above the critical value

$$m_{\mathrm{cr}}(Gr) := \left\lfloor \frac{\Gamma_y}{\pi}\sqrt{\frac{Gr}{8}} \right\rfloor. \tag{B.10}$$

Consequently, to compute the critical Grashoff number Gr_E for energy stability one can replace the constraint $\mathcal{E}\{\boldsymbol{u}\} \geq 0$ in (B.1) with the requirement that $\mathcal{E}_m\{U_m, W_m\}$ is positive semidefinite for all $m \leq m_{\mathrm{cr}}(Gr)$, and then solve

$$\begin{aligned} Gr_E = \sup_{Gr} \quad & Gr, \\ \text{s.t.} \quad & \mathcal{E}_m\{U_m, W_m\} \geq 0, \quad m = 1, 2, \ldots, m_{\mathrm{cr}}(Gr). \end{aligned} \tag{B.11}$$

Note that each constraint in this problem should be understood as being imposed for all functions U_m and W_m that satisfy the BCs (B.8). Note also that problem (B.11) is independent of the period in the x direction (Γ_x) due to the assumption of streamwise invariance, and one is interested in the variation of Gr_E with Γ_y, the period in the cross-stream direction.

As for the two-dimensional energy stability problem, Gr_E can be bounded from above and from below with semidefinite programming after each integral inequality in (B.11) is replaced with LMIs derived using the methods of chapter 4. Moreover, the same iterative procedure described in the previous section should be employed because the number of constraints in (B.11) depends on the decision variable and is unknown *a priori*.

Upper and lower bounds on Gr_E were computed for a range of values $\Gamma_y \leq 8$, using QUINOPT and SDPT3 to set up and solve all relevant SDPs. As in the two-dimensional

[2]Strictly speaking, the modes in the Fourier expansion of $\boldsymbol{u}$ are complex-valued, but in order to enforce the non-negativity of the functional $\mathcal{E}_m$ it suffices to consider real-valued functions. This can be justified using an argument analogous to that outlined in section 5.3.4.

TABLE B.2: Upper and lower bounds on Gr_E for the three-dimensional problem and two values of the horizontal period, $\Gamma_x = 2$ and $\Gamma_x = 3$, under the assumption that critical modes are streamwise invariant. The tabulated upper and lower bounds were computed, respectively, through the solution of outer and inner SDP approximations of (B.11) set up with QUINOPT using degree-N Legendre expansions.

	$\Gamma_y = 2$		$\Gamma_y = 3$	
N	Lower bound	Upper bound	Lower bound	Upper bound
2	0.000000	∞	51.575997	∞
4	57.195992	57.251133	51.729563	51.783832
6	57.198881	57.199042	51.730510	51.730546
8	57.198882	57.198883	51.730510	51.730511
10	57.198882	57.198882	51.730510	51.730510
12	57.198882	57.198882	51.730510	51.730510

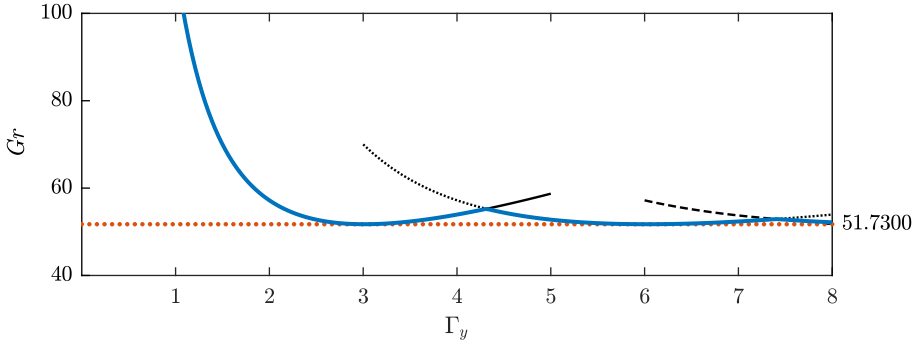


FIGURE B.2: Converged lower bounds on Gr_{E} for the three-dimensional problem, computed assuming that the critical modes are streamwise invariant, as a function of the horizontal period Γ_y in the cross-stream direction (——). Also plotted are the asymptotic value $Gr_{\mathrm{E}} \approx 51.7300$ computed by Hagstrom & Doering (2014) (·····) and converged lower bounds on the critical Grashoff number for energy stability of individual Fourier modes: $m = 1$ (——), $m = 2$ (·····), and $m = 3$ (- - -).

case, the upper and lower bounds converge to each other as N, the degree of the Legendre series expansions carried out by QUINOPT to set up the SDPs, is raised. Table B.2 demonstrates this for $\Gamma_y = 2$ and $\Gamma_y = 3$, but the same was observed for all other values Γ_y considered. Converged lower bounds on Gr_E, computed by increasing N until the optimal value of the inner SDP approximation of (B.11) increased by less than 0.1%, are plotted in figure B.2. The critical Grashoff number for energy stability of the first three Fourier modes ($m = 1$, 2, and 3) is also shown to demonstrate that sharp corners as Γ_y is increased correspond to changes in the critical Fourier mode. Similar to the two-dimensional case, the converged bounds on Gr_E approach the asymptotic value $Gr_E \approx 51.7300$ computed by Hagstrom & Doering (2014), and the local minima saturating this bound correspond to pairs (m, Γ_y) such that $\beta_m \approx 2.0856$. This value agrees extremely well with the critical wavenumber 2.085586 found by Hagstrom & Doering (2014) in the limit of infinite Γ_y.